

TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN



KỶ YẾU HỘI THẢO KHOA HỌC VÀ CÔNG NGHỆ



Tp. Hồ Chí Minh, ngày 24 tháng 5 năm 2025

HỘI THẢO KHOA HỌC VÀ CÔNG NGHỆ

KHOA CÔNG NGHỆ THÔNG TIN

Ban tổ chức

Trưởng ban

TS. Nguyễn Hồng Vũ

Trường Đại học Công Thương TP. HCM

Phó Trưởng ban

TS. Nguyễn Thanh Long

Trường Đại học Công Thương TP. HCM

Phó Trưởng ban (TT)

TS. Nguyễn Thị Định

Trường Đại học Công Thương TP. HCM

Thành viên

TS. Phùng Thế Bảo

Trường Đại học Công Thương TP. HCM

TS. Nguyễn Thị Bích Ngân

Trường Đại học Công Thương TP. HCM

TS. Vũ Đức Thịnh

Trường Đại học Công Thương TP. HCM

TS. Trần Việt Hùng

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Văn Lễ

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Thế Hữu

Trường Đại học Công Thương TP. HCM

ThS. Trần Đình Toàn

Trường Đại học Công Thương TP. HCM

ThS. Vũ Văn Vinh

Trường Đại học Công Thương TP. HCM

ThS. Lâm Thị Hoà Mi

Trường Đại học Công Thương TP. HCM

ThS. Trần Đức Tốt

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Quốc Sử

Trường Đại học Công Thương TP. HCM

Ban chương trình

Trưởng ban

TS. Nguyễn Thị Định

Trường Đại học Công Thương TP. HCM

Phó Trưởng ban

TS. Nguyễn Thị Bích Ngân

Trường Đại học Công Thương TP. HCM

Phó trưởng ban (TT)

ThS. Vũ Văn Vinh

Trường Đại học Công Thương TP. HCM

Thành viên

PGS.TS. Vũ Thanh Nguyên

Trường Đại học Công Thương TP. HCM

TS. Nguyễn Hồng Vũ

Trường Đại học Công Thương TP. HCM

TS. Nguyễn Thanh Long

Trường Đại học Công Thương TP. HCM

TS. Nguyễn Thị Định

Trường Đại học Công Thương TP. HCM

TS. Phạm Nguyễn Huy Phương

Trường Đại học Công Thương TP. HCM

ThS. Ngô Dương Hà

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Văn Lễ

Trường Đại học Công Thương TP. HCM

ThS. Trần Đình Toàn

Trường Đại học Công Thương TP. HCM

Ban thư ký và hậu cần

Trưởng ban

ThS. Lê Thị Thuỳ Lan

Trường Đại học Công Thương TP. HCM

Phó Trưởng ban

ThS. Trần Thị Vân Anh

Trường Đại học Công Thương TP. HCM

Thành viên

ThS. Đinh Thị Mận

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Phương Hạc

Trường Đại học Công Thương TP. HCM

ThS. Phan Thị Ngọc Mai

Trường Đại học Công Thương TP. HCM

ThS. Lâm Thị Họa Mi

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Thị Thu Tâm

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Thị Thanh Thủy

Trường Đại học Công Thương TP. HCM

ThS. Nguyễn Thị Hồng Thảo

Trường Đại học Công Thương TP. HCM

CN. Lương Thị Quỳnh Mai

Trường Đại học Công Thương TP. HCM

MỤC LỤC

1. TÌM KIẾM ẢNH DỰA TRÊN ONTOLOGY KẾT HỢP MÔ HÌNH HỌC SÂU	
<i>Võ Thái Bình, Phạm Minh Thuận, Nguyễn Trần Duy Khánh, Nguyễn Thị Định*</i>	
.....	1
2. TỐI ƯU HÓA CHI PHÍ VÀ THỜI GIAN VẬN CHUYỂN HÀNG HÓA BẰNG THUẬT TOÁN MULTI-VERSE OPTIMIZATION (MVO)	
<i>Nguyễn Thị Thùy Trang, Phan Thị Ngọc Mai*</i>	17
3. TỐI ƯU HÓA CHIM HẢI ÂU BẮC CỰC TRONG CÁC BÀI TOÁN TỐI ƯU	
<i>Bạch Ngọc Vy, Phùng Thị Ly, Đinh Nguyễn Trọng Nghĩa*</i>	30
4. TỐI ƯU HÓA THUẬT TOÁN ĐÀO TẠO ĐỘI BÓNG ĐÁ VÀ ỨNG DỤNG TRONG BÀI TOÁN TỐI ƯU HÓA	
<i>Trần Đình Hải Nguyên, Nguyễn Đình Đức, Đinh Nguyễn Trọng Nghĩa*</i>	44
5. THUẬT TOÁN KEPLER VÀ ỨNG DỤNG CHO CÁC BÀI TOÁN TỐI ƯU	
<i>Nguyễn Thế Hữu, Hồ Nguyễn Linh Chi, Nguyễn Thị Ngọc, Huỳnh Thị Bích Ngọc, Lê Thị Như Quỳnh, Đinh Nguyễn Trọng Nghĩa*</i>	60
6. THUẬT TOÁN TỐI ƯU DILOPHOSAURUS (DOA): NÂNG CAO HIỆU SUẤT CỦA THUẬT TOÁN T-REX (TROA)	
<i>Nguyễn Thế Hữu, Lều Công Trung Hiếu, Đặng Nguyên Hoài, Đinh Nguyễn Trọng Nghĩa*</i>	73
7. MÔ HÌNH ĐỀ XUẤT KẾT HỢP COATI VÀ MVO TRONG HỆ THỐNG HỖ TRỢ QUYẾT ĐỊNH ĐẦU TƯ	
<i>Huỳnh Thị Châu Lan*, Trần Huỳnh Phụng Quyên, Dương Trọng Đức, Hứa Vĩnh Khang, Đinh Nguyễn Trọng Nghĩa</i>	84
8. BÀI TOÁN K-ANONYMITY VÀ PHƯƠNG PHÁP GIẢI BẰNG META-HEURISTIC	
<i>Lê Hoàng Minh Nhật *, Đinh Nguyễn Trọng Nghĩa</i>	95
9. PHƯƠNG PHÁP LẬP LỊCH TRƯỜNG ĐẠI HỌC BẰNG THUẬT TOÁN METAHEURISTIC	
<i>Phạm Nguyễn, Đinh Nguyễn Trọng Nghĩa*</i>	105
10. AI VỚI HỌC TẬP TRỰC TUYẾN	
<i>Nguyễn Phương Hạc, Nguyễn Thị Hồng Thảo, Trần Võ Phong, Nguyễn Thành Hậu, Lương Bảo Hòa</i>	121

11. KẾT HỢP TỐI ƯU HÓA TRƯỜNG ĐIỆN TỪ VÀ THUẬT GIẢI DI TRUYỀN: MỘT PHƯƠNG PHÁP METAHEURISTIC ĐỂ TỐI ƯU HÓA TĂNG CƯỜNG	
<i>Nguyễn Thị Thu Tâm*, Đinh Nguyễn Trọng Nghĩa</i>	130
12. ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG HỌC TẬP CỦA SINH VIÊN TẠI CÁC TRƯỜNG ĐẠI HỌC HIỆN NAY	
<i>Nguyễn Thị Thu Tâm*, Lê Thị Thùy Lan</i>	141
13. THUẬT TOÁN CONFIGURATION FILTERING HỖ TRỢ GỢI Ý VIDEO TRÊN NỀN TẢNG GIẢI TRÍ HTB	
<i>Huỳnh Công Huy, Nguyễn Quốc Tiến, Nguyễn Quốc Bảo, Nguyễn Hải Yến*</i>	151
14. TỐI ƯU PHÂN CỤM MỒ VỚI WHALE OPTIMIZATION ALGORITHM	
<i>Nguyễn Thị Thùy Trang, Phan Thị Ngọc Mai*</i>	163
15. ỨNG DỤNG CÁC CÔNG CỤ TÍCH HỢP AI TRONG THIẾT KẾ WEBSITE:	
XU HƯỚNG VÀ TIỀM NĂNG PHÁT TRIỂN	
<i>Lâm Thị Họa Mi *</i>	173
16. AI – GIẢI PHÁP TỐI ƯU HÓA WEBSITE	
<i>Dương Thị Mộng Thùy*</i>	181
17. XÂY DỰNG HỆ THỐNG KHUYẾN NGHỊ CHỌN MỸ PHẨM DỰA TRÊN MÔ HÌNH HỌC MÁY	
<i>Nguyễn Thị Bích Ngân*, Nguyễn Thị Bích Tuyền, Đinh Thị Tâm</i>	190
18. MÔ HÌNH NHẬN DIỆN VÀ CHÚ THÍCH BIỂN BÁO GIAO THÔNG DÙNG MẠNG NƠON HỌC SÂU	
<i>Bùi Công Danh*</i>	205
19. MỘT SỐ PHƯƠNG PHÁP TRONG XÂY DỰNG HỆ THỐNG GỢI Ý	
<i>Đinh Thị Tâm</i>	217
20. KHẢO SÁT TÁC ĐỘNG CỦA AI ĐẾN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN TẠI ĐẠI HỌC CÔNG THƯƠNG TP.HỒ CHÍ MINH	
<i>Nguyễn Quốc Sử*, Trần Thị Bích Vân</i>	224
21. TÍCH HỢP CÔNG NGHỆ THỦ KÍNH ẢO THỜI GIAN THỰC TRÊN WEBSITE BÁN MẮT KÍNH THỜI TRANG DỰA TRÊN FACEMESH VÀ DLIB	
<i>Nguyễn Trọng Hoàng, Phúc Ngô Dương Hà, Trần Như Ý*</i>	235
22. MỘT KHẢO SÁT VỀ CÁC PHƯƠNG PHÁP XÁC ĐỊNH NGUỒN CHO BÀI TOÁN KHAI THÁC TẬP HỮU ÍCH CAO	
<i>Nguyen Van Lê *</i>	251

23. XÁC THỰC VÀ BẢO MẬT HÌNH ẢNH BẰNG CẤP DÙNG BLOCKCHAIN HYPERLEDGER FABRIC	
<i>Lê Quang Vinh*, Nguyễn Thị Thanh Thảo, Huỳnh Minh An, Lâm Nguyễn Anh Hào, Nguyễn Ngọc Nghĩa, Đặng Ngọc Tài, Hà Huỳnh Ánh Ngân, Huỳnh Công Tiến, Phan Công Hiệu, Nguyễn Thị Bích Ngân.....</i>	259
24. NGHIÊN CỨU THUẬT TOÁN KHAI THÁC TẬP HỮU ÍCH CAO TRÊN CƠ SỞ DỮ LIỆU ÂM DƯƠNG THAY ĐỔI VÀ ỨNG DỤNG VÀO WEBSITE BÁN SẢN PHẨM ĐIỆN TỬ	
<i>Nguyễn Huy Hoàng, Nguyễn Công Dũng, Hà Trọng Thắng, Vũ Văn Vinh*</i>	270
25. HỆ THỐNG ĐIỂM DANH SINH VIÊN BẰNG KHUÔN MẶT TẠI KÝ TÚC XÁ TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH	
<i>Bùi Hùng Phương, Huỳnh Thanh Triệu, Huỳnh Vĩ Khang, Trần Thị Vân Anh*</i>	285
26. XÂY DỰNG MÔ HÌNH PHÂN LOẠI CẢM XÚC DỰA TRÊN HỌC SÂU VÀ ỨNG DỤNG TRONG MẠNG XÃ HỘI ĐỂ PHÁT HIỆN VÀ NGĂN CHẶN BÌNH LUẬN KHÔNG PHÙ HỢP	
<i>Hà Trọng Thắng, Hồ Thị Kiều Oanh, Nguyễn Thanh Vi, Nguyễn Huy Hoàng, Nguyễn Công Dũng, Trần Thị Vân Anh*</i>	295
27. DỰ ĐOÁN GIÁ THỊ TRƯỜNG BITCOIN THÔNG QUA PHÂN TÍCH CẢM XÚC TIN TỨC DÙNG MONGODB	
<i>Trần Việt Hùng*, Ngô Ngọc Huỳnh Như</i>	307
28. ĐÁNH GIÁ HIỆU SUẤT CỦA MÔ HÌNH SINH KHUẾCH TÁN	
<i>Hoàng Thế Anh, Trần Đình Toàn*</i>	316
29. NGHIÊN CỨU THUẬT TOÁN KHAI THÁC TẬP HỮU ÍCH CAO CHÉO (CROSS-LEVEL HUI) TRÊN CƠ SỞ DỮ LIỆU PHÂN CẤP	
<i>Vũ Văn Vinh*, Phạm Tấn Thuận, Khúc Hồng Nhung, Trần Ngọc Hiếu Thảo</i>	325
30. GIẢI PHÁP BẢO MẬT CHO MẠNG XÃ HỘI THU NHỎ	
<i>Phạm Đình Văn, Nhiêu Hồng Phúc, Nguyễn Như Quang Tùng, Trần Thị Vân Anh, Đinh Thị Mận*</i>	337
31. HỆ THỐNG QUẢN LÝ DINH DƯỠNG THÔNG MINH	
<i>Võ Ngọc Nguyên Chương, Trần Trí Cường, Phạm Vũ Lương Bằng, Trần Vinh Hiển, Đỗ Quang Vinh, Trần Văn Thọ*</i>	351

TÌM KIẾM ẢNH DỰA TRÊN ONTOLOGY KẾT HỢP MÔ HÌNH HỌC SÂU

Võ Thái Bình¹, Phạm Minh Thuận¹, Nguyễn Trần Duy Khánh¹,
Nguyễn Thị Định^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: dinhnt@huit.edu.vn

Ngày nhận bài: 30/3/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Bài toán truy vấn ảnh dựa trên ngữ nghĩa ngày càng được ứng dụng trong nhiều lĩnh vực như giáo dục, y tế, du lịch, sinh học môi trường và thực hiện theo nhiều phương pháp khác nhau như sử dụng các mô hình máy học, kết hợp các mô hình transformer với nhiều loại dữ liệu khác nhau. Trong bài báo này, một cấu trúc ontology được xây dựng cho tập ảnh đa nhãn và áp dụng cho hệ truy vấn hình ảnh theo mô tả ngữ nghĩa được đề xuất. Để thực hiện bài toán này, mỗi ảnh đa nhãn được phân đoạn thành các ảnh đơn nhãn bằng mạng học sâu YOLO. Phương pháp đề xuất khai thác thông tin ngữ nghĩa từ các chú thích (captions) đi kèm hình ảnh để xây dựng một khung tri thức (ontology) có cấu trúc. Quá trình này bao gồm việc tiền xử lý dữ liệu văn bản, áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên như POS tagging và dependency parsing nhằm trích xuất các thành phần ngữ nghĩa, phân loại từ thành các nhóm thực thể (entities), thuộc tính (attributes) và mối quan hệ giữa các thực thể (relations). Sau đó, các thông tin thu được được tổ chức thành một ontology và lưu trữ trên GraphDB. Hệ thống cho phép thực hiện truy vấn qua hai hình thức: mô tả ngôn ngữ tự nhiên và thông qua ảnh. Kết quả thử nghiệm trên bộ dữ liệu Flickr30k cho thấy phương pháp đề xuất cải thiện đáng kể độ chính xác và hiệu quả truy vấn ảnh theo ngữ nghĩa, mở ra hướng đi mới trong lĩnh vực truy vấn ảnh dựa trên tri thức ngữ nghĩa.

Từ khóa: Truy vấn ảnh, ngữ nghĩa ảnh, Ontology, NLP, học sâu.

1. MỞ ĐẦU

Trong bối cảnh dữ liệu hình ảnh từ mạng xã hội và từ các kho dữ liệu số bùng nổ, việc quản lý và truy xuất thông tin hình ảnh trở thành thách thức lớn. Các hệ thống truy vấn ảnh theo nội dung (CBIR) dựa trên đặc trưng trực quan như màu sắc, kết cấu qua kỹ thuật BoVW hay học sâu bị hạn chế trong việc hiểu nội dung do khoảng cách ngữ nghĩa giữa đặc trưng mức thấp và khái niệm con người mong đợi [1, 2]. Nghiên cứu đề xuất xây dựng ontology cho tập ảnh đa nhãn kết hợp chú thích hình ảnh, chuyển dữ liệu phi cấu trúc thành tri thức có cấu trúc, hỗ trợ truy vấn ngữ nghĩa (SBIR) hiệu quả hơn. Đồng thời, những nghiên cứu vượt trội trong xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) [3] và các phương

pháp xây dựng ontology đã mở ra khả năng ứng dụng các phương pháp SBIR nhằm cải thiện hiệu suất truy vấn ảnh theo ngữ nghĩa.

Trong bài báo này, một phương pháp đề xuất xây dựng ontology cho tập ảnh đa nhãn, kết hợp mô tả ngữ nghĩa ảnh đầu vào để xây dựng hệ truy vấn ảnh theo ngữ nghĩa trên bộ dữ liệu Flickr30k được đánh giá là hiệu quả, khả thi. Kết hợp các mô hình học máy đã triển khai hệ thống phù hợp đáp ứng nhu cầu người dùng. Trong các công trình trước đây của nhóm tác giả [4, 5] chỉ xây dựng ontology cho tập ảnh đơn đối tượng, vì vậy tập ảnh đa đối tượng có nhiều cải tiến cần đề xuất và hoàn thiện.

Đóng góp của công trình này là (1) xây dựng một khung ontology cho tập ảnh đa nhãn, mô tả ngữ nghĩa và mối quan hệ giữa các đối tượng trên ảnh bởi ontology; (2) kết hợp mô hình học máy và xử lý ngôn ngữ tự nhiên cho phân tích ngữ nghĩa mô tả bằng text cho ảnh đầu vào; (3) sinh câu truy vấn SPARQL từ mô tả ảnh để truy vấn trên ontology; (4) thực nghiệm trên bộ ảnh Flickr30k [6] hiệu quả cao hơn một số công trình đã công bố được chọn làm so sánh trên cùng bộ dữ liệu.

Phần còn lại của bài báo gồm: **Phần 2** khảo sát các công trình nghiên cứu liên quan; **Phần 3** trình bày phương pháp xây dựng ontology cho tập ảnh đa nhãn; **Phần 4** xây dựng thực nghiệm và đánh giá kết quả; kết luận và hướng phát triển được trình bày trong **Phần 5**

2. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Zakaria hamane và cộng sự (2023) [7] đã tiến hành đánh giá tài liệu có hệ thống về việc so khớp ontology bằng các phương pháp học sâu, bao gồm các bài báo được xuất bản từ năm 2005 đến năm 2022. Phương pháp này để phân loại văn bản tài liệu và chi tiết các bài báo dựa trên phương pháp học sâu được sử dụng ontology, điều này cho thấy phân loại văn bản bằng ontology là một phương pháp được đánh giá hiệu quả, khả thi. Từ đó có thể áp dụng cho việc xây dựng ontology cho mô tả ảnh bằng văn bản nhằm áp dụng cho bài toán tìm kiếm ảnh dựa vào văn bản.

Haoyu Liu và cộng sự (2024) [8] đã đề xuất một phương pháp tìm kiếm ảnh dựa trên text cho tập ảnh đa đối tượng. Trong công trình này, tác giả đã xây dựng một tập dữ liệu thử thách truy vấn nhỏ gọn và phân mảnh mới (có tên là Flickr30K-CFQ) để mô hình hóa tác vụ truy xuất văn bản-hình ảnh khi xem xét nhiều nội dung và phong cách truy vấn, bao gồm ngữ liệu quan hệ thực thể nhỏ gọn và chi tiết. Nhóm tác giả cũng đề xuất một phương pháp truy vấn nâng cao dựa trên LLM sử dụng kỹ thuật nhắc nhở dựa trên LLM. Các thí nghiệm cho thấy bộ dữ liệu Flickr30-CFQ do nhóm tác giả đề xuất có sự thiếu hụt của các tập dữ liệu ngôn ngữ thị giác hiện có trong các mô tả văn bản-hình ảnh thực tế. Phương pháp Query-enhanced dựa trên LLM được áp dụng trên các mô hình truy xuất văn bản-hình ảnh hiện có khác nhau giúp cải thiện hiệu suất hiểu truy vấn trên cả bộ dữ liệu công khai Flickr30k và bộ thử thách Flickr30-CFQ. Kết quả thực nghiệm đã cải thiện hiệu suất tăng 0,9% và 2,4% trên các bộ dữ liệu Flickr, Flickr30-CFQ tương ứng.

Haitham Samih và cộng sự (2020) [9], bài báo này trình bày một phương pháp tiếp cận ngữ nghĩa hai giai đoạn đối với các câu truy vấn ngôn ngữ tự nhiên. Ở giai đoạn đầu tiên, phân tích cú pháp Stanford và một công cụ trích xuất quan hệ dựa trên quy tắc được thiết kế được sử dụng trong quy trình trích xuất bộ ba để trích xuất hiệu quả các thuộc tính, trường hợp và mối quan hệ chủ thích ngôn ngữ tự nhiên liên quan đến các đối tượng trên ảnh đã trích xuất. Giai đoạn thứ hai tích hợp các mối quan hệ được trích xuất với nguồn ConceptNet để cung cấp ngữ nghĩa cấp cao cho các bộ ba được trích xuất. Thực nghiệm được tiến hành từ tập dữ liệu Flickr8k. Kết quả cho thấy phương pháp tiếp cận này được đề xuất thành công

trong việc trích xuất các bộ ba quan hệ với giá trị độ chính xác trung bình là 97% đối với các loại mối quan hệ chú thích khác nhau: thuộc tính và quan hệ trường hợp, quan hệ phụ thuộc nhiều từ và quan hệ ngữ nghĩa. Kết quả thực nghiệm được trình bày với TopK ($K = 1, 5, 10$) thì độ chính xác tương ứng là 91%, 96% và 98%.

Qi Zhang và cộng sự (2020) [10] đề xuất phương pháp truy xuất song hướng hình ảnh-văn bản phụ thuộc rất nhiều vào phép đo độ tương đồng và học nhúng cho từng cặp hình ảnh-văn bản. Trong công trình này, nhóm tác giả đề xuất một mạng chú ý nhận thức ngữ cảnh thống nhất (CAAN), tập trung có chọn lọc vào các đoạn cục bộ quan trọng (vùng hình ảnh và từ vựng) bằng cách tổng hợp ngữ cảnh, xem xét các tương tác giữa hình ảnh và câu trong quá trình truy xuất, các tương quan nội phương thức được suy ra từ sự chú ý vùng-từ thay vì so sánh trực quan khoảng cách giữa các đặc điểm hình ảnh gốc. Phương pháp này đạt được kết quả khá cạnh tranh trên hai tập dữ liệu truy xuất hình ảnh-văn bản chung Flickr30K và MS-COCO, đồng thời kết quả thu được đã đánh giá cao phương pháp đề xuất cho giải pháp tìm kiếm ảnh theo nhiều phương pháp kết hợp văn bản, hình ảnh với các mô hình học máy tiên tiến.

Jalila Filali và cộng sự (2020) [11] đã đề xuất một phương pháp xây dựng ontology cho ảnh đa đối tượng dựa trên phân lớp và Annotation của hình ảnh. Trong công trình này, nhóm tác giả đã khắc phục được những khuyết điểm trước kia mà phương pháp xây dựng ontology đề xuất, cụ thể những đóng góp là: (1) khai thác các mối quan hệ giữa các lớp cho quá trình phân loại hình ảnh; (2) kết hợp các đầu ra của bộ phân loại hypernym-hyponym để nâng cao hiệu suất phân lớp. Mục đích của bài báo là cải thiện phân loại hình ảnh. Một kỹ thuật đề xuất trong bài báo đã được thử nghiệm và kết quả thu được đã cải thiện phân loại hình ảnh khá tốt.

Qua một số công trình khảo sát về phương pháp tìm kiếm ảnh dựa trên text, kỹ thuật nâng cao hiệu quả tìm kiếm ảnh bằng cách sử dụng ontology, phân tích ngữ nghĩa hình ảnh bằng học sâu đã cải thiện đáng kể độ chính xác cho tìm kiếm ảnh. Trên cơ sở này, bài báo chúng tôi đề xuất phương pháp tìm kiếm ảnh dựa trên ảnh đầu vào và văn bản mô tả ảnh đầu vào.

3. PHƯƠNG PHÁP XÂY DỰNG ONTOLOGY CHO TẬP ẢNH ĐA NHÂN

3.1. Cấu trúc ontology

Ontology đóng vai trò quan trọng trong việc xây dựng một khung tri thức có cấu trúc, giúp biểu diễn dữ liệu một cách có tổ chức và hỗ trợ khả năng suy diễn tự động giữa các đối tượng liên quan và mô tả ngữ nghĩa cho hình ảnh. Trong hệ thống truy vấn hình ảnh, ontology không chỉ đơn thuần là một tập hợp dữ liệu được lưu trữ có cấu trúc, mà còn là nền tảng để hệ thống hiểu ngữ nghĩa, liên kết thông tin và trả về các truy vấn phức tạp. Trong bài báo này, một ontology đề xuất xây dựng gồm các thành phần sau:

Class (Lớp): Tập hợp các đối tượng có đặc điểm chung thuộc về một phân lớp trong phân lớp ảnh, Dog, Car, Women,

Object Property (Thuộc tính đối tượng): Mô tả thuộc tính mối quan hệ giữa các lớp, như nextTo, hasPart, ...

Data Property (Thuộc tính dữ liệu): Mô tả thuộc tính của lớp dữ liệu bởi giá trị cụ thể như color, size,

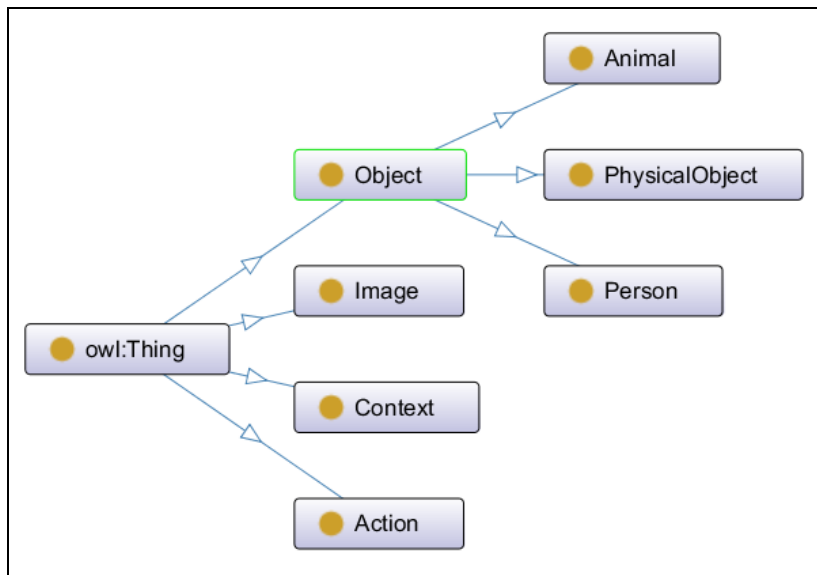
Individual (Cá thể): Thực thể cụ thể thuộc một lớp, đây chính là đối tượng chính của hình ảnh cần xây dựng cho ontology như Fido, Eiffel Tower, ...

Axiom (Tiên đề): Quy tắc logic xác định ràng buộc, chẳng hạn Dog là subclass của Animal.

Ontology tổ chức các khái niệm liên quan đến hình ảnh (như đối tượng, màu sắc, vị trí) và mối quan hệ giữa chúng, giúp hệ thống hiểu rõ ngữ cảnh của hình ảnh. Vì vậy, ontology giúp hệ thống có thể trả lời các câu hỏi phức tạp như "tìm hình ảnh có chó nâu cạnh xe đỏ" bằng cách suy luận logic dựa trên mối quan hệ và thuộc tính.

3.2. Các thành phần của Ontology

Một thành phần quan trọng cho toàn bộ hệ thống truy vấn ảnh theo ngữ nghĩa là ontology. Ontology là khung tri thức để xác định ý nghĩa của tập dữ liệu nên cần phải xây dựng với cấu trúc phù hợp. Khi thực hiện phân tích, tìm hiểu về tập dữ liệu Flickr30k cho thấy đây là tập hình ảnh khá phong phú về chủ đề con người, động vật, hoạt động xã hội, cảnh vật, Do đó, trong bài báo này cấu trúc ontology được đề xuất với cấu trúc phân lớp 'cha – con' như sau:



Hình 1. Cấu trúc phân lớp trong ontology bộ ảnh Flickr30k

Hình 1 mô tả cấu trúc lớp cha, lớp con và mối quan hệ giữa các phân lớp thuộc bộ ảnh Flickr30k gồm:

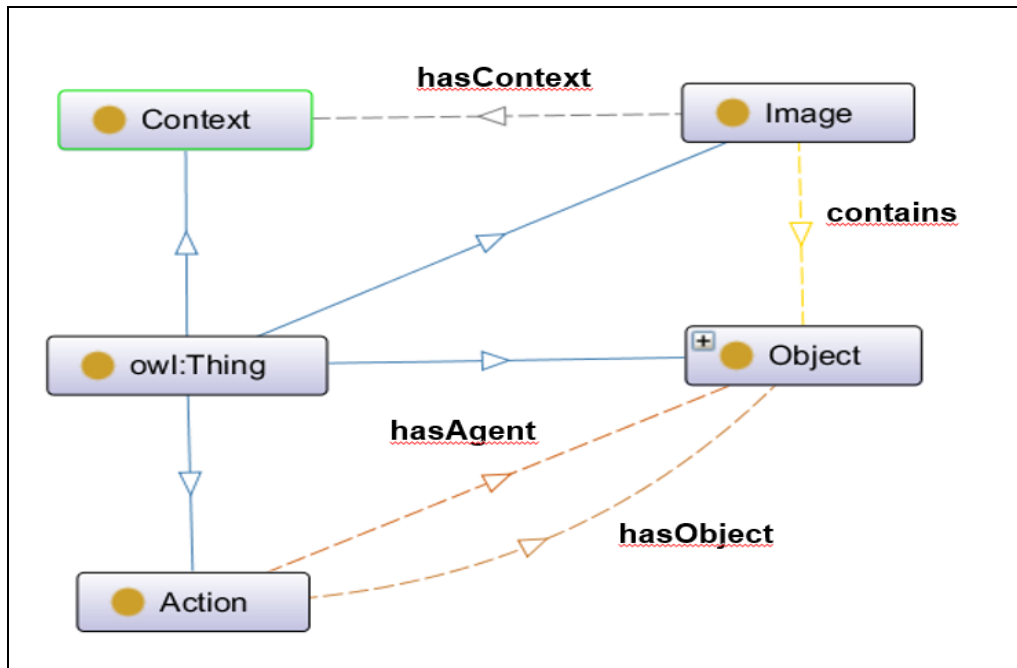
Image: Đại diện cho các ảnh trong tập Flickr30k, là thực thể trung tâm cần mô tả.

Object: Lớp cha chung cho các đối tượng xuất hiện trong ảnh, bao gồm: **Person** mô tả con người, phù hợp với chủ đề "con người". **Animal** mô tả động vật, phù hợp với chủ đề "động vật". **PhysicalObject** mô tả các vật thể vật lý khác (như đồ vật, công cụ, ...) hỗ trợ mô tả các đối tượng không phụ thuộc về Person hoặc Animal.

Action: Đại diện cho các hành động hoặc hoạt động xã hội (chạy, nhảy, giao tiếp...)

Context: Mô tả bối cảnh của ảnh (thời tiết, vị trí) phù hợp với chủ đề "cảnh vật".

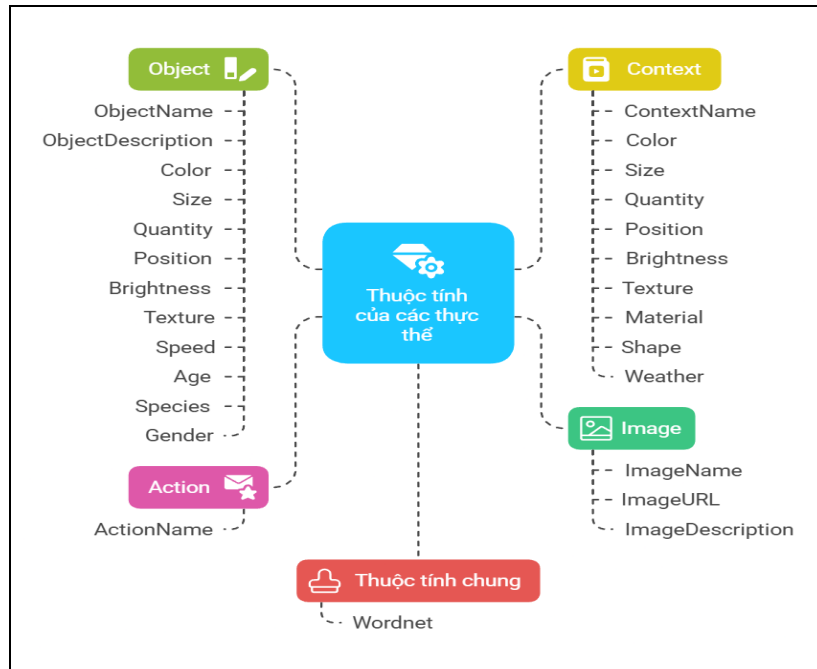
Bên cạnh xác định các lớp cha, lớp con việc xây dựng thuộc tính đối tượng (Object property) và các thuộc tính của lớp đó cũng phải phù hợp, mang tính bao hàm toàn bộ tập ảnh. Hình 2 mô tả mối quan hệ giữa Image – Object (contains), Image – Context (hasContext), Object – Action (hasAgent, hasObject)... và nhiều mối quan hệ khác như quan hệ giữa Object – Object, Image – Image.



Hình 2. Minh họa các mối quan hệ trong ontology được xây dựng

Các thuộc tính liên quan của các thành phần cần lưu trữ cho mô tả ontology gồm:

- Image:** ImageName, ImageURL, ImageDescription cung cấp thông tin cơ bản và mô tả tổng quát về ảnh.
- Object:** ObjectName, ObjectDescription, Age, Color, Size, Position, v.v. mô tả chi tiết các đặc điểm của con người, động vật, hoặc vật thể.
- Person/Animal:** Gender, Emotion, Species, phù hợp để mô tả đặc trưng của con người và động vật.
- Action:** ActionName, Wordnet, hỗ trợ định danh hành động và liên kết với từ điển ngữ nghĩa như WordNet.
- Context:** ContextName, Brightness, Weather, Material, v.v. mô tả bối cảnh môi trường của ảnh.



Hình 3. Thuộc tính class trong ontology

Hình 3 mô tả các thành phần mô tả thuộc tính cho mỗi class trong ontology như thuộc tính cho ảnh, thuộc tính Action, thuộc tính Object và Context. Các thuộc tính này là cơ sở cho việc xác định thành phần mối quan hệ trong ontology cần xây dựng.

3.3. Phương pháp xây dựng ontology

Hiện nay có nhiều phương pháp tổ chức dữ liệu và xây dựng ontology cho hình ảnh, tuy nhiên để mang lại hiệu quả truy xuất ngữ nghĩa cho tập ảnh đa nhãn từ ontology thì một phương pháp đề xuất xây dựng ontology gồm các bước được đề xuất như sau:

Bước 1: Thu thập và tiền xử lý dữ liệu

- (1) Đọc dữ liệu từ file CSV bằng pandas, chia thành các đoạn (chunks) để tối ưu bộ nhớ.
- (2) Tiền xử lý: Chuẩn hóa caption: Chuyển về chữ thường, loại bỏ stopwords (dùng NLTK). Dùng spaCy ("en_core_web_sm") để thực hiện POS Tagging và Dependency Parsing, trích xuất thực thể: Danh từ (NOUN, PROPN). Thuộc tính: Tính từ, số (ADJ, NUM). Quan hệ: Động từ, giới từ (VERB, ADP).
- (3) Kết quả: Dữ liệu văn bản đã xử lý, chứa thực thể, thuộc tính, quan hệ.

Bước 2: Phân loại và hợp nhất thông tin

- (1) Phân loại: Dùng SentenceTransformer ("all-MiniLM-L6-v2") và FAISS: Thực thể mã hóa thành vector, so sánh với từ điển danh từ để gán nhãn (Person, Object, Context). Thuộc tính: So sánh với từ điển tính từ (Color, Size, Quantity). Quan hệ: Phân loại thành "action" hoặc tạo thực thể mới (PhysicalObject).

- (2) Hợp nhất: Tổng hợp dữ liệu từ 5 caption, loại bỏ trùng lặp.
- (3) Kết quả: Dữ liệu phân loại (entities, attributes, relations) đã hợp nhất.

Bước 3: Xây dựng và lưu trữ ontology

- (1) Cấu trúc lớp (Class):

Image (Đại diện ảnh); Object (Person, Animal, PhysicalObject); Action (Hành động chạy, nhảy); Context (Bối cảnh thời tiết, vị trí).

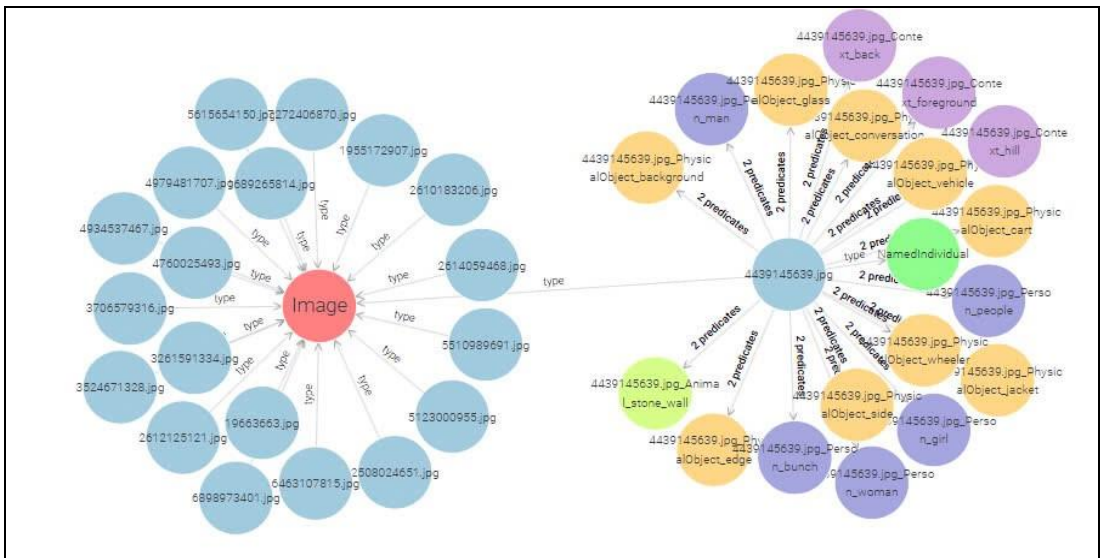
- (2) Mỗi quan hệ (Object Property): hasObject, performsAction, locatedIn.

Thuộc tính (Data Property) gồm: Image (ImageName, ImageURL); Object (ObjectName, Color, Size); Person/Animal (Gender, Species); Action (ActionName); Context (ContextName, Weather).

- (3) Lưu trữ: Dùng owlready2 để xây dựng Ontology, lưu trên GraphDB.

- (4) Kết quả: Ontology hoàn chỉnh, sẵn sàng cho truy vấn

Từ các bước đề xuất xây dựng ontology cho tập ảnh đa nhân Flickr30k, trong công trình này chúng tôi đã nghiên cứu xây dựng thành công một ontology hoàn chỉnh, tạo ra khung tri thức có cấu trúc rõ ràng. Khung này gồm các thực thể, thuộc tính và mối quan hệ được trích xuất, sắp xếp gọn gàng, thể hiện đầy đủ thông tin ngữ nghĩa. Chất lượng được kiểm tra qua độ chính xác, thời gian xử lý và sự đồng nhất, cho thấy ontology vừa đầy đủ vừa đáng tin, hỗ trợ tốt cho truy vấn hình ảnh dựa trên ngữ nghĩa. Cấu trúc ontology xây dựng được minh họa như hình 4.



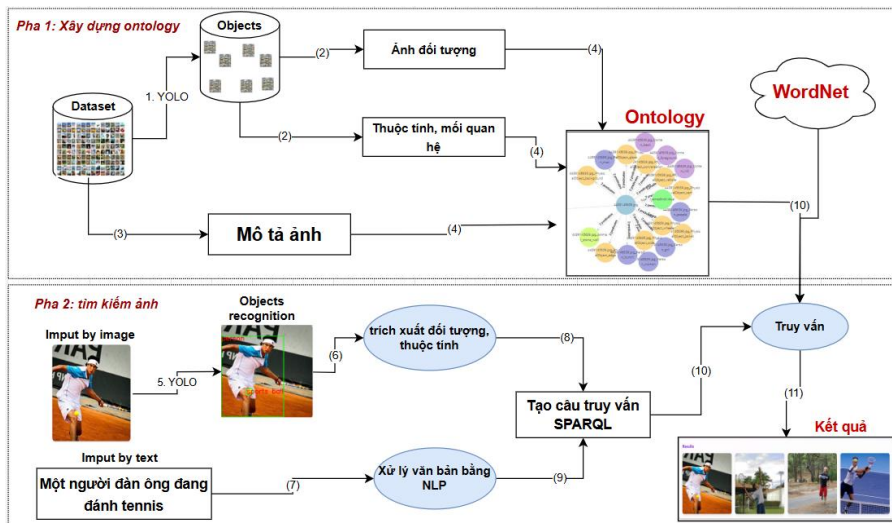
Hình 4. Minh hoạ ontology cho tập ảnh Flickr30k

Hình 4 minh họa một ontology cho tập ảnh đa nhãn Flickr30k trên GraphDB, trong đó mỗi node trên ontology là một ảnh, mỗi ảnh có nhiều thuộc tính, mỗi quan hệ giữa các node, thuộc tính node. Mỗi node có một nhãn (label), các node cùng label thì cùng định dạng. Dựa trên ontology đã xây dựng, thuật toán truy vấn ảnh liên quan đến ảnh đầu vào hoặc mô tả ảnh đầu vào được trích xuất bằng câu truy vấn SPARQL.

4. MÔ HÌNH ĐỀ XUẤT

Để tìm kiếm những ảnh tương tự như mô tả ảnh hoặc ảnh đầu vào gồm những đối tượng và mô tả thuộc tính đối tượng, mô hình tìm kiếm ảnh được đề xuất (Hình 5) với các bước như sau:

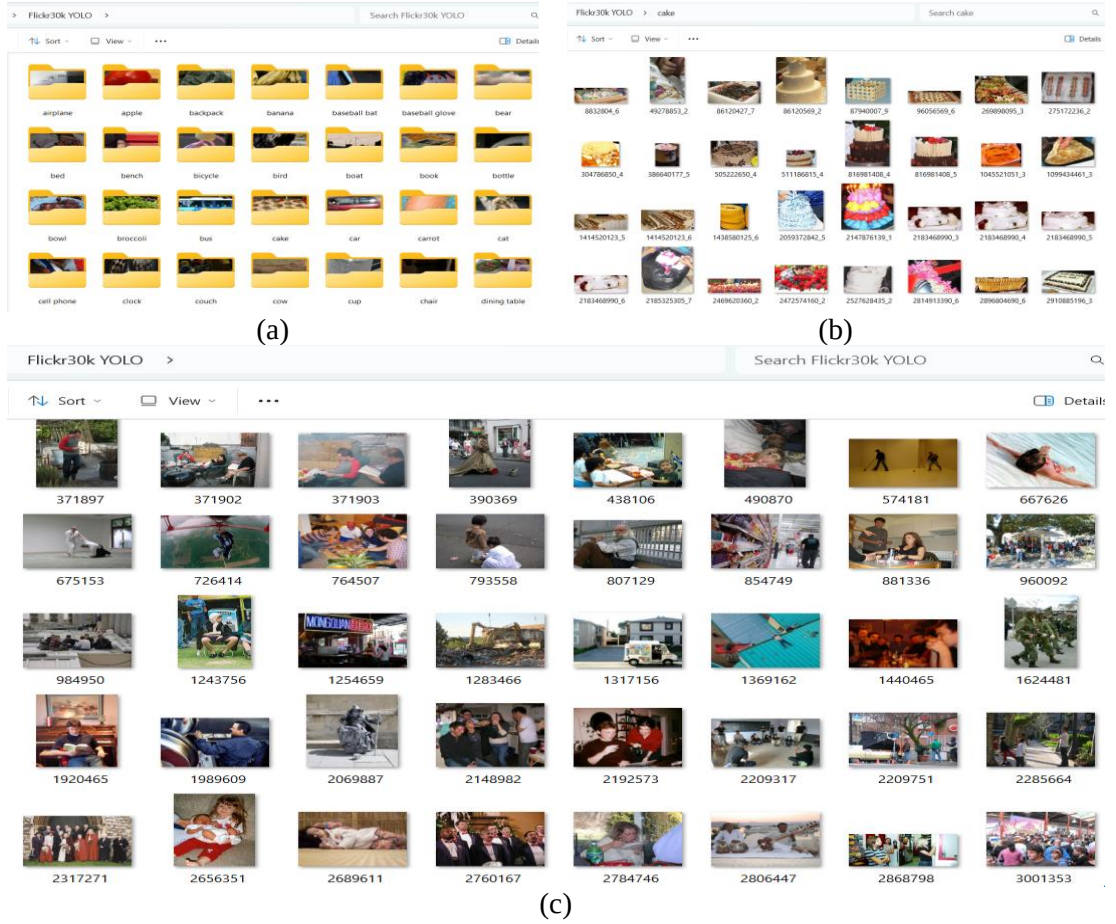
- (1) Phân đoạn và nhận diện đối tượng trên tập ảnh thực nghiệm bằng mô hình học sâu YOLO;
- (2) Trích xuất từng đối tượng, thuộc tính và mối quan hệ giữa các đối tượng;
- (3) Trích xuất mô tả ảnh (annotation) cho tập ảnh thực nghiệm, làm căn cứ xây dựng ontology;
- (4) Xây dựng ontology dựa trên đối tượng, thuộc tính và mối quan hệ giữa các đối tượng;
- (5) Mỗi ảnh đầu vào, phân đoạn ảnh bằng mô hình YOLO;
- (6) Trích xuất đối tượng, thuộc tính cho ảnh đầu vào;
- (7) Xử lý văn bản mô tả ảnh đầu vào;
- (8) Xây dựng câu truy vấn SPARQL từ ảnh input;
- (9) Xây dựng câu truy vấn từ mô tả ảnh;
- (10) Thực hiện truy vấn trên ontology kết hợp WordNet;
- (11) Trích xuất tập ảnh tương tự với ảnh đầu vào hoặc mô tả ảnh đầu vào bằng text.



Hình 5. Mô hình tìm kiếm ảnh dựa trên ontology và học sâu

5. THỰC NGHIỆM

5.1. Mô tả bộ dữ liệu thực nghiệm



Hình 6. Tập ảnh Flickr30k

Flickr30k là bộ dữ liệu chuẩn để đánh giá mô hình tạo mô tả ảnh và truy xuất thông tin đa phương thức, nhằm nghiên cứu mối liên hệ giữa ảnh và văn bản qua các tác vụ như sinh mô tả, tìm ảnh từ văn bản, và liên kết từ với vùng ảnh. Bộ ảnh gồm 31,783 ảnh từ Flickr, đa dạng về chủ đề (con người, động vật, đồ vật, cảnh quan), kèm 158,915 câu mô tả (5 câu/ảnh) do người dùng Amazon Mechanical Turk viết, tập trung vào đối tượng chính và ngữ cảnh. Chi tiết bộ ảnh Flickr30k được minh họa như Hình 6, trong đó ảnh sau khi phân đoạn và phân lớp bằng YOLO (a), ảnh sau khi phân đoạn theo đối tượng bằng mạng học sâu YOLO (b) gốc ban đầu chưa tính và tập ảnh ban đầu chưa phân đoạn (c).

Tập dữ liệu Flickr30k các chủ đề ảnh đa dạng, bao gồm ảnh chứa đối tượng là con người (40%), chứa động vật động vật (15%), chứa đồ vật đồ vật (25%), ảnh phong cảnh (20%). Độ dài câu mô tả ảnh trung bình 10–15 từ/câu, kết hợp mô tả vật thể, hành động và quan hệ không gian. Phần mô tả ảnh được sử dụng cho quá trình trích xuất thông tin mô tả hình ảnh.

5.2. Kết quả thực nghiệm

Ontology được xây dựng từ dữ liệu Flickr30K, chia thành các lớp như Image, Object, Action, lưu dưới dạng OWL và triển khai trên GraphDB để hỗ trợ truy vấn SPARQL hiệu quả, cho phép quản lý và suy diễn ngữ nghĩa.

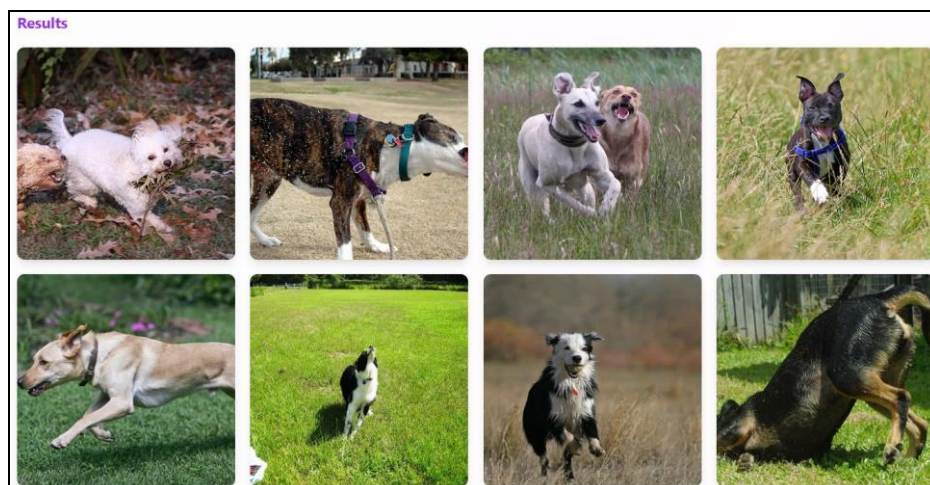
Ứng dụng dùng FastAPI, hỗ trợ hai cách truy vấn hình ảnh liên quan với dữ liệu input:

Văn bản (text): Người dùng nhập mô tả, trường hợp này minh hoạ mô tả ảnh bằng text là “a dog is running on the grass”, hệ thống phân tích câu, phân loại từ điển, chuyển thành câu truy vấn SPARQL. Quá trình này mất trung bình 55ms để xử lý và truy xuất những ảnh phù hợp với mô tả text.

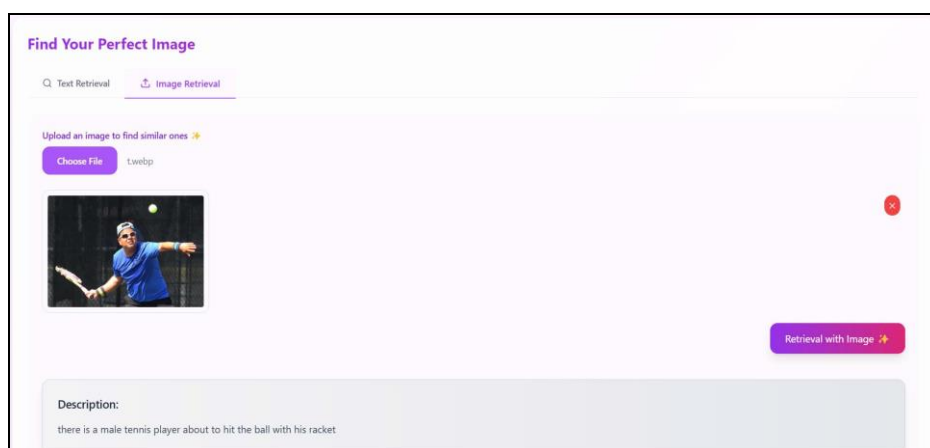
Hình ảnh (input image): Người dùng tải ảnh đầu vào, hệ thống tự tạo mô tả, sinh câu truy vấn SPARQL. Toàn bộ quy trình từ xử lý đến truy xuất kết quả mất trung bình 6 giây và kết quả hiển thị ảnh tương đồng với ảnh input.

Hình 7. Giao diện truy vấn input là text

Hình 7, mô tả giao diện truy vấn ảnh dựa trên dữ liệu đầu vào là text mô tả ảnh cần tìm. Trong giao diện này minh hoạ là ‘a dog is running on the grass’, như vậy hệ thống xử lý văn bản tìm ra các đối tượng là ‘dog’, ‘grass’, sinh câu truy vấn SPARQL để tìm hững hình ảnh thoả điều kiện này. Thời gian xử lý truy vấn văn bản trung bình là 55ms và được kết quả trả về như Hình 8.



Hình 8. Giao diện hiển thị kết quả trả về input là text



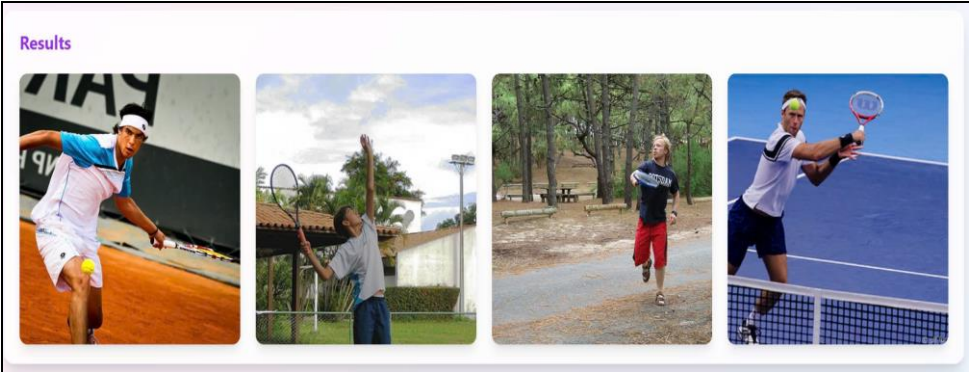
Hình 9. Giao diện hiển thị image retrieval input là ảnh

Hình 9, mô tả giao diện truy vấn ảnh dựa trên dữ liệu đầu vào là ảnh. Trong giao diện này truyền vào một ảnh, như vậy hệ thống sử dụng học sâu trích xuất đối tượng (man, sport ball) sinh câu truy vấn SPARQL (Hình 7) dựa trên các đối tượng nhận diện được để tìm hững hình ảnh thoả điều kiện. Quá trình này mất khoảng 5-6 giây, tùy thuộc vào độ phức tạp của ảnh. kết quả trả về như Hình 10.

Generated SPARQL Query:

```
1 PREFIX : <https://www.semanticweb.org/asus/ontologies/2025/1/untilted-ontology-26#> SELECT DISTINCT ?image ?url WHERE {
2   ?PhysicalObjectName_racket a :PhysicalObject.
3   ?image :contains ?PhysicalObjectName_racket.
4   VALUES ?synonym_PhysicalObjectName_racket { "racquet" "dissonance" "noise" "racket" }
5   OPTIONAL { ?PhysicalObjectName_racket :wordnet ?synonym_PhysicalObjectName_racket. }
6   ?PhysicalObjectName_racket :Objectname ?synonym_PhysicalObjectName_racket.
7   ?image a :Image;
8     :ImageURL ?url.
9   OPTIONAL {
10    ?PhysicalObjectName_ball a :PhysicalObject.
11    ?image :contains ?PhysicalObjectName_ball.
12    VALUES ?synonym_PhysicalObjectName_ball { "lump" "testis" "glob" "orb" "globe" "bollock" "egg" "clod" "chunk" "musket ball" "ball" }
13    OPTIONAL { ?PhysicalObjectName_ball :wordnet ?synonym_PhysicalObjectName_ball. }
14    OPTIONAL { ?PhysicalObjectName_ball :Objectname ?synonym_PhysicalObjectName_ball. }
15    ?PhysicalObjectName_tennis_player a :PhysicalObject.
16    ?image :contains ?PhysicalObjectName_tennis_player.
17    VALUES ?synonym_PhysicalObjectName_tennis_player { "tennis player" }
18    OPTIONAL { ?PhysicalObjectName_tennis_player :wordnet ?synonym_PhysicalObjectName_tennis_player. }
19    OPTIONAL { ?PhysicalObjectName_tennis_player :Objectname ?synonym_PhysicalObjectName_tennis_player. }
20    VALUES ?synonym_Age_male { "male" }
21    OPTIONAL { ?PhysicalObjectName_tennis_player :Age ?synonym_Age_male. }
22  }
23 }
```

Hình 10. Giao diện mô tả và câu truy vấn liên quan đến hình ảnh



Hình 11. Giao diện hiển thị kết quả trả về input là ảnh

5.3. So sánh kết quả thực nghiệm

Để minh chứng cho kết quả thực nghiệm mô hình đề xuất tìm kiếm ảnh đa nhãn, kết quả thực nghiệm trong công trình này được tiến hành so sánh với một số kết quả thực nghiệm trên cùng bộ dữ liệu được trình bày trong Bảng 1 như sau:

Bảng 1. So sánh kết quả thực nghiệm trên bộ ảnh Flcikr30k với các phương pháp khác

Phương pháp	MAP
IRSGS-GCN, (Yoon, S.), 2021 [11]	0,5670
CAMP, (Wang Z, et al.), 2019 [12]	0,7710
BGAN- 48 bit, (Song et al., 2018) [13]	0,7030
SR-KDT (Re KD-Tree) [14]	0,7794
IT_SBOD (by input image)	0,7925
IT_SBOD (by input text)	0,7834

Kết quả so sánh độ chính xác tìm kiếm ảnh trên bộ dữ liệu Flickr30k với một số công trình khác cho thấy, hệ truy vấn ảnh đa nhãn IT-SBOD cao hơn một số công trình được chọn làm so sánh vì các lý do:

1. Hệ truy vấn ảnh đa nhãn ảnh IT-SBOD kết hợp được đặc trưng cấp cao được tích hợp từ ontology, điều này giúp phân tích ngữ nghĩa hình ảnh chính xác, làm tăng kết quả phân tích ngữ nghĩa từ nội dung hình ảnh, đây là một yếu tố giúp cải thiện độ chính xác tìm kiếm ảnh theo ngữ nghĩa.
2. Hệ truy vấn ảnh đa nhãn ảnh IT-SBOD sử dụng mạng học sâu YOLOv8, đây là phiên bản khá hoàn thiện với độ chính xác nhận diện đối tượng cao, phân vùng đối tượng khá chính xác. Điều này góp phần tăng hiệu suất nhận diện tối tưng trên ảnh đa nhãn.
3. Hệ truy vấn ảnh đa nhãn ảnh IT-SBOD sử dụng mô hình BLIP giúp quá trình chuyển ảnh sang text khá chính xác, từ đó xây dựng câu truy vấn SPARQL cho tìm kiếm trên ontology đã xây dựng.
4. Hệ truy vấn ảnh đa nhãn ảnh IT-SBOD kết hợp xử lý ngôn ngữ tự nhiên, đáp ứng nhu cầu đa chức năng của người dùng trong bối cảnh hiện nay, đây là một công cụ ứng dụng mô hình hình ngôn ngữ lớn trong AI được nhiều nhóm nghiên cứu quan tâm.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu này đã xây dựng thành công ontology cho tập ảnh đa nhãn Flickr30k, hỗ trợ truy vấn ảnh theo ngữ nghĩa hiệu quả hơn tìm kiếm ảnh theo nội dung nhờ kết hợp mô hình xử lý ngôn ngữ tự nhiên (NLP) hiệu quả và ontology xây dựng trên GraphDB tổ chức lưu trữ dữ liệu. Hệ thống cho phép tìm kiếm linh hoạt bằng văn bản và ảnh đầu cho dữ liệu đầu vào, có ứng dụng thực tiễn cao. Hướng phát triển sẽ cải thiện mô hình sinh mô tả bằng học sâu và mở rộng quy mô dữ liệu để nâng cao hiệu suất cả về độ chính xác và thời gian đáp ứng nhu cầu người dùng. Xây dựng khung ontology có cấu trúc hiệu quả, bài báo đã thành công trong việc xây dựng một ontology có tổ chức, với các lớp chính như Image, Object (Person, Animal, PhysicalObject), Action, và Context, cùng các mối quan hệ (hasObject, performsAction) và thuộc tính (Color, Size, Gender). Ontology được lưu trên GraphDB, hỗ trợ câu truy vấn SPARQL và suy diễn tự động, giúp hệ thống hiểu ngữ nghĩa và trả lời các truy vấn phức tạp như “a dog is running on the grass” một cách chính xác (Phần 3.1.4). Truy vấn ngữ nghĩa chính xác và linh hoạt: Hệ thống SBIR cải thiện độ chính xác so với CBIR nhờ tích hợp NLP (POS Tagging, Dependency Parsing) và Ontology. Hệ thống hỗ trợ cả truy vấn bằng mô tả và hình ảnh, với giao diện FastAPI thân thiện. Việc ánh xạ truy vấn tự nhiên thành SPARQL và sử dụng WordNet, SBERT để mở rộng ngữ nghĩa đã tăng khả năng khớp và độ bao quát của kết quả.

TÀI LIỆU THAM KHẢO

1. Li, X., Yang, J., & Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). *Neurocomputing*, 452, 675-689.
2. Sunitha, T., & Sivarani, T. S. (2022). Novel content based medical image retrieval based on BoVW classification method. *Biomedical Signal Processing and Control*, 77, 103678.

3. Ashok Kumar, P. M., Subha Mastan Rao, T., Arun Raj, L., & Pugazhendi, E. (2021). An Efficient Text-Based Image Retrieval Using Natural Language Processing (NLP) Techniques. In *Intelligent System Design: Proceedings of Intelligent System Design: INDIA 2019* (pp. 505-519). Springer Singapore.
4. Sain, A., Bhunia, A. K., Koley, S., Chowdhury, P. N., Chattopadhyay, S., Xiang, T., & Song, Y. Z. (2023). Exploiting unlabelled photos for stronger fine-grained SBIR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6873-6883).
5. Saavedra, J. M., Morales, J., & Murrugarra-Llerena, N. (2023). SBIR-BYOL: a self-supervised sketch-based image retrieval model. *Neural Computing and Applications*, 35(7), 5395-5408.
6. Sankesara, H. Flickr Dataset. Available at <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset>. Accessed 10 April 2024.
7. Hamane, Z., Samih, A., & Fennan, A. (2023). ONTOLOGY MATCHING USING DEEP LEARNING. *Journal of Theoretical and Applied Information Technology*, 101(10).
8. Liu, H., Song, Y., Wang, X., Zhu, X., Li, Z., Song, W., & Li, T. (2024, July). Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval. In *International Conference on Database Systems for Advanced Applications* (pp. 419-434). Singapore: Springer Nature Singapore.
9. Samih, H., Rady, S., Ismail, M., & Gharib, T. (2021). * Improving Natural Language Queries Search and Retrieval through Semantic Image Annotation Understanding. *International Journal of Intelligent Computing and Information Sciences*, 20(2), 67-78.
10. Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3536-3545).
11. Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., & Kim, E. S. (2021). Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence* 35(12) (pp. 10718-10726).
12. Z W., et al. (2019). Camp cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5764-5773).
13. Song, J., et al. (2018). Binary generative adversarial networks for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
14. Dinh, Nguyen Thi, Thanh The Van, & Thanh Manh Le. (2022). Semantic relationship-based image retrieval using KD-tree structure. In *Asian Conference on Intelligent Information and Database Systems* (pp. X-Y). Cham: Springer International Publishing.
15. Huang, J., Wang, Z., Liang, H., Lin, L., Yu, B., Dong, F., & Xu, Y. Accepted Manuscript *International Journal of Pattern Recognition and Artificial Intelligence*.
16. Jabnoun, J., Haffar, N., Zrigui, A., Nsir, S., Nicolas, H., & Trigui, A. (2022, September). An image retrieval system using deep learning to extract high-level

- features. In *International Conference on Computational Collective Intelligence* (pp. 167-179). Cham: Springer International Publishing.
17. Gkelios, S., Boutalis, Y., & Chatzichristofis, S. A. (2021, July). Investigating the vision transformer model for image retrieval tasks. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)* (pp. 367-373). IEEE.
18. Viridiano, M., Lorenzi, A., Torrent, T. T., Matos, E. E., Pagano, A. S., Sigiliano, N. S., & de Freitas, M. H. P. (2024, May). Framed Multi30K: A frame-based multimodal-multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 7438-7449).
19. Dost, S., Serafini, L., Rospocher, M., Ballan, L., & Sperduti, A. (2020). Vt-linker: Visual-textual-knowledge entity linker. In *ECAI 2020* (pp. 2897-2898). IOS Press.

ABSTRACT

IMAGE RETRIEVAL USING ONTOLOGY COMBINED WITH A DEEP LEARNING MODEL

Vo Thai Binh¹, Pham Minh Thuan¹, Nguyen Tran Duy Khanh¹, Nguyen Thi Dinh^{1,*}

¹*Ho Chi Minh City University of Industry and Trade*

*Email: dinght@huit.edu.vn

The problem of semantic-based image retrieval is increasingly applied in many fields, such as education, healthcare, tourism, and environmental biology, and implemented by many different methods, such as using machine learning models and combining transformer models with many different types of data. In this paper, an ontology structure is built for a multi-label image set and applied to an image retrieval system according to the proposed semantic description. To implement this solution, each multi-label image is segmented into single-label images using the YOLO deep learning network. The proposed method exploits semantic information from captions accompanying the image to build a structured knowledge framework (ontology). This process includes preprocessing text data, applying natural language processing techniques such as POS tagging and dependency parsing to extract semantic components, and classifying words into groups of entities, attributes, and relationships between entities. The obtained information is then organized into an ontology and stored in GraphDB. The system allows querying in two forms: natural language description and through images. Experimental results on the Flickr30k dataset show that the proposed method significantly improves the accuracy and efficiency of semantic image querying, opening up a new direction in the field of image querying based on semantic knowledge.

Keywords: Image retrieval, semantics of image, Ontology, NLP, Deep learning.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Võ Thái Bình

Số điện thoại liên hệ: 0772587205

TỐI ƯU HÓA CHI PHÍ VÀ THỜI GIAN VẬN CHUYỂN HÀNG HÓA BẰNG THUẬT TOÁN MULTI-VERSE OPTIMIZATION

Nguyễn Thị Thùy Trang¹, Phan Thị Ngọc Mai^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: maiphantn@huit.edu.vn

Ngày nhận bài: 12/3/2025 ; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Bài báo này đề xuất ứng dụng thuật toán Multi-Verse Optimization (MVO) để tối ưu hóa chi phí và thời gian vận chuyển hàng hóa trong chuỗi cung ứng. Mô hình thử nghiệm được xây dựng với dữ liệu bao gồm nhà cung cấp, điểm tiêu thụ, phương tiện vận chuyển và ma trận chi phí – thời gian. Thuật toán MVO được triển khai và so sánh với các phương pháp khác như Genetic Algorithm (GA) và Particle Swarm Optimization (PSO). Kết quả thực nghiệm cho thấy MVO giúp giảm chi phí vận chuyển, cải thiện thời gian giao hàng và giảm vi phạm ràng buộc. Ngoài ra, MVO có tốc độ hội tụ nhanh hơn so với GA và PSO, giúp tối ưu hóa tuyến đường và lựa chọn phương tiện hiệu quả hơn. Phương pháp này có khả năng thích ứng với sự thay đổi trong nhu cầu và điều kiện chuỗi cung ứng. Kết quả cho thấy MVO có tiềm năng ứng dụng cao trong logistics, giúp doanh nghiệp giảm chi phí vận hành, nâng cao hiệu suất và đảm bảo chất lượng dịch vụ.

Từ khóa: Multi-Verse Optimization, tối ưu chuỗi cung ứng, vận chuyển hàng hóa, tối ưu chi phí, logistics

1. GIỚI THIỆU

Trong bối cảnh toàn cầu hóa và sự phát triển nhanh chóng của thương mại điện tử, tối ưu hóa chuỗi cung ứng đóng vai trò quan trọng trong việc nâng cao hiệu quả vận chuyển hàng hóa. Các doanh nghiệp luôn tìm kiếm giải pháp để giảm chi phí vận chuyển mà vẫn đảm bảo thời gian giao hàng, đặc biệt khi đối mặt với những thách thức như biến động nhu cầu, chi phí nhiên liệu tăng, và các ràng buộc về thời gian giao hàng. Để giải quyết vấn đề này, các phương pháp tối ưu hóa thông minh đã được áp dụng nhằm tìm ra các chiến lược vận chuyển hiệu quả nhất.

Nhiều nghiên cứu đã tập trung vào tối ưu hóa vận chuyển hàng hóa bằng các thuật toán metaheuristic. GA là một trong những phương pháp được sử dụng phổ biến trong tối ưu hóa logistics nhờ khả năng tìm kiếm không gian giải pháp rộng [1]. Tuy nhiên, GA có thể gặp vấn đề về hội tụ chậm và dễ rơi vào cực tiểu cục bộ khi giải quyết các bài toán lớn. PSO [2] cũng

được áp dụng trong lập kế hoạch vận chuyển, giúp cải thiện hiệu suất so với GA nhưng vẫn có hạn chế về tốc độ hội tụ khi tối ưu trong không gian lớn.

Một số nghiên cứu khác đã sử dụng Ant Colony Optimization (ACO) để tối ưu hóa tuyến đường vận chuyển [3], giúp tìm ra lộ trình ngắn nhất giữa các điểm tiêu thụ. Tuy nhiên, phương pháp này đòi hỏi chi phí tính toán cao và có thể không phù hợp với các hệ thống chuỗi cung ứng quy mô lớn. Bên cạnh đó, thuật toán Whale Optimization Algorithm (WOA) [4] cũng đã được thử nghiệm trong lĩnh vực logistics nhưng vẫn chưa đạt được hiệu quả cao khi cân bằng giữa chi phí và thời gian vận chuyển.

Gần đây, thuật toán MVO [5] đã thu hút sự quan tâm do khả năng tối ưu hóa mạnh mẽ, tốc độ hội tụ nhanh và hiệu suất cao trong các bài toán phức tạp. MVO lấy cảm hứng từ lý thuyết đa vũ trụ, mô phỏng sự trao đổi vật chất giữa các vũ trụ để tìm ra giải pháp tốt nhất. Một số nghiên cứu đã chứng minh rằng MVO có thể đạt hiệu suất cao hơn so với GA, PSO và các thuật toán khác trong tối ưu hóa mạng lưới giao thông và logistics [6]. Tuy nhiên, vẫn chưa có nhiều nghiên cứu áp dụng MVO vào bài toán tối ưu hóa vận chuyển hàng hóa trong chuỗi cung ứng, đặc biệt trong việc tìm kiếm phương án cân bằng giữa chi phí vận chuyển và thời gian giao hàng.

Dựa trên các nghiên cứu trước đây, bài báo này đề xuất ứng dụng MVO vào bài toán lập kế hoạch vận chuyển hàng hóa nhằm giảm chi phí, tối ưu thời gian giao hàng và đảm bảo ràng buộc thực tế. Mô hình thực nghiệm sử dụng dữ liệu từ hệ thống chuỗi cung ứng với các thông số về nhà cung cấp, điểm tiêu thụ, phương tiện vận chuyển và ma trận chi phí – thời gian. Thuật toán MVO sẽ được so sánh với GA và PSO để đánh giá mức độ hiệu quả về chi phí, thời gian giao hàng và số lần vi phạm ràng buộc.

Mục tiêu của nghiên cứu là chứng minh rằng MVO có thể giúp doanh nghiệp tối ưu hóa vận chuyển hàng hóa tốt hơn so với các thuật toán khác, góp phần nâng cao hiệu suất chuỗi cung ứng, giảm thiểu chi phí vận hành và cải thiện chất lượng dịch vụ logistics.

2. TỔNG QUAN VỀ NGHIÊN CỨU

Tối ưu hóa chuỗi cung ứng là một lĩnh vực nghiên cứu quan trọng nhằm cải thiện hiệu quả vận hành, giảm chi phí và tối đa hóa tốc độ giao hàng. Nhiều thuật toán tối ưu hóa đã được áp dụng trong lĩnh vực này, bao gồm các phương pháp metaheuristic như GA, PSO, Ant Colony Optimization (ACO) và gần đây là MVO. Phần này sẽ tổng quan các nghiên cứu liên quan đến tối ưu hóa chuỗi cung ứng, đặc biệt là các thuật toán metaheuristic, và dẫn dắt vào phương pháp MVO như một cách tiếp cận mới và hiệu quả hơn.

2.1. Các phương pháp tối ưu hóa trong chuỗi cung ứng

GA là một thuật toán tiến hóa mô phỏng cơ chế chọn lọc tự nhiên. GA sử dụng các toán tử như chọn lọc, lai ghép và đột biến để tạo ra các thế hệ mới của lời giải và tiến tới tối ưu hóa [7]. GA đã được áp dụng rộng rãi trong tối ưu hóa chuỗi cung ứng, giúp tìm ra các tuyến đường vận chuyển hiệu quả nhất. Tuy nhiên, GA có một số hạn chế như dễ rơi vào cực tiểu cục bộ và tốc độ hội tụ chậm khi giải quyết các bài toán lớn.

PSO là một thuật toán dựa trên mô phỏng hành vi bầy đàn, nơi mỗi phần tử (hạt) trong quần thể di chuyển trong không gian tìm kiếm theo vị trí tốt nhất của chính nó và của bầy đàn [8]. PSO đã được chứng minh là có hiệu quả cao trong tối ưu hóa chuỗi cung ứng nhờ tốc độ hội tụ nhanh hơn GA và khả năng thích ứng với thay đổi trong dữ liệu đầu vào. Tuy nhiên,

PSO có thể bị ảnh hưởng bởi hiện tượng quá sớm hội tụ và khó khăn trong việc tìm kiếm giải pháp tối ưu toàn cục khi không gian tìm kiếm rộng lớn.

Ant Colony Optimization (ACO) dựa trên hành vi tìm kiếm đường đi của loài kiến trong tự nhiên, nơi mỗi cá thể tìm kiếm con đường tối ưu thông qua cơ chế gửi tín hiệu pheromone [9]. ACO đã được áp dụng trong nhiều bài toán tối ưu hóa chuỗi cung ứng, đặc biệt là trong tối ưu hóa tuyến đường vận chuyển. Phương pháp này có khả năng tìm kiếm tuyến đường tối ưu một cách hiệu quả, tuy nhiên đòi hỏi tài nguyên tính toán lớn và có tốc độ hội tụ chậm khi không gian tìm kiếm trở nên phức tạp.

2.2. Hạn chế của các phương pháp trước và động lực nghiên cứu

Mặc dù các thuật toán metaheuristic như GA, PSO và ACO đã được ứng dụng rộng rãi trong nhiều lĩnh vực tối ưu hóa, bao gồm cả chuỗi cung ứng, nhưng chúng vẫn tồn tại những hạn chế đáng kể cần được cân nhắc. Cụ thể, thuật toán GA có nguy cơ bị mắc kẹt trong cực tiểu cục bộ và thường cho thấy tốc độ hội tụ chậm trong các không gian tìm kiếm phi tuyến và có độ phức tạp cao [7]. Trong khi đó, thuật toán PSO mặc dù có khả năng hội tụ nhanh và hiệu quả trong không gian liên tục, nhưng lại thiếu ổn định khi xử lý các bài toán với nhiều ràng buộc phức tạp và điều kiện khắt khe [8]. Bên cạnh đó, thuật toán ACO yêu cầu chi phí tính toán lớn và thường gặp khó khăn trong việc duy trì tốc độ hội tụ, đặc biệt khi quy mô bài toán tăng lên đáng kể [9].

Xuất phát từ những hạn chế này, bài báo đề xuất sử dụng MVO— một thuật toán tối ưu hóa lấy cảm hứng từ lý thuyết đa vũ trụ, trong đó các vũ trụ trao đổi vật chất để tìm kiếm giải pháp tối ưu. MVO được xem như một phương pháp thay thế đầy triển vọng với hiệu suất cao hơn, giúp tối ưu hóa chi phí và thời gian vận chuyển trong chuỗi cung ứng.

3. MÔ HÌNH BÀI TOÁN TỐI ƯU VẬN CHUYỂN HÀNG HÓA

3.1. Mô tả bài toán

Bài toán tối ưu vận chuyển hàng hóa nhằm xác định phương án vận chuyển tối ưu từ nhà cung cấp đến các điểm tiêu thụ, đảm bảo chi phí thấp nhất và đáp ứng thời gian giao hàng. Quá trình này đòi hỏi xem xét nhiều yếu tố như khoảng cách, phương tiện vận chuyển, năng lực cung ứng và nhu cầu của điểm tiêu thụ. Trong nghiên cứu này, mục tiêu là tìm ra giải pháp vận chuyển tối ưu sao cho tổng chi phí và thời gian giao hàng được giảm thiểu đồng thời, đồng thời vẫn đảm bảo các ràng buộc về cung - cầu.

3.2. Tập hợp và biến số bài toán

Cho một hệ thống chuỗi cung ứng bao gồm:

- m nhà cung cấp $S_1, S_2, \dots, S_m$ với năng lực cung ứng giới hạn.
- n trung tâm phân phối $D_1, D_2, \dots, D_n$ với nhu cầu hàng hóa cụ thể.
- Ma trận chi phí vận chuyển $C_{i,j}$: Chi phí vận chuyển một đơn vị hàng hóa từ nhà cung cấp S_i đến trung tâm phân phối D_j
- Ma trận thời gian vận chuyển $T_{i,j}$: Thời gian cần để vận chuyển một đơn vị hàng hóa từ S_i đến D_j

Mục tiêu là tìm ra phương án vận chuyển tối ưu sao cho tổng chi phí vận chuyển và tổng thời gian giao hàng được tối thiểu hóa đồng thời, trong khi vẫn đáp ứng các ràng buộc về cung - cầu.

3.3. Hàm mục tiêu

Hàm Fitness được thiết kế để tối ưu hóa đồng thời thời chi phí và thời gian vận chuyển, đồng thời tuân thủ các ràng buộc về thời gian giao hàng và giới hạn trọng tải của xe. Để đảm bảo hiệu quả sử dụng phương tiện, mỗi xe tải chỉ được vận hành khi lượng hàng hóa chở trên xe đạt ít nhất $k_{qd} \times$ tải trọng tối đa. Nói cách khác, nếu một xe tải có tải trọng tối đa là W_{max} , thì lượng hàng hóa vận chuyển trên xe phải tối thiểu bằng $k_{qd} \times W_{max}$

Hàm đánh giá Fitness:

$$\text{Fitness} = \alpha Z_1 + \beta Z_2 + \gamma P_L + \mu P_M + \delta P_S \quad (1)$$

Trong đó:

α : Hệ số trọng số điều chỉnh mức độ ưu tiên của chi phí.

β : Hệ số trọng số điều chỉnh mức độ ưu tiên của thời gian.

γ : Hệ số phạt khi vi phạm thời gian.

μ : Hệ số phạt khi vi phạm ràng buộc trọng tải.

δ : Hệ số phạt khi vi phạm năng lực cung ứng.

a. Tổng chi phí vận chuyển (Z_1)

Tối ưu chi phí vận chuyển:

$$Z_1 = \sum_{i=1}^m \sum_{j=1}^n X_{i,j} \times C_{i,j} \quad (2)$$

Trong đó:

$X_{i,j}$: Số lượng hàng hóa vận chuyển từ nhà cung cấp S_i đến trung tâm phân phối D_j

$C_{i,j}$: Chi phí vận chuyển 1 đơn vị hàng hóa từ S_i đến D_j

b. Tổng thời gian giao hàng (Z_2)

$$Z_2 = \sum_{i=1}^m \sum_{j=1}^n X_{i,j} \times T_{i,j} \quad (3)$$

Trong đó:

$T_{i,j}$: Thời gian vận chuyển từ S_i đến D_j .

Tối ưu hóa để đảm bảo giao hàng đúng hạn

c. Ràng buộc về thời gian giao hàng:

$$P_L = \sum_{i=1}^m \sum_{j=1}^n X_{i,j} \times \max(0, T_{i,j} - T_{max}) \quad (4)$$

Trong đó:

T_{max} : Thời gian vận chuyển tối đa cho phép.

Nếu thời gian giao hàng vượt quá giới hạn T_{max} , sẽ bị phạt để làm giảm độ ưu tiên của phương án đó.

d. Ràng buộc về tải trọng tối thiểu: Xe tải chỉ được xuất phát nếu chở ít nhất $k_{qd} * \text{tải trọng tối đa}$.

$$P_M = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^{N_{i,j}} \max(0, k_{qd} * W_{max} - X_{i,j,k}) \quad (5)$$

Trong đó:

k : Chỉ số xe tải trên tuyến.

$N_{i,j}$: Số xe tải phục vụ tuyến $i \rightarrow j$.

$X_{i,j,k}$: Khối lượng hàng hóa xe tải k chở trên tuyến $i \rightarrow j$

e. Hình phạt khi nhà cung cấp giao vượt quá năng lực tối đa:

$$P_S = \sum_{i=1}^m \max(0, \sum_{j=1}^n X_{i,j} - \text{Capacity}_i) \quad (6)$$

Trong đó

$X_{i,j}$: Số lượng hàng hóa vận chuyển từ nhà cung cấp S_i đến trung tâm phân phối D_j

Capacity_i : Năng lực cung ứng tối đa của nhà cung cấp i .

Hàm fitness tối ưu hóa cả chi phí và thời gian vận chuyển, đảm bảo đơn hàng được giao đúng hạn mà không vi phạm giới hạn thời gian. Đồng thời, nó kiểm soát tải trọng phương tiện, giảm lãng phí và hỗ trợ xác định chính xác số lượng xe cần sử dụng. Ngoài ra, hàm fitness giúp cân bằng giữa chi phí và tốc độ giao hàng, tùy theo mức độ ưu tiên do các hệ số điều chỉnh $\delta, \beta, \gamma, \mu$ xác định. Việc áp dụng hàm này giúp doanh nghiệp nâng cao hiệu suất vận hành, tối ưu hóa nguồn lực logistics, làm cho chuỗi cung ứng linh hoạt hơn và thích ứng tốt với biến động thị trường.

4. PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN TỐI ƯU VẬN CHUYỂN BẰNG MVO

4.1. Giới thiệu thuật toán MVO

a. Nguyên lý hoạt động của MVO (mô phỏng lý thuyết đa vũ trụ)

Thuật toán MVO được đề xuất bởi Mirjalili (2016) và lấy cảm hứng từ lý thuyết đa vũ trụ trong vật lý thiên văn. Theo lý thuyết này, vũ trụ không chỉ tồn tại một phiên bản duy nhất mà có vô số vũ trụ song song, mỗi vũ trụ có mức năng lượng và trạng thái vật chất khác nhau. MVO mô phỏng quá trình vật chất di chuyển giữa các vũ trụ nhằm tìm ra một trạng thái tối ưu nhất trong không gian tìm kiếm [4].

Trong MVO, các vũ trụ được coi như các lời giải tiềm năng của bài toán tối ưu hóa. Mỗi vũ trụ có một mức năng lượng (fitness) tương ứng với chất lượng lời giải của nó. Các vũ trụ

có mức năng lượng cao hơn sẽ hút vật chất từ các vũ trụ có năng lượng thấp hơn, giúp cải thiện lời giải theo thời gian.

MVO sử dụng ba khái niệm quan trọng để mô phỏng cơ chế vận hành:

- White Hole:

Các vũ trụ có mức năng lượng tốt hơn (tức lời giải tối ưu hơn) sẽ đóng vai trò là White Hole, cho phép vật chất từ các vũ trụ khác di chuyển vào. Điều này giúp các lời giải kém hiệu quả dần tiến hóa về lời giải tối ưu hơn.

- Black Hole:

Các vũ trụ kém hiệu quả hơn sẽ mất vật chất và trở nên ít ảnh hưởng trong quá trình tối ưu. Điều này giúp loại bỏ các lời giải kém chất lượng và tập trung vào các hướng tìm kiếm tiềm năng hơn.

- Wormhole Tunnel:

+ Vật chất trong các vũ trụ liên tục được tái phân phối giữa các vũ trụ thông qua một cơ chế trao đổi xác suất.

+ Những vũ trụ có lời giải tốt sẽ thu hút vật chất từ những vũ trụ yếu hơn, giúp cải thiện mức độ tối ưu của giải pháp.

+ Một số yếu tố ngẫu nhiên cũng được thêm vào để đảm bảo thuật toán không bị mắc kẹt tại cực tiểu cục bộ.

Dưới đây là các bước thực hiện thuật toán MVO:

1. Khởi tạo số lượng vũ trụ N, giới hạn trên dưới của biến
2. Khởi tạo quần thể vũ trụ ngẫu nhiên
3. Tính toán giá trị hàm mục tiêu (Fitness) cho từng vũ trụ
4. While (chưa đạt điều kiện dừng) do:
5. Sắp xếp vũ trụ theo giá trị hàm mục tiêu
6. Cập nhật vị trí của các vũ trụ theo cơ chế White Hole
7. Cập nhật vị trí của các vũ trụ theo cơ chế WormHole
8. Cập nhật quần thể và tính toán lại giá trị hàm mục tiêu
9. End While
10. Trả về vũ trụ có giá trị tốt nhất

Trong đó, điều kiện dừng phổ biến trong MVO:

- Số lần lặp tối đa (Max Iterations):

Dừng khi số vòng lặp đạt đến giá trị được định nghĩa trước.

- Ngưỡng hội tụ (tùy chọn):

Dừng khi độ chênh lệch fitness giữa hai thế hệ liên tiếp nhỏ hơn ε , tức:

$$\left| Fitness_{best}^{(t)} - Fitness_{best}^{(t-1)} \right| < \varepsilon \quad (7)$$

Nhờ các cơ chế này, MVO có thể duy trì sự cân bằng giữa khả năng khai thác (exploitation) và khả năng khám phá (exploration), giúp thuật toán hội tụ nhanh hơn so với các phương pháp metaheuristic khác như GA và PSO.

b. Các bước cập nhật vũ trụ trong thuật toán Multi-Verse Optimization

- Cơ chế White Hole (Trao đổi vật chất giữa các vũ trụ)

+ Mô phỏng: Các vũ trụ có năng lượng tốt hơn (fitness thấp hơn) hút vật chất từ các vũ trụ yếu hơn (fitness cao hơn).

+ Mục tiêu: Giúp vũ trụ yếu tiến hóa dần về lời giải tối ưu hơn.

Công thức cập nhật White Hole

$$X_{i,j}^{new} = X_{k,j} \text{ nếu } r_1 < P_i \quad (8)$$

$X_{i,j}^{new}$: Giá trị mới của biến j trong vũ trụ i.

P_i : Xác suất trao đổi vật chất, tính theo:

$X_{k,j}$: Giá trị từ một vũ trụ tốt hơn k.

$P_i = \frac{1}{i+1}$ (Các vũ trụ tốt hơn có xác suất cao hơn để trao đổi thông tin với vũ trụ khác.)

r_2 : Số ngẫu nhiên trong khoảng [0,1]

- Cơ chế Black Hole (Loại bỏ vũ trụ yếu)

+ Mô phỏng: Nếu một vũ trụ có fitness quá tệ, nó sẽ bị thay thế hoàn toàn bằng một vũ trụ mới.

+ Mục tiêu: Giúp thuật toán loại bỏ lời giải kém, mở rộng phạm vi tìm kiếm.

Công thức cập nhật Black Hole

$$X_{i,j}^{new} = X_{random,j} \text{ nếu } F_i > F_{threshold} \quad (9)$$

$X_{random,j}$: Giá trị mới ngẫu nhiên trong phạm vi của biến $X_{i,j}$

F_i : Giá trị fitness của vũ trụ X_i .

$F_{threshold}$: Ngưỡng fitness, có thể là trung bình của toàn bộ quần thể.

- Cơ chế Wormhole Tunnel trong MVO

Trong thuật toán MVO, Wormhole Tunnel là cơ chế khai thác cục bộ, cho phép một số vũ trụ dịch chuyển đến vùng lân cận của vũ trụ tốt nhất hiện tại. Xác suất để cơ chế này được áp dụng gọi là Wormhole Existence Probability (WEP), được cập nhật tăng tuyến tính theo số vòng lặp:

$$WEP = WEP_{min} + \left(\frac{1}{L}\right) \times (WEP_{max} - WEP_{min}) \quad (10)$$

Khi $r_2 < WEP$, vị trí của biến được cập nhật theo:

$$X_{i,j}^{new} = X_{i,j} + TDR \times (UB_j - LB_j) \times r$$

trong đó TDR giảm dần theo thời gian để tăng độ tinh chỉnh. Cơ chế này giúp tăng khả năng khai thác và tránh kẹt tại cực trị địa phương.

c. Mã giả của thuật toán MVO

Mã giả dưới đây trình bày quy trình triển khai thuật toán MVO được sử dụng trong nghiên cứu này [4].

1. Tạo ngẫu nhiên tập hợp các vũ trụ U
2. Khởi tạo WEP và TDR
3. Lưu vũ trụ tốt nhất hiện tại (Best_universe)
4. Trong khi chưa đạt điều kiện dừng:
 - 4.1. Đánh giá hàm mục tiêu (fitness) cho tất cả các vũ trụ
 - 4.2. Chuẩn hóa Inflation rate thành NI(U)
 - 4.3. Sắp xếp các vũ trụ theo NI(U)
 - 4.4. Cập nhật WEP và TDR theo công thức (10)
 - 4.5. Với mỗi vũ trụ i:
 - a. Với mỗi biến j trong vũ trụ i:
 - Sinh số ngẫu nhiên $r_1 \in [0, 1]$

- Nếu $r_1 < NI(U[i])$:
 - Chọn vũ trụ khác k bằng Roulette Wheel Selection (dựa trên $-NI$)
 - Cập nhật theo cơ chế White Hole, công thức (8):

$$x_{ij}^{new} = x_{kj}$$
 - Sinh số ngẫu nhiên $r_2 \in [0, 1]$
 - Nếu $r_2 < WEP$:
 - Sinh thêm $r_3, r_4 \in [0, 1]$
 - Nếu $r_3 < 0.5$:

$$x_{ij}^{new} = Best_universe[j] + TDR \times ((ub_j - lb_j) \times r_4 + lb_j)$$
 - Ngược lại:

$$x_{ij}^{new} = Best_universe[j] - TDR \times ((ub_j - lb_j) \times r_4 + lb_j)$$
 - b. Nếu $fitness(U[i]) > threshold$:
 - Cập nhật theo cơ chế Black Hole theo công thức (9):
5. Trả về vũ trụ tốt nhất sau khi điều kiện dừng được thỏa mãn

4.2. Thuật toán MVO-COST giải bài toán tối ưu vận chuyển bằng MVO

Thuật toán MVO-COST được xây dựng dựa trên nguyên lý của MVO, nhằm tối ưu hóa chi phí và thời gian vận chuyển, đồng thời đảm bảo các ràng buộc về thời gian giao hàng và trọng tải phương tiện.

Input	Mô tả
$S = \{S_1, S_2, \dots, S_m\}$	Danh sách nhà cung cấp
$D = \{D_1, D_2, \dots, D_n\}$	Danh sách trung tâm phân phối
$Capacity_i$	Năng lực cung ứng của mỗi nhà cung cấp S_i
$Demand_j$	Nhu cầu hàng hóa của mỗi trung tâm phân phối D_j
$C_{i,j}$	Ma trận chi phí vận chuyển
$T_{i,j}$	Ma trận thời gian vận chuyển
W_{max}	Trọng tải tối đa của phương tiện
T_{max}	Giới hạn thời gian giao hàng
$\alpha, \beta, \gamma, \delta, \mu$	Hệ số trọng số trong hàm fitness
Output	Mô tả
$X_{i,j}^*$	Phương án vận chuyển tối ưu: Ma trận phân bổ hàng hóa giữa các nhà cung cấp và trung tâm phân phối
Z_1^*	Tổng chi phí vận chuyển tối ưu: Tổng chi phí vận chuyển hàng hóa thấp nhất có thể đạt được
Z_2^*	Tổng thời gian giao hàng tối ưu: Tổng thời gian vận chuyển hàng hóa tối thiểu
Số lượng xe tải sử dụng	Tổng số xe tải cần thiết
P_L, P_M, P_S	Vi phạm thời gian giao hàng P_L , vi phạm tải trọng xe P_M , vi phạm năng lực cung ứng P_S

Dưới đây là các bước thực hiện thuật toán MVO-COST:

1. KHỞI TẠO:

1.1. Khởi tạo tập hợp N vũ trụ (lời giải ban đầu) với các giá trị $X_{i,j,k}$ ngẫu nhiên trong phạm vi hợp lệ.

1.2. Đánh giá giá trị hàm mục tiêu (Fitness) cho mỗi vũ trụ.

2. ĐÁNH GIÁ FITNESS:

2.1. Tính tổng chi phí vận chuyển: Z_1

2.2. Tính tổng thời gian vận chuyển: Z_2

2.3. Kiểm tra vi phạm thời gian giao hàng: P_L

2.4. Kiểm tra vi phạm trọng tải phương tiện: P_M

2.5. Hình phạt khi nhà cung cấp giao vượt quá năng lực tối đa: P_S

2.6. Tính giá trị fitness:

$$\text{Fitness} = \alpha Z_1 + \beta Z_2 + \gamma P_L + \mu P_M + \delta P_S$$

3. SẮP XẾP vũ trụ theo giá trị fitness từ tốt nhất đến kém nhất.

4. BẮT ĐẦU VÒNG LẶP (tối ưu hóa)

While (chưa đạt $\max_{\text{iterations}}$) do:

4.1. Cập nhật vị trí của vũ trụ theo cơ chế WHITE HOLE:

- Chọn vũ trụ X_i và thay thế một phần giá trị bằng X_k từ vũ trụ tốt hơn.

$$X_{i,j,k}^{\text{new}} = X_{k,j,k} \text{ nếu } r < P_i, \text{ với } P_i = \frac{1}{i+1}$$

4.2. Cập nhật vị trí của vũ trụ theo cơ chế BLACK HOLE:

- Nếu fitness của $X_i > F_{\text{threshold}}$, thay X_i bằng một vũ trụ ngẫu nhiên.

$$X_{i,j,k}^{\text{new}} = \text{Random}(X_{\text{lower}}, X_{\text{upper}})$$

4.3. Cập nhật vị trí của vũ trụ theo cơ chế WORMHOLE TUNNEL:

- Dịch chuyển ngẫu nhiên trong không gian tìm kiếm:

$$X_{i,j,k}^{\text{new}} = X_{i,j,k} + \text{TDR} * (UB_j - LB_j) * r$$

4.4. Tính lại giá trị fitness của vũ trụ mới.

4.5. Nếu tìm thấy lời giải có fitness tốt hơn, cập nhật lời giải tốt nhất.

5. KẾT THÚC vòng lặp, trả về lời giải tối ưu $X_{i,j,k}^*$

5. THỰC NGHIỆM

Thực nghiệm được thực hiện nhằm đánh giá hiệu quả của thuật toán MVO trong bài toán tối ưu hóa vận chuyển hàng hóa. Cụ thể, nghiên cứu sẽ tiến hành so sánh MVO với các thuật toán phổ biến khác như GA và PSO dựa trên các tiêu chí quan trọng, bao gồm chi phí vận chuyển, thời gian giao hàng và số lần vi phạm ràng buộc.

Bài toán vận chuyển hàng hóa được xây dựng trong một hệ thống chuỗi cung ứng gồm n nhà cung cấp ($S_1, S_2, S_3, \dots, S_n$) và m trung tâm phân phối ($D_1, D_2, D_3, \dots, D_m$). Mỗi trung tâm phân phối có một mức nhu cầu hàng hóa xác định, trong khi các nhà cung cấp hoạt động với giới hạn cung ứng tối đa, đảm bảo đáp ứng nhu cầu nhưng không vượt quá khả năng phân phối của từng đơn vị.

Thông tin về nhu cầu hàng hóa tại các trung tâm phân phối (tính theo kg) được trình bày chi tiết trong Bảng 1.

Bảng 1. Nhu cầu hàng hóa tại các trung tâm phân phối (kg)

Trung tâm phân phối	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆	D ₇	D ₈	D ₉
Nhu cầu (kg)	350	400	500	380	450	520	360	420	480

Giới hạn năng lực cung ứng của từng nhà cung cấp được trình bày chi tiết trong Bảng 2.

Bảng 2. Năng lực cung ứng của các nhà cung cấp (kg)

Nhà cung cấp	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉
Giới hạn cung ứng	450	500	500	470	520	530	490	510	525

Phương tiện vận chuyển: Số xe tải: 15 xe. Tải trọng tối đa mỗi xe: 300 kg.

Bảng 3 thể hiện ma trận chi phí vận chuyển một đơn vị hàng hóa giữa các nhà cung cấp và trung tâm phân phối.

Bảng 3. Chi phí vận chuyển một đơn vị hàng hóa

Nhà cung cấp → Trung tâm	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆	D ₇	D ₈	D ₉
S ₁	5	7	9	6	8	10	4	7	9
S ₂	6	8	7	5	9	11	6	8	10
S ₃	4	6	5	7	9	12	5	6	8
S ₄	5	7	6	8	10	11	4	5	7
S ₅	7	9	8	6	8	12	6	7	9
S ₆	6	8	9	7	10	13	5	8	10
S ₇	5	7	6	8	9	11	6	7	8
S ₈	4	6	5	6	7	9	5	6	7
S ₉	5	7	6	7	9	10	4	5	6

Bảng 4 trình bày ma trận thời gian vận chuyển (giờ) giữa các nhà cung cấp và trung tâm phân phối, dùng làm cơ sở đánh giá ràng buộc thời gian trong mô hình.

Bảng 4. Thời gian vận chuyển giữa nhà cung cấp và trung tâm phân phối (giờ)

Nhà cung cấp → Trung tâm	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆	D ₇	D ₈	D ₉
S ₁	2	3	5	4	6	7	3	4	5
S ₂	3	4	6	2	5	8	4	5	6
S ₃	1	2	4	3	5	7	2	3	4
S ₄	2	3	5	4	6	8	3	4	5
S ₅	3	5	6	4	6	9	3	5	6
S ₆	2	4	5	3	6	8	4	5	7
S ₇	3	5	7	4	6	9	3	4	6
S ₈	1	3	4	2	5	7	2	3	5
S ₉	2	4	6	3	6	8	3	5	6

Kết quả thực nghiệm trên hệ thống chuỗi cung ứng (9 nhà cung cấp, 9 trung tâm phân phối) cho thấy MVO vượt trội hơn GA và PSO về chi phí, thời gian giao hàng và tuân thủ ràng buộc logistics.

Cụ thể, MVO đạt chi phí thấp nhất (27,150\$) và thời gian giao hàng ngắn nhất (5.4 giờ), trong khi GA và PSO có chi phí cao hơn (28,900\$ và 30,300\$) và thời gian giao hàng dài hơn (5.8 giờ và 6.3 giờ).

Về vi phạm ràng buộc, MVO không vi phạm tải trọng hay năng lực cung ứng, trong khi GA có 5 lần vi phạm tải trọng, 3 lần vi phạm cung ứng và PSO có 7 lần vi phạm tải trọng, 4 lần vi phạm cung ứng. Điều này chứng tỏ MVO tối ưu hơn trong quản lý ràng buộc logistics, đảm bảo phương án vận chuyển khả thi.

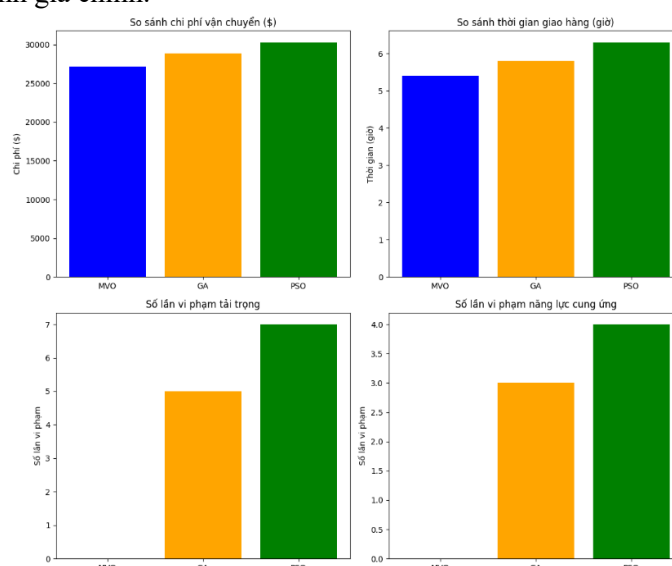
Ngoài ra, MVO có tốc độ hội tụ nhanh hơn, giúp tìm ra lời giải tối ưu hiệu quả hơn so với GA và PSO, khẳng định MVO là phương pháp tối ưu trong bài toán vận chuyển hàng hóa quy mô lớn.

Bảng 5 trình bày kết quả thực nghiệm của ba thuật toán tối ưu (MVO, GA, PSO) theo các tiêu chí: chi phí, thời gian giao hàng và số lần vi phạm ràng buộc.

Bảng 5. So sánh kết quả thực nghiệm giữa các thuật toán tối ưu

Thuật toán	Chi phí vận chuyển (\$)	Thời gian giao hàng (giờ)	Vi phạm tải trọng	Vi phạm cung ứng
MVO	27,150	5.4	0	0
GA	28,900	5.8	5	3
PSO	30,300	6.3	7	4

Biểu đồ trực quan hóa kết quả sẽ giúp minh họa rõ hơn sự khác biệt giữa các thuật toán về chi phí, thời gian giao hàng và số lần vi phạm ràng buộc. Kết quả thực nghiệm này khẳng định MVO là giải pháp hiệu quả hơn trong việc tối ưu hóa vận chuyển hàng hóa trên hệ thống chuỗi cung ứng mở rộng. Hình 1 thể hiện biểu đồ trực quan hóa kết quả giữa ba thuật toán theo các tiêu chí đánh giá chính.



Hình 1. Biểu đồ trực quan hóa kết quả giữa 3 thuật toán MVO, GA, PSO

6. KẾT LUẬN

Kết quả nghiên cứu cho thấy thuật toán MVO là phương pháp tối ưu hóa hiệu quả trong bài toán vận chuyển hàng hóa. So với các thuật toán phổ biến như GA và PSO, MVO giúp giảm chi phí vận chuyển, rút ngắn thời gian giao hàng và đảm bảo tuân thủ các ràng buộc logistics. MVO không chỉ tìm ra lời giải tối ưu nhanh hơn mà còn hạn chế tối đa các vi phạm về tải trọng và năng lực cung ứng, đảm bảo tính khả thi của phương án vận chuyển. Với những ưu điểm này, MVO có tiềm năng ứng dụng rộng rãi trong thực tế, giúp doanh nghiệp nâng cao hiệu suất vận hành, tối ưu hóa chuỗi cung ứng và đáp ứng tốt hơn nhu cầu thị trường. Trong tương lai, nghiên cứu có thể mở rộng để kiểm tra hiệu suất của MVO trên các hệ thống logistics phức tạp hơn, bao gồm nhiều biến động về nguồn cung, nhu cầu và điều kiện vận chuyển thực tế.

TÀI LIỆU THAM KHẢO

- [1] Bai, W., Zhang, W., Meng, F., Enyekwe, I., Reyes, P. M., & Allmendinger, R. (2024). A Two-Stage Metaheuristic Technique Constructed by Particle Swarm Optimization and Simulated Annealing for a Complex Supply Chain Management Model. SSRN.
- [2] Dzalbs, I., & Kalganova, T. (2020). Accelerating Supply Chains with Ant Colony Optimization across Range of Hardware Solutions. arXiv preprint arXiv:2001.08102.
- [3] Zhang, H., Guo, Y. W., Hou, Y., Tang, L., & Deveci, M. (2024). Improved Whale Optimization Algorithm for Supply Chain Financial Risk Assessment of Cloud Warehouse Platform. *Advances in Production Engineering & Management*, 19(3), 395–407.
- [4] Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-Verse Optimizer: A Nature-Inspired Algorithm for Global Optimization. *Knowledge-Based Systems*, 89, 286–301.
- [5] Rosales Muñoz, A. A., Grisales-Noreña, L. F., Montano, J., Montoya, O. D., & Perea-Moreno, A.-J. (2022). Application of the Multiverse Optimization Method to Solve the Optimal Power Flow Problem in Alternating Current Networks. *Electronics*, 11(8), 1287.
- [6] Siegfried, P., & Biedermann, H. (2021). Optimization of Supply Chain Network using Genetic Algorithms based on Bill of Materials. Munich Personal RePEc Archive.
- [7] Bai, W., Zhang, W., Meng, F., Enyekwe, I., Reyes, P. M., & Allmendinger, R. (2024). A Two-Stage Metaheuristic Technique Constructed by Particle Swarm Optimization and Simulated Annealing for a Complex Supply Chain Management Model. SSRN.
- [8] Dzalbs, I., & Kalganova, T. (2020). Accelerating Supply Chains with Ant Colony Optimization across a Range of Hardware Solutions. *Computers & Industrial Engineering*, 147, 106668.
- [9] Zhang, H., Guo, Y. W., Hou, Y., Tang, L., & Deveci, M. (2024). Improved Whale Optimization Algorithm for Supply Chain Financial Risk Assessment of Cloud Warehouse Platform. ResearchGate.

ABSTRACT

OPTIMIZATION OF TRANSPORTATION COST AND TIME USING THE MULTI-VERSE OPTIMIZATION ALGORITHM

Nguyen Thi Thuy Trang¹, Phan Thi Ngoc Mai^{1,*}

¹*Ho Chi Minh City University of Industry and Trade*

*Email: maiphantn@huit.edu.vn

This paper proposes the application of the Multi-Verse Optimization (MVO) algorithm to optimize transportation cost and delivery time in supply chains. The experimental model is constructed using data that includes suppliers, distribution centers, transportation vehicles, and cost–time matrices. The MVO algorithm is implemented and compared with other methods

such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Experimental results show that MVO significantly reduces transportation costs, improves delivery time, and minimizes constraint violations. In addition, MVO demonstrates faster convergence than GA and PSO, leading to more effective route planning and vehicle selection. This method also adapts well to changes in demand and supply chain conditions. The findings suggest that MVO holds strong potential for logistics applications, helping businesses reduce operational costs, improve efficiency, and ensure service quality.

Keywords: Multi-Verse Optimization, supply chain optimization, freight transportation, cost optimization, logistics

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thị Thùy Trang

Số điện thoại liên hệ: 0936522047

TỐI ƯU HÓA CHIM HẢI ÂU BẮC CỰC TRONG CÁC BÀI TOÁN TỐI ƯU HÓA

Bạch Ngọc Vy¹, Phùng Thị Ly¹, Đinh Nguyễn Trọng Nghĩa^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: nghiadnt@hufi.edu.vn

Ngày nhận bài: 11/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Bài báo giới thiệu thuật toán tối ưu hóa tên là “Arctic Puffin Optimization” (APO), lấy cảm hứng từ hành vi kiếm ăn và lặn tìm mồi của hải âu. Thuật toán này được thiết kế để tìm kiếm giải pháp tối ưu cho các bài toán phức tạp một cách nhanh chóng và hiệu quả. APO sử dụng các thông số như số lần lặp tối đa, kích thước quần thể, phạm vi tìm kiếm, chiến lược cân bằng giữa khám phá (tìm kiếm trên không) và khai thác (lặn sâu vào vùng tiềm năng). Mặt khác, APO mô phỏng hành vi của các cá thể trong quần thể, phân bố đều trong không gian tìm kiếm và tập trung vào các khu vực có tiềm năng lớn để cải thiện hiệu suất tìm kiếm. Thuật toán cũng có khả năng giảm thiểu rủi ro “mắc kẹt” ở các cực trị cục bộ nhờ sự kết hợp giữa khám phá và khai thác linh hoạt. Thuật toán APO được đề xuất cho các ứng dụng thực tế như tối ưu hóa phi tuyến và bài toán phức hợp, giúp tìm kiếm giải pháp tối ưu với độ tin cậy cao và thời gian xử lý nhanh.

Từ khóa: Arctic Puffin Optimization, tối ưu hóa, tìm kiếm trên không, kiếm ăn dưới nước

1. MỞ ĐẦU

Thuật toán Arctic Puffin Optimization (APO) là một thuật toán tối ưu hóa mô phỏng quá trình sinh tồn và tìm kiếm thức ăn của loài chim hải âu Bắc Cực. Thuật toán này khai thác hành vi tìm kiếm, hợp tác và săn mồi của chim hải âu để tìm kiếm giải pháp tối ưu. Chim hải âu sẽ di chuyển dựa trên một quy trình sinh tồn, tìm kiếm thức ăn trong môi trường. Để tránh bị kẹt trong các cực tiểu cục bộ, thuật toán APO duy trì sự đa dạng trong quần thể chim hải âu, cho phép tìm kiếm các khu vực chưa được khai thác trong không gian tìm kiếm. Chim hải âu sẽ điều chỉnh vị trí của mình thông qua một hàm mục tiêu dựa trên hiệu suất của vị trí hiện tại và một yếu tố ngẫu nhiên để tránh bị mắc kẹt trong các cực tiểu địa phương. Thuật toán tối ưu hóa APO có ưu điểm là dễ triển khai và điều chỉnh tham số, hiệu quả cao trong việc giải các bài toán đa cực trị và có thể kết hợp với các thuật toán khác để tăng hiệu quả [1].

Trong nhiều trường hợp, các thuật toán metaheuristic là một trong những lựa chọn ưu tiên để giải quyết các vấn đề tối ưu hóa, đặc biệt là khi xử lý các vấn đề đặc trưng bởi không gian tìm kiếm cực kỳ phức tạp. Các thuật toán metaheuristic thường được thiết kế dựa trên các hiện tượng tự nhiên, dựa trên các thuật toán tiến hóa như thuật toán di truyền

(Genetic Algorithm – GA) [2], thuật toán tối ưu hóa số học (Arithmetic Optimization Algorithm - AOA) [3], thuật toán tìm kiếm hòa âm (Harmony Search – HS) [4], thuật toán tối ưu hóa bầy kiến (Ant Colony Optimization – ACO) [5]. Tuy các thuật toán metaheuristic vẫn tồn tại một số những hạn chế nhất định nhưng vẫn có ưu điểm chung là tìm nghiệm tốt cho bài toán phức tạp mà các phương pháp tối ưu truyền thống khó xử lý. Các thuật toán này được sử dụng phổ biến để giải quyết các bài toán tối ưu hóa trong nhiều lĩnh vực AI, logistics, mạng viễn thông, điều khiển robot, hệ thống năng lượng, v.v...

Thuật toán APO được lấy cảm hứng từ cách chim hải âu Bắc cực tìm kiếm thức ăn và sinh tồn. Giống như các thuật toán trí thông minh bầy đàn khác, thuật toán APO cân bằng tốt giữa khám phá và khai thác, giúp tránh mắc kẹt ở cực trị cục bộ. Bên cạnh đó, APO có ít tham số hơn, dễ điều chỉnh và ứng dụng rộng rãi trong tối ưu hóa lịch trình, điều hướng robot và tối ưu mạng nơ-ron. Nhờ cơ chế thích nghi linh hoạt của quần thể chim hải âu, thuật toán này có thể duy trì sự đa dạng trong không gian tìm kiếm và tăng hiệu quả trong việc giải quyết các bài toán tối ưu hóa đa mục tiêu và phi tuyến.

Bài báo cáo này sẽ trình bày về thuật toán Arctic Puffin Optimization và ý tưởng áp dụng của nó trong bài toán tối ưu hóa.

2. THUẬT TOÁN TỐI ƯU HÓA CHIM HẢI ÂU BẮC CỰC

2.1. Đặc điểm của thuật toán

Chim hải âu Bắc Cực là loài chim nhỏ, sinh tồn trong môi trường khắc nghiệt với nguồn thức ăn chính là cá nhỏ và các loài sinh vật biển khác. Hải âu có cách kiếm ăn đặc biệt, chúng không chỉ bay trên mặt biển để kiếm thức ăn mà còn lặn sâu dưới nước nhờ đôi cánh khỏe mạnh được sử dụng như mái chèo. Các cá thể và quần thể chim hải âu được thể hiện trong Hình 1.



Hình 1. Cá thể và quần thể chim hải âu Bắc Cực

Chim hải âu thích bay thành bầy ở rìa đại dương, bơi trên bề mặt và kiếm cá. Loài chim này thường kiếm ăn theo nhóm hoặc bầy với kỹ năng tìm kiếm, đánh bắt cá, khả năng bay tuyệt vời và làm việc theo nhóm để kiếm ăn trong đại dương. Bài báo này khai thác hành vi tìm kiếm, hợp tác và săn mồi của chim hải âu để tìm kiếm giải pháp tối ưu.

2.2. Ý tưởng và mô hình toán học của thuật toán

Chim hải âu có thể nhắm mục tiêu vào thức ăn và lao xuống để bắt khi đang bay. Khi phát hiện nguồn thức ăn tiềm năng, chúng sẽ hợp tác trong đàn để tối ưu hóa hiệu quả

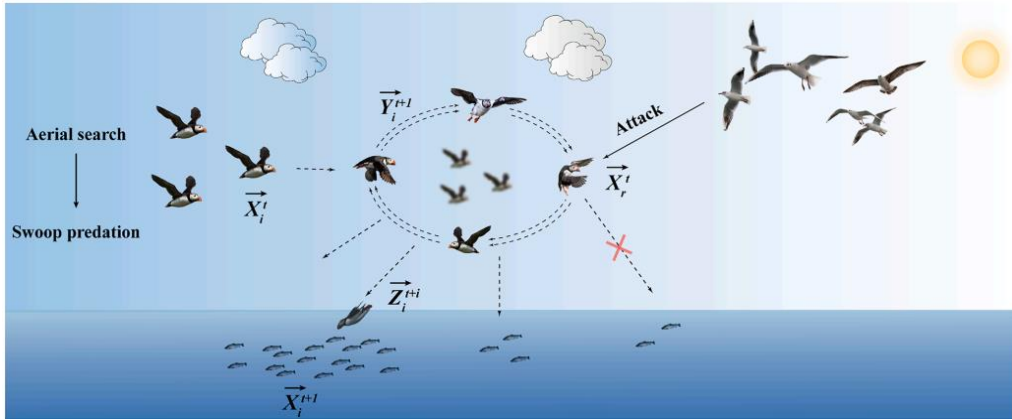
tìm kiếm. Không gian tìm kiếm được mô phỏng như môi trường sinh tồn của chim hải âu, chúng sẽ không biết trước được cấu trúc không gian tìm kiếm, nhưng thông qua việc khám phá và chia sẻ thông tin, chúng định vị các điểm tiềm năng trong không gian và cải thiện giải pháp tổng thể như được minh họa trong Hình 2. Hành vi chính của chim hải âu trong thuật toán này được chia thành ba giai đoạn: khởi tạo quần thể (population initialization), thăm dò (exploration) và hành vi khai thác (exploitation), trong đó hành vi thăm dò gồm 2 chiến lược: giai đoạn tìm kiếm (Aerial Search) và lao xuống săn mồi (Swooping Predation) và hành vi khai thác gồm 3 chiến lược: thu thập kiếm ăn (Gathering Foraging), tăng cường tìm kiếm (Intensifying Search) và tránh kẻ săn mồi (Avoiding Predators).

Ngoài ra, nó liên quan đến sự chuyển đổi giữa hai chiến lược này, được gây ra bởi yếu tố biến đổi hành vi B của hải âu. Thuật toán APO thiết kế hệ số chuyển tiếp hành vi B để đạt được sự chuyển đổi suôn sẻ từ tìm kiếm toàn cầu sang khai thác cục bộ. Nó được định nghĩa như sau:

$$B = 2 * \log(1 / rand) * (1 - t/T) \quad (18)$$

Trong thuật toán APO, B là một tham số thay đổi thích ứng cho phép điều chỉnh động khi các lần lặp lại liên tục, do đó thích ứng với nhu cầu tìm kiếm ở các giai đoạn khác nhau.

Bên cạnh đó, thuật toán APO còn giới thiệu một tham số C, so sánh B với C để xác định chiến lược tìm kiếm nên được thực hiện ở giai đoạn lặp hiện tại. Trong bài báo này, tham số C được đặt thành 0.5. Thiết kế này giúp thuật toán đào sâu hơn vào các giải pháp chất lượng cao tiềm năng và cải thiện tốc độ hội tụ và độ chính xác của nó, giúp cải thiện hiệu suất và độ bền của thuật toán.



Hình 2. Hành vi kiếm ăn trên không của hải âu Bắc Cực

Các mô hình toán học cho mỗi giai đoạn được mô tả như sau:

2.2.1. Khởi tạo quần thể

Mỗi chim hải âu Bắc Cực đại diện cho một giải pháp tiềm năng tham gia vào quá trình tối ưu hóa. Quá trình tạo ra quần thể khởi tạo được mô tả bằng phương trình sau:

$$\vec{X}_i^t = rand * (ub - lb) + lb, i = 1, 2, 3, \dots, N \quad (1)$$

Trong đó $\overrightarrow{X}_i^t$ đại diện cho vị trí của cá nóc Bắc Cực, rand tạo ra một số ngẫu nhiên từ 0 đến 1; ub và lb lần lượt đại diện cho giới hạn trên và dưới, N là số cá thể trong quần thể.

2.2.2. Hành vi thăm dò

Trước hết là *chiến lược tìm kiếm trên không*. Trong mô hình này, chim hải âu thường tham gia vào các chuyến bay phối hợp, phương trình cập nhật vị trí liên quan đến hành vi kiếm ăn trên không được thể hiện như sau:

$$\overrightarrow{Y}_i^{t+1} = \overrightarrow{X}_i^t + (\overrightarrow{X}_i^t - \overrightarrow{X}_r^t) * L(D) + R \quad (2)$$

Trong đó, $\overrightarrow{Y}_i^{t+1}$ là vị trí mới của cá thể hải âu i trong không gian tìm kiếm, $L(D)$ là biểu thị một số ngẫu nhiên được tạo ra theo phân phối Levy (bước nhảy Levy) được tính theo công thức:

$$L(D) = \frac{u}{|v|^{\frac{1}{\beta}}} \text{ với } u \sim N(0, \sigma_u^2), v \sim N(0, 1), \beta \in (0, 2]$$

, D là số chiều không gian. Cá thể hải âu di chuyển theo một cá thể tham chiếu, nếu khoảng cách giữa vị trí hiện tại $\overrightarrow{X}_i^t$ và vị trí tham chiếu $\overrightarrow{X}_r^t$ lớn, $L(D)$ sẽ giúp di chuyển xa hơn, khám phá nhiều vùng mới. Ngược lại, thuật toán sẽ điều chỉnh vị trí để khai thác tốt hơn khu vực có lời giải tối ưu với hệ số bổ sung R được tính theo công thức:

$$R = \text{round}(0.5 * (0.05 + \text{rand})) * \alpha \quad (3)$$

Sự kết hợp của sinh số ngẫu nhiên từ 0 đến 1 của rand và sinh số ngẫu nhiên theo phân phối chuẩn $N(0, 1)$ α đảm bảo rằng thuật toán không bị quá phụ thuộc vào bước nhảy Levy.

Chiến lược thứ hai là sà xuống săn mồi. Để mô hình hóa hành vi lao xuống săn mồi, hệ số vận tốc S được sử dụng để điều chỉnh độ dịch chuyển của hải âu trong quá trình lao xuống. Dưới đây là phương trình cập nhật vị trí của hải âu:

$$\overrightarrow{Z}_i^{t+1} = \overrightarrow{Y}_i^{t+1} * S \quad (4)$$

$$S = \tan((\text{rand} - 0.5) * \pi) \quad (5)$$

Trong chiến lược bay này, chim hải âu điều chỉnh độ dịch chuyển của mình trong giai đoạn đầu tiên bằng cách đưa vào hệ số vận tốc S . Việc đưa vào các đặc tính phân phối của tham số S làm tăng tính ngẫu nhiên, đa dạng của thuật toán. Việc thiết lập giai đoạn săn mồi sà xuống giúp cải thiện hơn nữa hiệu quả tìm kiếm của thuật toán trong không gian giải pháp.

Để đạt được kết quả tối ưu, thuật toán chọn hợp nhất các vị trí được tạo ra ở hai mô hình trong giai đoạn tìm kiếm trên không thành một giải pháp mới. Các giải pháp này sau đó được sắp xếp theo giá trị fitness và N cá thể hàng đầu được chọn để hình thành quần thể mới. Các phương trình được mô tả như sau:

$$\overrightarrow{P}_i^{t+1} = \overrightarrow{Y}_i^{t+1} \cup \overrightarrow{Z}_i^{t+1} \quad (6)$$

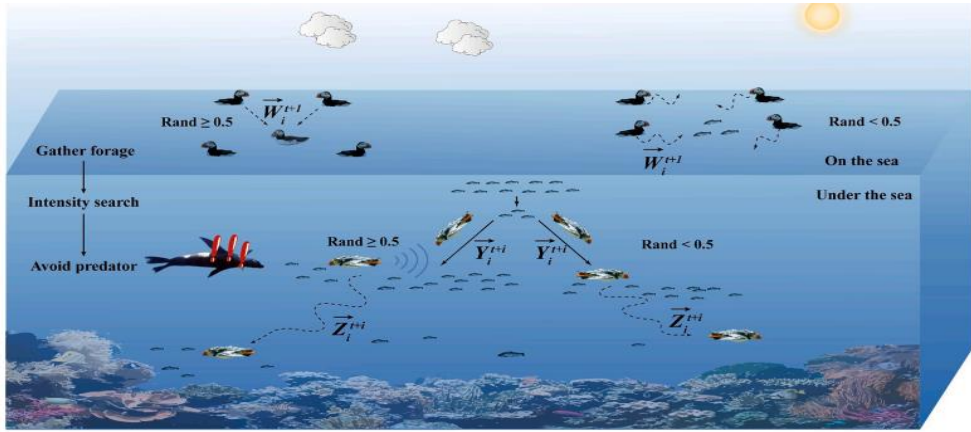
$$new = sort(\overrightarrow{P_i^{t+1}}) \quad (7)$$

$$\overrightarrow{X_i^{t+1}} = new(1:N) \quad (8)$$

Trong đó, hàm sort là sắp xếp quần thể mới theo giá trị fitness từ nhỏ đến lớn để chọn quần thể mới $\overrightarrow{X_i^{t+1}}$.

2.2.3. Hành vi khai thác dưới nước

Tìm kiếm ăn dưới nước bao gồm ba chiến lược chính, mỗi chiến lược được sử dụng trong các bối cảnh cụ thể để nâng cao hiệu quả săn mồi được minh họa trong Hình 3.



Hình 3. Hành vi kiếm ăn dưới nước của chim hải âu

Thu thập kiếm ăn là chiến lược khai thác dưới nước đầu tiên, những con chim ở trên mặt biển sẽ quan sát hành vi của các thành viên khác để xác định các điểm lặn hoặc nguồn thức ăn. Dưới đây là phương trình mô tả cách cập nhật vị trí của hải âu:

$$\overrightarrow{W_i^{t+1}} = \{\overrightarrow{X_{r1}^t} + F * L(D) * (\overrightarrow{X_{r2}^t} - \overrightarrow{X_{r3}^t}) \mid rand \geq 0.5\} \quad (9)$$

$$\overrightarrow{X_{r1}^t} + F * (\overrightarrow{X_{r2}^t} - \overrightarrow{X_{r3}^t}) \mid rand < 0.5 \quad (10)$$

Trong các phương trình này, khi $rand < 0.5$, chim hải âu tham gia kiếm ăn hợp tác với các hải âu khác $\overrightarrow{X_{r1}^t}, \overrightarrow{X_{r2}^t}, \overrightarrow{X_{r3}^t}$ với các biến r_1, r_2, r_3 là các số nguyên ngẫu nhiên từ 1 đến $N - 1$ (không bao gồm i), sử dụng hệ số hợp tác $F = 0.5$ để khám phá môi trường xung quanh. Ngược lại, một chim hải âu có thể thay đổi vị trí để tham gia vào một nhóm săn mồi có lợi hơn để hợp tác tốt hơn. Do đó, bước nhảy Levy được sử dụng một lần nữa để nâng cao khả năng săn mồi nhanh chóng và hiệu quả.

Tăng cường tìm kiếm là chiến lược dưới nước thứ hai. Khi quá trình săn mồi diễn ra, chim hải âu phải thay đổi vị trí dưới nước của mình để tìm kiếm thêm cá hoặc các nguồn thức ăn dưới nước khác. Phương trình cập nhật vị trí cho chiến lược này được thể hiện như sau:

$$\overrightarrow{Y_i^{t+1}} = \overrightarrow{W_i^{t+1}} * (1 + f) \quad (11)$$

$$f = 0.1 * (rand - 1) * \frac{T-t}{T} \quad (12)$$

Trong đó T biểu thị tổng số lần lặp lại và t biểu thị số lần lặp hiện tại. Tham số f là một yếu tố thích ứng được sử dụng để điều chỉnh vị trí của hải âu trong nước.

Tránh kẻ săn mồi là chiến lược dưới nước thứ ba. Chiến lược này được sử dụng để mô tả hành vi của hải âu khi chúng phát hiện ra kẻ săn mồi ở gần. Phương trình cập nhật vị trí cho chiến lược này được thể hiện như sau:

$$\overrightarrow{X_i^{t+1}} = \{\overrightarrow{X_i^t} + F * L(D) * (\overrightarrow{X_{r1}^t} - \overrightarrow{X_{r2}^t})\} rand \geq 0.5 \quad (13)$$

$$\overrightarrow{X_i^t} + \beta * (\overrightarrow{X_{r1}^t} - \overrightarrow{X_{r2}^t}) rand < 0.5 \quad (14)$$

Trong đó β là số ngẫu nhiên phân phối đều giữa (0,1). Khi $rand \geq 0,5$ chỉ ra rằng có kẻ săn mồi ở gần, chim hải âu chọn né tránh ngay lập tức. Ngược lại, hành vi của chim hải âu có xu hướng thận trọng hơn bằng cách thay đổi vị trí ngẫu nhiên. Do đó, chiến lược hành vi kép được mô phỏng trong thuật toán để đảm bảo tính mạnh mẽ và khả năng thích ứng.

Thuật toán chọn hợp nhất các vị trí ứng viên từ ba phương trình vị trí khác nhau thành một giải pháp mới để có được kết quả tối ưu trong các tình huống khác nhau giống với giai đoạn trên. Phương trình cập nhật vị trí được mô tả như sau:

$$\overrightarrow{P_i^{t+1}} = \overrightarrow{W_i^{t+1}} \cup \overrightarrow{Y_i^{t+1}} \cup \overrightarrow{Z_i^{t+1}} \quad (15)$$

$$new = sort(\overrightarrow{P_i^{t+1}}) \quad (16)$$

$$\overrightarrow{X_i^{t+1}} = new(1:N) \quad (17)$$

Chiến lược toàn diện này cho phép mô hình xem xét nhiều kịch bản kiểm ăn cùng lúc, cuối cùng chọn vị trí phù hợp nhất làm giải pháp tối ưu.

Thuật toán: APO

Input:

N: Số lượng cá thể trong quần thể

T: Số vòng lặp tối đa

lb: Giới hạn dưới của không gian tìm kiếm

ub: Giới hạn trên của không gian tìm kiếm

dim: Số chiều của bài toán tối ưu hóa

rand: Sinh số ngẫu nhiên từ 0 đến 1

fobj: Hàm mục tiêu dùng để đánh giá độ tốt của mỗi lời giải

F: Hệ số hành vi

C: Tham số

Khởi tạo quần thể:

// Tạo ma trận vị trí ban đầu và lưu giá trị hàm mục tiêu của từng cá thể

for i = 1 to N do

PopPos(i) = rand * (ub - lb) + lb

PopFit(i) = fobj(PopPos(i))

End for

Vòng lặp tối ưu hóa chính

for $I_t = 1$ to T do

for i = 1 to N do

$B = 2 * \log(1 / \text{rand}) * (1 - I_t / T)$

//Giai đoạn thăm dò

if $B > C$

$Y = \text{PopPos}(i) + \text{Levy}(\text{dim}) * (\text{PopPos}(i) - \text{PopPos}(r)) + \text{round}(0.5 * (0.05 + \text{rand})) * \text{rand}$

$Z = Y * \tan((\text{rand} - 0.5) * \pi)$

NewPop = [Y;Z]

NewPopfit = [fobj(Y);fobj(Z)]

sort(NewPopfit)

newPopPos = NewPop(sorted_indexes(1))

else

//Giai đoạn khai thác

$F = (0.1 * (\text{rand} - 1) * (T - I_t)) / T$

if $\text{rand} < 0.5$

$W = \text{PopPos}(\text{RandInd}(1)) + F * (\text{PopPos}(i) - \text{PopPos}(r))$

else

$W = \text{PopPos}(\text{RandInd}(1)) + F * \text{Levy}(\text{dim}) * (\text{PopPos}(i) - \text{PopPos}(r))$

End if

$Y = (1 + f) * W$

```
Epsilon = unifrnd(0, 1)
if rand < 0.5
    Z = PopPos(i) + Epsilon * ( PopPos(r1) - PopPos(r2))
else
    Z = PopPos(i) + F * Levy(dim) * ( PopPos(r1) - PopPos(r2))
End if
NewPop = [W;Y;Z]
NewPopfit = [fobj(W);fobj(Y);fobj(Z)]
sort(NewPopfit)
newPopPos = NewPop(sorted_indexes(1))
end if
end for
//Cập nhật cá thể tốt nhất toàn cục
BestF = newPopFit(1)
BestX = newPopPos(1)
end for
```

Output: Trả về vị trí tốt nhất **BestX** và giá trị tối ưu **BestF**

3. MINH HỌA THUẬT TOÁN

Lấy ví dụ tìm min value hàm tối ưu được định nghĩa bởi:

$$f(x, y, z) = \left(\frac{x}{4}\right)^2 + \left(\frac{y}{2}\right)^2 + z^2 \quad (19)$$

Trong đó x, y, z sẽ nằm trong khoảng $[-2, 2]$.

Bước đầu tiên khởi tạo quần thể hải âu và hệ số hợp tác ngẫu nhiên $C = 0.5$. Ở đây bài toán có 3 biến x, y, z với số lượng hải âu trong quần thể là 10 cá thể. Sau khi khởi tạo, tiến hành tính toán độ thích nghi (fitness) và hệ số chuyển đổi hành vi ta được *Bảng 1* với hệ số chuyển đổi hành vi $B = 1.177$

Bảng 1. Tập hợp các cá thể hải âu sau khi khởi tạo.

x	y	z	fitness
0.00939	0.20608	-0.75362	0.57856
0.78453	0.27180	-1.25196	1.62433

0.97617	2.00000	-0.44066	1.25373
-1.99289	-1.29971	0.54254	0.96489
0.04226	-1.77581	-0.11033	0.80066
-1.29593	-1.65542	-1.26571	2.39209
0.21160	0.61971	-0.49905	0.34787
0.55083	1.61536	0.48402	0.90559
-2.00000	-0.29844	0.01814	0.27259
1.70743	-1.76717	1.72610	3.94235

Từ hệ số B và C ta dễ dàng nhận thấy $B > C$ nên ta sẽ thực hiện chiến lược tìm kiếm trên không. Thuật toán sẽ sử dụng quần thể vừa khởi tạo để tính toán vị trí mới của cá thể hải âu i trong không gian tìm kiếm theo giá trị $Levy = 0.35224$ và $R = 0.00057$ thu được kết quả ở Bảng 2 như sau:

Bảng 2. Quần thể hải âu sau khi cập nhật lại vị trí tìm kiếm trên không

x	y	z	fitness
-1.86455	-0.02460	0.11296	0.23020
1.66845	1.61386	-2.00000	4.82512
2.00000	-0.72029	-1.72839	3.36703
-2.00000	1.91904	-0.42882	1.35456
-2.00000	0.24097	-0.19836	0.30386
-1.43284	-1.91319	1.27330	2.66469
-0.69076	0.10240	-2.00000	4.03244
-1.79913	0.09500	0.12980	0.22141
-1.91475	0.85936	1.92848	4.13280
2.00000	1.94736	-1.73667	4.21408

Sau khi có vị trí mới trong không gian tìm kiếm, thuật toán sẽ tiếp tục tính vị trí mới của cá thể i lao xuống săn mồi dựa vào vị trí của quần thể trong không gian tìm kiếm và hệ số vận tốc $S = -1.14227$ (điều chỉnh độ dịch chuyển của hải âu trong quá trình lao xuống) thu được kết quả ở Bảng 3 như sau:

Bảng 3. Quần thể chim hải âu sau khi cập nhật vị trí và xuống săn mồi

x	y	z	fitness
1.60617	2.00000	0.21205	1.20620
-2.00000	0.24983	-0.20897	0.30927
1.75929	-1.62290	-0.33706	0.96550
-0.59767	-2.00000	-1.14934	2.34332
-1.83300	-1.68101	-0.46843	1.13587
0.61062	-0.40796	-2.00000	4.06491

-1.43784	-1.76151	0.33477	1.01701
1.88446	-0.19249	-1.87406	3.74330
1.13503	2.00000	-1.54787	3.47642
0.58428	0.96533	-1.38418	2.17026

Sau khi cập nhật vị trí mới của quần thể, nếu giá trị của các biến x, y, z vượt khỏi khoảng $[-2, 2]$ thì sẽ tự động cập nhật lại về giá trị biên. Tập hợp ba quần thể trên lại thành một và sắp xếp theo thứ tự tăng dần của độ thích nghi (fitness) và chọn ra 10 cá thể đầu tiên làm quần thể khởi tạo mới. Khi kết thúc, kết quả đạt được thể hiện ở *Bảng 4*:

Bảng 4. Quần thể chim hải âu sau khi hoàn thành vòng lặp 1

x	y	z	fitness
-1.79913	0.09500	0.12980	0.22141
-1.86455	-0.02460	0.11296	0.23020
-2.00000	-0.29844	0.01814	0.27259
-2.00000	0.24097	-0.19836	0.30386
-2.00000	0.24983	-0.20897	0.30927
0.21160	0.61971	-0.49905	0.34787
0.00939	0.20608	-0.75362	0.57856
0.04226	-1.77581	-0.11033	0.80066
0.55083	1.61536	0.48402	0.90559
-1.99289	-1.29971	0.54254	0.96489

Tiến hành lặp lại 100 vòng lặp sẽ được một tập hợp các vị trí cá thể hải âu như trong *Bảng 5* dưới đây.

Bảng 5. Quần thể chim hải âu sau 100 vòng lặp

x	y	z	fitness
-3.98086E-11	-4.23145E-11	1.39025E-12	5.48607E-22
2.21596E-10	-1.61821E-11	-1.2908E-11	3.30113E-21
3.56737E-11	5.75045E-11	-4.96256E-11	3.36893E-21
2.90823E-10	-4.63876E-11	2.33914E-11	6.37123E-21
5.75517E-11	-3.43557E-11	-7.92088E-11	6.77613E-21
3.42747E-11	-1.56625E-11	-8.80154E-11	7.88146E-21
6.43301E-11	-1.83344E-10	8.27348E-12	8.7309E-21
2.87056E-10	-1.32165E-10	2.50751E-11	1.01457E-20
3.20812E-10	-2.25884E-10	-3.32588E-11	2.02946E-20
8.2276E-11	-2.10233E-10	-9.4741E-11	2.04484E-20

Sau khi quá trình hoàn tất, giá trị tối ưu nhất đầu tiên trong bảng là kết quả của bài toán cần tìm.

4. ỨNG DỤNG VÀO BÀI TOÁN TỐI ƯU HÓA

4.1. Giới thiệu bài toán tối ưu hóa

Bài toán tối ưu hóa (The Optimization Problem) là bài toán tìm kiếm giá trị tốt nhất (cực đại hoặc cực tiểu) của một hàm mục tiêu trong một tập hợp các ràng buộc cho trước trong tất cả các lời giải khả thi. Bài toán tối ưu hóa có thể được chia thành hai loại tùy thuộc vào việc các biến là liên tục hay rời rạc. [6]

Định Nghĩa 1: Hàm mục tiêu: Hàm cần tối ưu sao cho hàm đó đạt giá trị nhỏ nhất (tối thiểu hóa) hoặc lớn nhất (tối đa hóa). Ví dụ: Trong bài toán hồi quy tuyến tính, ta cần tối thiểu hóa tổng bình phương sai số (Mean Squared Error - MSE).

Định Nghĩa 2: Biến quyết định: Các giá trị mà ta cần điều chỉnh để đạt giá trị tối ưu của hàm mục tiêu. Ví dụ: Trong bài toán lập lịch sản xuất, biến quyết định có thể là số lượng sản phẩm được sản xuất trong ngày.

Định Nghĩa 3: Tập khả thi: Tập các giá trị hợp lệ của biến quyết định. Ví dụ: Trong bài toán phân bổ ngân sách, tổng số tiền đầu tư không được vượt quá ngân sách B, ở đây B chính là tập khả thi.

Định Nghĩa 4: Ràng buộc: Điều kiện giới hạn giá trị của biến quyết định. Ví dụ: Trong bài toán phân bổ ngân sách, tổng số tiền đầu tư không được vượt quá ngân sách B, tức là $\sum x_i \leq B$.

Định Nghĩa 5: Nghiệm tối ưu: Giá trị của biến quyết định cho kết quả tối ưu (nhỏ nhất hoặc lớn nhất). Ví dụ: Cho phương trình $f(x): x - 1 = 0$, nghiệm $x = 1$ là nghiệm tối ưu của $f(x)$.

Một bài toán tối ưu hóa tổng quát có dạng:

$$x^* = \operatorname{argmin} f_0(x) \text{ sao cho } \begin{cases} f_i(x) \leq 0, & i = 1, 2, 3, \dots, m \\ h_j(x) \leq 0, & j = 1, 2, 3, \dots, m \end{cases} \quad (20)$$

trong đó hàm $f_i(x)$ và $h_i(x)$ là các ràng buộc của biến x để tối thiểu hàm $f_0(x)$. Khi $m = p = 0$, bài toán (20) được gọi là bài toán tối ưu không ràng buộc. Giá trị tối ưu của bài toán tối ưu (20) được định nghĩa là:

$$p^* = \inf \inf \{f_0(x) \mid f_i(x) \leq 0, i = 1, 2, 3, \dots, m; h_j(x) = 0, j = 1, 2, 3, \dots, p\} \quad (21)$$

trong đó $\inf$ là viết tắt của hàm infimum. p^* có thể nhận các giá trị $\pm\infty$. Nếu bài toán là không khả thi, ta coi $p^* = +\infty$, nếu hàm mục tiêu không bị chặn dưới trong tập xác định, ta coi $p^* = -\infty$.

Trong toán tối ưu (thường là không gian nhiều chiều), gọi một điểm x là cực tiểu nếu tồn tại một giá trị (thường được gọi là bán kính) R sao cho:

$$f_0(x) = \inf \{f_0(z) \mid f_i(z) \leq 0, i = 1, 2, 3, \dots, m; h_j(z) = 0, j = 1, 2, 3, \dots, p, \|z - x\|_2 \leq R\} \quad (22)$$

Nếu một điểm x thỏa mãn $f_i(x) = 0$, ta nói rằng bất đẳng thức ràng buộc thứ i : $f_i(x) = 0$ là khả dụng. Nếu $f_i(x) < 0$, ta nói rằng ràng buộc này là không khả dụng tại x .

Mặc dù trong định nghĩa bài toán tối ưu (20) là cho bài toán tối thiểu hàm mục tiêu với các ràng buộc thỏa mãn các điều kiện nhỏ hơn hoặc bằng 0, các bài toán tối ưu với tối đa hàm mục tiêu và điều kiện ràng buộc ở dạng khác đều có thể đưa về được dạng này:

$$\max f_0(x) \leftrightarrow \min -f_0(x) \quad (23)$$

$$f_i(x) \leq g(x) \leftrightarrow f_i(x) - g(x) \leq 0 \quad (24)$$

$$f_i(x) \geq 0 \leftrightarrow -f_i(x) \leq 0 \quad (25)$$

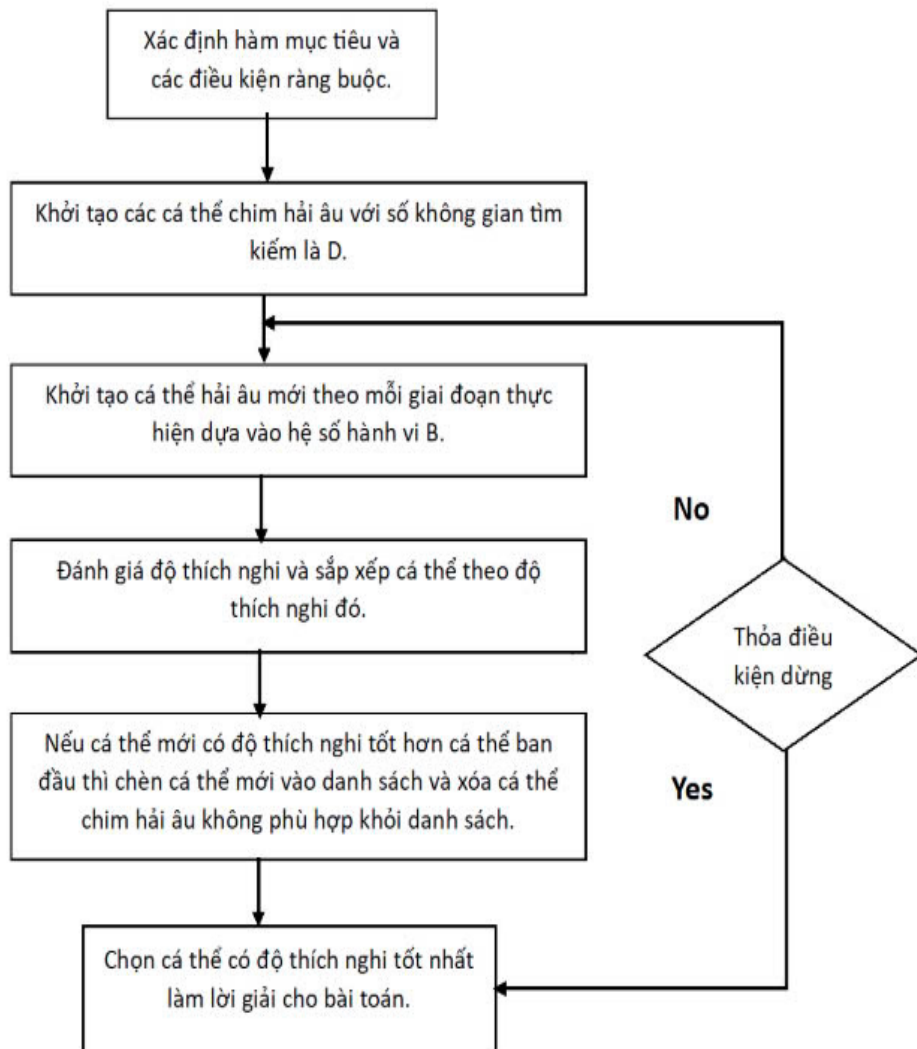
$$a \leq f_i(x) \leq b \leftrightarrow f_i(x) - b \leq 0 \text{ và } a - f_i(x) \leq 0 \quad (26)$$

Trong đó, $f_i(x) \leq 0 \leftrightarrow f_i(x) + s = 0$ và $s_i \geq 0$. s_i được gọi là biến rời rạc, phép biến đổi đơn giản này trong nhiều trường hợp lại tỏ ra hiệu quả vì bất đẳng thức $s_i \geq 0$ thường dễ giải quyết hơn là $f_i(x) \leq 0$.

4.2. Ứng dụng

Trong thuật toán APO, cuối mỗi vòng lặp t , người ta tìm ra tập lời giải $\Omega(t)$ và tập thuật toán tìm kiếm địa phương $\mathcal{A}(t)$. Thay vì sử dụng một phương pháp cố định cho toàn bộ vòng lặp, thuật toán sẽ linh hoạt lựa chọn phương pháp phù hợp tùy theo ràng buộc bài toán. Kết quả thực nghiệm cho thấy việc áp dụng tìm kiếm địa phương đa dạng và linh hoạt ở mỗi bước lặp tùy theo các ràng buộc và đặc điểm hàm mục tiêu cải thiện đáng kể chất lượng thuật toán so với các thuật toán sử dụng đơn điệu một phương pháp tìm kiếm cho mọi bước lặp.

Sau khi xác định được hàm mục tiêu $f_0(x)$ và các điều kiện ràng buộc $\begin{cases} f_i(x) \leq 0, & i = 1, 2, 3, \dots, m \\ h_j(x) \leq 0, & j = 1, 2, 3, \dots, m \end{cases}$ cho bài toán cần tối ưu, khởi tạo quần thể hải âu (tập khả thi) X với số lượng N cá thể sao cho mỗi cá thể đại diện cho một giải pháp x với n chỉ số (thỏa điều kiện ràng buộc) và các tham số, các cá thể sẽ được đánh giá độ thích nghi thông qua $f_0(x)$. Dựa vào tham số vừa khởi tạo để xác định hướng giải pháp cho bài toán và điều chỉnh các chỉ số của x sao cho phù hợp với hướng giải pháp vừa xác định. Cụ thể, trong mỗi vòng lặp t của thuật toán APO áp dụng cho bài toán tối ưu, các giải pháp mới x_i^{t+1} sẽ được cập nhật dựa trên giải pháp hiện tại x_i^t của chúng kết hợp với giải pháp khác tại thời điểm đó x_j^t và các tham số ngẫu nhiên. Sau khi cập nhật và đánh giá độ thích nghi của x_i^{t+1} ta sẽ có tập khả thi mới X^{t+1} , kết hợp với X và chọn ra N giải pháp có độ thích nghi tối ưu nhất làm X cho vòng lặp tiếp theo. Quá trình này được tiếp tục cho đến khi đáp ứng được tiêu chuẩn dừng hoặc đã đạt đến số lần lặp tối đa được thiết lập. Bài toán tối ưu hóa sử dụng thuật toán APO được biểu thị bởi sơ đồ ở Hình 4.



Hình 4. Sơ đồ phương pháp giải bài toán tối ưu sử dụng APO

5. KẾT LUẬN VÀ PHƯƠNG HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi đã nghiên cứu và áp dụng thuật toán Arctic Puffin Optimization (APO) vào bài toán tối ưu hóa, trong đó việc kết hợp tìm kiếm địa phương linh hoạt giúp cải thiện đáng kể chất lượng lời giải, giảm nguy cơ mắc kẹt tại nghiệm cục bộ và tăng tốc độ hội tụ. Tuy nhiên, vẫn còn nhiều hướng nghiên cứu tiềm năng để nâng cao hiệu suất của APO. Một số hướng đi triển vọng bao gồm phát triển APO song song và phân tán nhằm xử lý các bài toán lớn; cũng như tích hợp với học máy để tự động điều chỉnh tham số theo từng loại bài toán. Với những hướng nghiên cứu này, thuật toán APO hứa hẹn trở thành một công cụ mạnh mẽ trong các bài toán tối ưu hóa phức tạp, giúp cải thiện hiệu suất và chất lượng lời giải trong nhiều lĩnh vực ứng dụng.

Các kết quả thực nghiệm cụ thể sẽ được công bố trong các báo cáo tiếp theo.

TÀI LIỆU THAM KHẢO

1. Wen-chuan Wang, Wei-can Tian, Dong-mei Xu, Hong-fei Zang, Arctic Puffin Optimization: A Bio-inspired Metaheuristic Algorithm for Solving Engineering Design Optimization, *Advances in Engineering Software*, **195**, (2024).
2. Wolpert DH, Macready WG, No free lunch theorems for optimization, *IEEE Transac Evolution Computat*, **1**, (1997), 67–82.
3. Hashim FA, Hussain K, Houssein EH, Mabrouk MS, Al-Atabany W, Archimedes optimization algorithm: A new metaheuristic algorithm for solving optimization problems, *Appl Intelligence*, **51**, (2021), 1531–1551.
4. Zong Woo G, Joong Hoon K, Loganathan GV, A new heuristic optimization algorithm: Harmony search, *Simulation*, **76.2**, (2001), 60–68.
5. Dorigo M, Birattari M, Stutzle T, Ant Colony Optimization, *IEEE Comput Intell Mag*, **1**, (2006), 28–39.
6. Stephen P. Boyd, Lieven Vandenberghe, *Convex Optimization*, Tập 1, Cambridge University Press, 127-136, United Kingdom (2004).

ABSTRACT

ARCTIC PUFFIN OPTIMIZATION IN OPTIMIZATION PROBLEMS

Bach Ngọc Vy¹, Phùng Thị Ly¹, Đinh Nguyễn Trọng Nghĩa^{1,*}

¹*Ho Chi Minh City University of Industry and Trade*

**Email: nghiadnt@hufi.edu.vn*

This paper introduces an optimization algorithm named "Arctic Puffin Optimization" (APO), inspired by the feeding and diving behaviors of puffins. The algorithm is designed to quickly and efficiently find optimal solutions for complex problems. APO utilizes parameters such as the maximum number of iterations, population size, search range, and a strategy to balance exploration (Aerial Search) and exploitation (Underwater Foraging). Furthermore, APO simulates the behavior of individuals in a population, evenly distributed across the search space and focusing on high-potential areas to improve search performance. The algorithm also minimizes the risk of being trapped in local optima by combining flexible exploration and exploitation. APO is proposed for practical applications such as nonlinear optimization and complex problems, helping to find optimal solutions with high reliability and fast processing time.

Keyword: Arctic Puffin Optimization, Optimization Problems, Aerial Search, Underwater Foraging

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Bạch Ngọc Vy

Số điện thoại liên hệ: 0382214577

TỐI ƯU HÓA THUẬT TOÁN ĐÀO TẠO ĐỘI BÓNG ĐÁ VÀ ỨNG DỤNG TRONG BÀI TOÁN TỐI ƯU HÓA

Trần Đình Hải Nguyên¹, Nguyễn Đình Đức¹, Đinh Nguyễn Trọng Nghĩa^{1,*}

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: nghiadnt@huit.edu.vn

Ngày nhận bài: 11/03/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Bài báo giới thiệu về thuật toán tối ưu hóa dựa trên hành vi con người (human-based algorithm) và quy trình huấn luyện trong môn thể thao bóng đá, tên là “Thuật toán đào tạo đội bóng đá” (Football Team Training Algorithm - FTTA). FTTA là một thuật toán metaheuristic mới lấy cảm hứng từ các giai đoạn huấn luyện đội bóng đá, bao gồm huấn luyện tập thể, huấn luyện nhóm và huấn luyện bổ sung cho cá nhân, nhằm cải thiện khả năng tìm kiếm và tối ưu hóa các giải pháp trong không gian tìm kiếm phức tạp. Các cầu thủ trong đội bóng, tương ứng với các cá thể trong thuật toán, hợp tác và cạnh tranh để tối ưu hóa hiệu suất toàn đội. Thuật toán này không chỉ nhấn mạnh vào việc tìm kiếm giải pháp tối ưu toàn cục mà còn tăng cường khả năng thoát khỏi các cực trị cục bộ trong các bài toán tối ưu hóa. Kết quả thực nghiệm cho thấy FTTA vượt trội so với các thuật toán như GA, PSO về tốc độ hội tụ và chất lượng nghiệm. Bài báo này phân tích chi tiết cơ chế hoạt động của FTTA, cung cấp ví dụ minh họa, và đề xuất các ứng dụng thực tế, đồng thời đưa ra hướng phát triển để mở rộng và cải thiện thuật toán.

Từ khóa: Thuật toán đào tạo đội bóng đá, Thuật toán dựa trên con người, Siêu thuật toán, Tối ưu hóa

1. GIỚI THIỆU

Tối ưu hóa là một lĩnh vực cốt lõi trong khoa học và kỹ thuật, đặc biệt trong việc giải quyết các bài toán phức tạp với không gian tìm kiếm đa chiều và phi tuyến tính. Các thuật toán metaheuristic, như Genetic Algorithm (GA), Particle Swarm Optimization (PSO) đã được sử dụng rộng rãi nhờ khả năng tìm kiếm nghiệm tối ưu trong những không gian này [1,2]. Nền tảng của các thuật toán này bắt nguồn từ các nghiên cứu ban đầu về hệ thống tự nhiên và nhân tạo [3]. Tuy nhiên, cả hai phương pháp đều có những hạn chế đáng kể. GA thường hội tụ sớm do thiếu đa dạng trong quá trình lai ghép và đột biến, dẫn đến việc bỏ sót các nghiệm tối ưu toàn cục [2]. PSO, với cơ chế di chuyển dựa trên thông tin tốt nhất của bầy

đàn, dễ mắc kẹt ở cực trị cục bộ khi không gian tìm kiếm có nhiều đỉnh [4]. Những hạn chế này đặt ra nhu cầu về một thuật toán mới, có khả năng cân bằng giữa khám phá (exploration) và khai thác (exploitation) một cách hiệu quả.

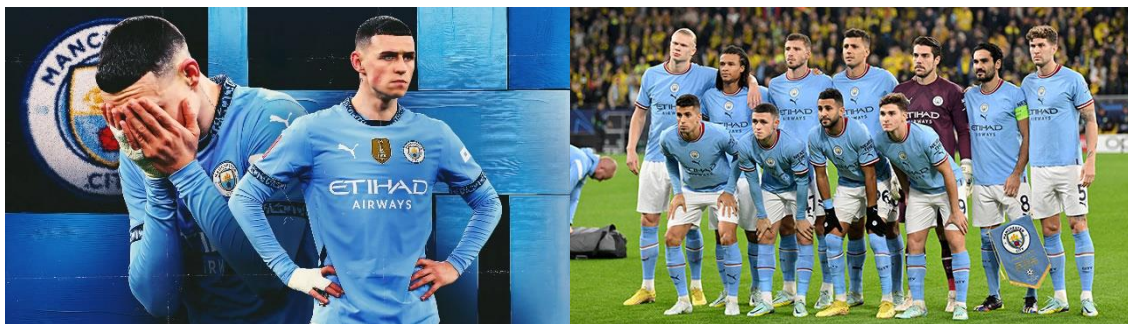
Để giải quyết vấn đề trên, bài báo đề xuất Football Team Training Algorithm (FTTA), một thuật toán metaheuristic lấy cảm hứng từ quá trình huấn luyện đội bóng đá. FTTA mô phỏng ba giai đoạn: huấn luyện tập thể (khám phá không gian tìm kiếm), huấn luyện nhóm (khai thác cục bộ), và huấn luyện bổ sung cá nhân (thoát khỏi cực trị cục bộ) [5]. Vấn đề nghiên cứu được đặt ra là: Làm thế nào để thiết kế một thuật toán tối ưu hóa vượt trội trong việc tìm kiếm nghiệm toàn cục và tránh mắc kẹt ở các cực trị cục bộ? Giả thuyết của nghiên cứu là FTTA, với cơ chế huấn luyện linh hoạt, có thể đạt hiệu quả cao hơn GA và PSO trong các bài toán tối ưu hóa phức tạp. Mục tiêu bao gồm: (1) phát triển và mô tả chi tiết FTTA, (2) đánh giá hiệu quả của FTTA qua thực nghiệm, và (3) khám phá ứng dụng trong các lĩnh vực như học máy, chuỗi cung ứng, và thiết kế kỹ thuật. Phạm vi nghiên cứu tập trung vào các bài toán tối ưu hóa liên tục với không gian tìm kiếm đa chiều.

So với GA (dựa trên tiến hóa) và PSO (dựa trên bầy đàn), FTTA mang lại sự linh hoạt nhờ kết hợp các chiến lược huấn luyện, giúp tối ưu hóa hiệu suất cả ở cấp độ cá nhân và tập thể. Bài báo trình bày cơ chế hoạt động của FTTA, kết quả thực nghiệm so sánh với GA và PSO, và các ứng dụng thực tế, từ đó làm rõ đóng góp khoa học của nghiên cứu. Các phần tiếp theo sẽ phân tích chi tiết thuật toán, ví dụ minh họa, ứng dụng, và hướng phát triển trong tương lai.

2. THUẬT TOÁN ĐÀO TẠO ĐỘI BÓNG ĐÁ

2.1. Đặc điểm của thuật toán

Thuật toán Football Team Training Algorithm (FTTA) thuộc nhóm các thuật toán tối ưu hóa metaheuristic lấy cảm hứng từ hành vi con người và tập trung vào việc mô phỏng quy trình hợp tác và cạnh tranh trong một đội bóng. Trong đội bóng, mỗi cầu thủ có hành vi và vai trò riêng nhưng vẫn phải phối hợp chặt chẽ với đồng đội để đạt mục tiêu chung. Thuật toán FTTA tái hiện quá trình này bằng cách xem mỗi cá thể trong không gian tìm kiếm như một cầu thủ, hoạt động dựa trên sự học hỏi, tương tác và cải thiện liên tục. Các cầu thủ và đội bóng được thể hiện trong Hình 1.



Hình 1. Cầu thủ và đội bóng

Trong quá trình huấn luyện tập thể, toàn đội bóng sẽ tham gia vào các hoạt động đồng bộ như chạy chiến thuật hay tập duyệt đội hình, giúp mở rộng không gian tìm kiếm. Đây

là giai đoạn mà các cá thể được đồng nhất về hướng đi, hỗ trợ thuật toán khám phá các khu vực tiềm năng mới trong không gian. Các cầu thủ trong đội (cá thể trong thuật toán) sẽ học hỏi từ hiệu suất chung, điều chỉnh hành vi để đạt được hiệu quả cao hơn.

Khi chuyển sang huấn luyện nhóm, đội bóng chia thành các nhóm nhỏ để cải thiện chi tiết hơn về chiến thuật và kỹ năng. Tương tự, FTTA cho phép các cá thể làm việc trong các vùng không gian nhỏ hơn, tập trung khai thác tiềm năng cục bộ. Giai đoạn này đảm bảo thuật toán không chỉ dừng lại ở việc khám phá mà còn tập trung vào khai thác hiệu quả những khu vực có triển vọng cao.

Cuối cùng, giai đoạn huấn luyện bổ sung tương tự như việc một cầu thủ tự rèn luyện cá nhân để khắc phục điểm yếu hoặc nâng cao kỹ năng đặc biệt. Trong FTTA, điều này được thực hiện bằng các bước nhảy ngẫu nhiên trong không gian tìm kiếm, giúp thuật toán thoát khỏi các cực trị cục bộ và tăng cường khả năng đa dạng hoá quần thể giải pháp.

Điểm nổi bật của FTTA là khả năng mô phỏng chính xác các hành vi học hỏi và phối hợp giữa các cá thể. Thuật toán này không chỉ cải thiện hiệu suất tổng thể mà còn duy trì tính linh hoạt cao, cho phép giải quyết các bài toán tối ưu hoá có không gian phức tạp, nhiều cực trị cục bộ và yêu cầu cân bằng giữa khám phá và khai thác [5,6].

2.2. Ý tưởng chính của thuật toán

Thuật toán đào tạo đội bóng (Football Team Training Algorithm - FTTA) được phát triển dựa trên việc mô phỏng quá trình huấn luyện trong môn thể thao bóng đá, nơi mà các cầu thủ liên tục học hỏi, cải thiện kỹ năng cá nhân và phối hợp với đồng đội để nâng cao hiệu suất tập thể. FTTA chuyển hoá quá trình này thành các chiến lược tối ưu hoá trong không gian tìm kiếm.

Mỗi cá thể trong thuật toán được xem như một cá nhân (cầu thủ) trong đội bóng, tương ứng với một giải pháp tiềm năng trong không gian tìm kiếm. Các cá thể này tương tác với nhau qua ba giai đoạn chính, tương ứng với quá trình huấn luyện trong bóng đá: huấn luyện tập thể (toàn đội), huấn luyện nhóm (chia nhóm để cải thiện chi tiết), và huấn luyện bổ sung (cải thiện cá nhân nhờ bước nhảy ngẫu nhiên). Sự phối hợp giữa các giai đoạn này giúp khai thác khả năng của cá nhân trong khi tối ưu hoá hiệu suất toàn đội.

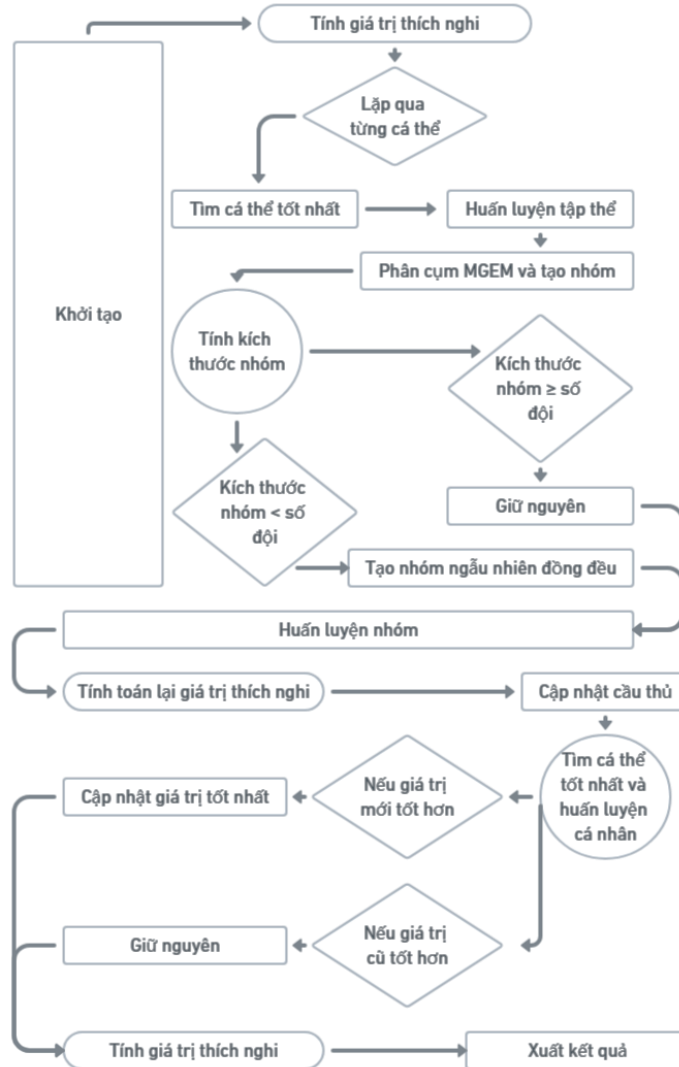
FTTA có khả năng cân bằng giữa hai quá trình quan trọng trong tối ưu hoá: khám phá để tìm kiếm các vùng giải pháp mới, và khai thác để tập trung tối ưu hoá tại các vùng tiềm năng. Quá trình huấn luyện tập thể giúp khám phá không gian toàn cục, trong khi huấn luyện nhóm khai thác tiềm năng trong các khu vực nhỏ, và huấn luyện bổ sung cung cấp cơ chế ngẫu nhiên giúp thoát khỏi cực trị cục bộ.

Trong bóng đá, quá trình huấn luyện bao gồm huấn luyện toàn đội nhằm nâng cao khả năng phối hợp, huấn luyện nhóm tăng cường chiến thuật trong từng khu vực cụ thể, và huấn luyện cá nhân để cải thiện kỹ năng đặc thù như chuyền bóng hoặc sút bóng. FTTA chuyển hoá những giai đoạn này thành các chiến lược tối ưu hoá trong không gian tìm kiếm, trong đó huấn luyện tập thể giúp mở rộng không gian, huấn luyện nhóm tối ưu hoá tập trung và huấn luyện bổ sung đảm bảo sự đa dạng của quần thể.

FTTA được thiết kế để giúp tìm giải pháp toàn cục trong các bài toán tối ưu hoá phức tạp. Bằng việc kết hợp tối ưu hoá cá nhân, tập thể và khai thác ngẫu nhiên, thuật toán đảm bảo tính hiệu quả trong không gian tìm kiếm và giúp thoát khỏi các cực trị cục bộ.

2.3. Mô hình toán học của thuật toán

Trong phần này sẽ trình bày phần mô hình toán học và thuật toán. Các mô hình toán học liên quan đến giai đoạn khởi tạo, giai đoạn huấn luyện nhóm, giai đoạn huấn luyện tập thể và giai đoạn huấn luyện bổ sung cho các cá nhân sẽ được cung cấp. Sau đó thuật toán FTTA sẽ được trình bày dưới dạng lưu đồ Hình 2.



Hình 2. Lưu đồ thuật toán FTTA

2.3.1. Giai đoạn khởi tạo

Cũng giống như các thuật toán metaheuristic khác, FTTA khởi tạo quần thể bằng cách tạo ngẫu nhiên các tác nhân N trong không gian vấn đề. Công thức được thể hiện như sau:

$$X_t^1 = lb + rand \times (ub - lb), i = 1, 2, \dots, N \quad (1)$$

Trong đó X_t^1 là tác nhân thứ nhất trong thời điểm t . ub và lb biểu thị ranh giới trên và dưới của không gian vấn đề. N biểu thị tổng số tác nhân được FTFA sử dụng trong quá trình tối ưu hóa. $rand$ biểu thị một số ngẫu nhiên giữa 0 và 1 giúp tăng khả năng phân tán các tác nhân trải đều trong không gian vấn đề. Quần thể ban đầu có thể được biểu diễn như sau:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,j} & \dots & X_{1,Dim} \\ X_{2,1} & X_{2,2} & \dots & X_{2,j} & \dots & X_{2,Dim} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ X_{i,1} & X_{i,2} & \dots & X_{i,j} & \dots & X_{i,Dim} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,j} & \dots & X_{N,Dim} \end{bmatrix} \quad (2)$$

Trong đó, $X_{i,j}$ biểu thị vị trí thứ j của tác nhân thứ i và Dim biểu thị kích thước chiều của bài toán đã cho.

2.3.2. Giai đoạn huấn luyện tập thể

Trong giai đoạn huấn luyện tập thể, FTFA phân ngẫu nhiên các tác nhân thành 4 loại: người theo dõi (followers), người khám phá (discoverers), người suy nghĩ (thinkers) và người dễ biến động (volatilities). Ngoài ra, tất cả các tác nhân đều chọn ngẫu nhiên một trong những loại này để cập nhật vị trí trong những lần lặp lại.

Đối với các tác nhân chọn vai trò người theo dõi, họ được chuyển ngẫu nhiên về phía tác nhân tối ưu trong mỗi chiều với công thức sau:

$$X_i^{t+1} = X_i^t + rand \times (X_{best} - X_i^t) \quad (3)$$

Trong đó X_{best} biểu thị tác nhân có độ phù hợp nhỏ nhất trong số tất cả các tác nhân (trong trường hợp bài toán tối thiểu hóa). t biểu thị số lần lặp hiện tại.

Đối với những người khám phá, các tác nhân này học hỏi từ cả những tác nhân tốt nhất và tệ nhất theo cách sau:

$$X_i^{t+1} = X_i^t + rand \times (X_{best} - X_i^t) - rand \times (X_{worst} - X_i^t) \quad (4)$$

Trong đó X_{worst} biểu thị tác nhân tệ nhất toàn cục.

Người suy nghĩ khác với người khám phá ở chỗ họ trực tiếp tìm hiểu khoảng cách giữa tác nhân tối ưu và tác nhân tệ nhất theo công thức sau:

$$X_i^{t+1} = X_i^t + rand \times (X_{best} - X_{worst}) \quad (5)$$

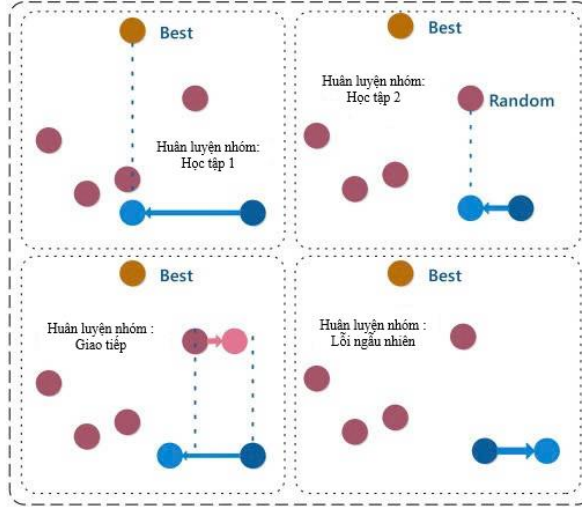
Thay vì học hỏi từ các tác nhân khác, người dễ biến động tự đào tạo theo cách được định nghĩa như sau:

$$X_i^{t+1} = X_i^t \times (1 + T(t)) \quad (6)$$

Trong đó $T(t)$ là một số ngẫu nhiên tuân theo phân phối t với số lần lặp t là bậc tự do.

2.3.3. Giai đoạn huấn luyện nhóm

Trong giai đoạn đào tạo nhóm, FTTA sẽ sử dụng các phương pháp phân cụm thích ứng kỳ vọng tối đa hỗn hợp Gauss (MGEM) để phân loại các tác nhân thành bốn danh tính: tiên đạo, tiên vệ, hậu vệ và thủ môn. Trong FTTA, có ít nhất 2 tác nhân của mỗi loại. Khi có một số loại tác nhân có ít hơn 2, FTTA sẽ thực hiện nhóm ngẫu nhiên đồng nhất tất cả các tác nhân. Sau khi nhóm hoàn tất, quá trình học tối ưu, học ngẫu nhiên và giao tiếp ngẫu nhiên được thực hiện ngẫu nhiên cho tất cả các tác nhân. Các phương pháp huấn luyện nhóm và sự di chuyển của các tác nhân trong giai đoạn này được thể hiện trong Hình 3.



Hình 3. Các phương pháp huấn luyện nhóm

Công thức cụ thể được thể hiện như sau:

Học tập tối ưu (Optimal learning) :

$$X_i^t = \begin{cases} X_i^t, & \text{nếu } rand > p_{study} \\ X_{best}, & \text{nếu } rand \leq p_{study} \end{cases} \quad (7)$$

Học tập ngẫu nhiên (Random learning) :

$$X_i^t = \begin{cases} X_i^t, & \text{nếu } rand > p_{study} \\ X_{random}, & \text{nếu } rand \leq p_{study} \end{cases} \quad (8)$$

Giao tiếp ngẫu nhiên (Random Communication) :

$$X_i^t = \begin{cases} X_i^t, & \text{nếu } rand > p_{comm} \\ X_{random}, & \text{nếu } rand \leq p_{comm} \end{cases} \quad (9)$$

Trong đó p_{study} biểu thị xác suất học tập. p_{comm} biểu thị xác suất giao tiếp. X_{random} biểu thị một cá nhân được chọn ngẫu nhiên. $rand$ là một số ngẫu nhiên theo phân phối chuẩn.

2.3.4. Giai đoạn huấn luyện bổ sung cá nhân

Sau khi hoàn thành giai đoạn huấn luyện nhóm, thuật toán FTTA sẽ đánh giá độ phù hợp (fitness) của từng tác tử. Đối với mỗi tác tử, giá trị fitness tốt hơn sẽ được chọn giữa tác tử con (offspring) và tác tử cha (parent). Tiếp theo, tác tử có fitness tối ưu nhất sẽ được huấn luyện bổ sung thông qua công thức sau.

$$X_t^{t+1} = X_t^t \times \left(1 + \left(1 - \frac{1}{t}\right) \times \text{Gauss} + \frac{1}{t} \times \text{Cauchy}\right) \quad (10)$$

Trong đó *Gauss* và *Cauchy* là các số ngẫu nhiên tuân theo phân phối Gaussian và Cauchy. Sau ba giai đoạn đào tạo, tất cả các tác nhân trước tiên thực hiện phán đoán điều kiện biên và phần vượt quá ranh giới sẽ được đặt lại về giá trị ranh giới. Sau đó, tính toán độ phù hợp được thực hiện để có được tác nhân tốt nhất cho đến khi điều kiện dừng được thỏa mãn và tác nhân tối ưu và độ phù hợp tối ưu được đưa ra.

Thuật toán : FTTA

```

Input:
lower_bound, upper_bound : giới hạn trên và dưới của không gian tìm kiếm
population : Tạo ngẫu nhiên pop_size cá thể trong [lower_bound, upper_bound] với
dim chiều
fitness : Tính objective_function cho mỗi cá thể
best_solution : Cá thể có fitness nhỏ nhất
best_fitness : Fitness của best_solution
worst_solution : Cá thể có fitness lớn nhất
fitness_history : Danh sách lưu lịch sử best_fitness
Vòng lặp chính
Vòng lặp t từ 1 đến max_iter
// Giai đoạn 1: Huấn luyện tập thể
    roles : Gán ngẫu nhiên vai trò cho mỗi cá thể: ['Follower', 'Discoverer',
'Thinker', 'Volatile']
    Với mỗi cá thể i trong population
        Nếu roles[i] = 'Follower'
            population[i] : population[i] + rand() * (best_solution -
population[i])
        Nếu roles[i] = 'Discoverer'
            population[i] : population[i] + rand() * (best_solution -
population[i]) - rand() * (worst_solution - population[i])
        Nếu roles[i] = 'Thinker'
            population[i] : population[i] + rand() * (best_solution -
worst_solution)
        Nếu roles[i] = 'Volatile'
            T : Lấy mẫu ngẫu nhiên từ phân phối t với bậc tự do t
            population[i] : population[i] * (1 + T)
            population[i] : Giới hạn trong [lower_bound, upper_bound]
            fitness[i] : objective_function(population[i])
    Kết thúc
// Giai đoạn 2: Huấn luyện nhóm
    
```

```
labels : Phân cụm population thành 4 nhóm (Striker, Midfielder, Defender,
Goalkeeper) bằng MGEM
Nếu bất kỳ nhóm nào có ít hơn 2 cá thể
    labels : Phân nhóm ngẫu nhiên thành 4 nhóm
Với mỗi cá thể i trong population
    strategy : Chọn ngẫu nhiên: ['optimal', 'random', 'communication']
    Nếu strategy = 'optimal' và rand() ≤ p_study
        population[i] : best_solution
    Nếu strategy = 'random' và rand() ≤ p_study
        population[i] : Cá thể ngẫu nhiên trong cùng nhóm với i
    Nếu strategy = 'communication' và rand() ≤ p_comm
        population[i] : Cá thể ngẫu nhiên trong cùng nhóm với i * (1 +
randn())

    population[i] : Giới hạn trong [lower_bound, upper_bound]
    fitness[i] : objective_function(population[i])

Kết thúc
//Giai đoạn 3: Huấn luyện bổ sung cá nhân
gauss : Lấy mẫu ngẫu nhiên từ phân phối Gaussian
cauchy : Lấy mẫu ngẫu nhiên từ phân phối Cauchy
best_idx : Chỉ số cá thể có fitness nhỏ nhất
best_solution : population[best_idx]
best_solution : best_solution * (1 + (1 - 1/t) * gauss + (1/t) * cauchy)
best_solution : Giới hạn trong [lower_bound, upper_bound]
best_fitness_new : objective_function(best_solution)
Nếu best_fitness_new < best_fitness
    population[best_idx] : best_solution
    fitness[best_idx] : best_fitness_new
    best_fitness : best_fitness_new

Kết thúc
worst_solution : Cá thể có fitness lớn nhất
fitness_history : Thêm best_fitness vào lịch sử

Kết thúc
Output:
Trả về best_solution, best_fitness, fitness_history
```

2.4. Ưu và nhược điểm của FTTA

Thuật toán Đào tạo Đội bóng đá (FTTA) sở hữu nhiều ưu điểm nổi bật, đặc biệt trong việc giải quyết các bài toán tối ưu hóa phức tạp. Nhờ mô phỏng linh hoạt ba giai đoạn huấn luyện (tập thể, nhóm, và bổ sung cá nhân), FTTA đạt được sự cân bằng hiệu quả giữa khám phá và khai thác, cho phép tìm kiếm nghiệm tối ưu toàn cục trong không gian đa chiều và phi tuyến tính. Cơ chế ngẫu nhiên hóa thông qua phân phối Gaussian và Cauchy trong giai đoạn huấn luyện bổ sung giúp FTTA dễ dàng thoát khỏi cực trị cục bộ, nâng cao khả năng hội tụ so với các thuật toán như GA, PSO. Tuy nhiên, FTTA cũng tồn tại một số hạn chế, bao gồm sự phụ thuộc vào việc điều chỉnh các tham số như xác suất học tập (p_{study}) và giao tiếp (p_{comm}), đòi hỏi kinh nghiệm hoặc thử nghiệm để đạt hiệu quả tối ưu. Ngoài ra, khi số lượng

cá thể hoặc vòng lặp tăng, yêu cầu tính toán của FTTA có thể trở nên đáng kể, đặc biệt trong các bài toán quy mô lớn. Những đặc điểm này nhấn mạnh tiềm năng của FTTA đồng thời chỉ ra các hướng cải thiện cần thiết trong tương lai.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Để đánh giá hiệu quả của thuật toán FTTA, nhóm thực hiện thực nghiệm trên một loạt các hàm thử nghiệm phổ biến bao gồm F1, F2, F3, F4, F5 và F6. Các hàm này được lựa chọn nhằm kiểm tra khả năng của thuật toán trong việc tìm kiếm cực trị toàn cục trong các không gian tìm kiếm phức tạp và đa dạng.

Dưới đây là bảng mô tả chi tiết các hàm thử nghiệm, bao gồm công thức, phạm vi biến, và số lượng biến sử dụng trong các thí nghiệm. Tất cả các kết quả kiểm chứng trong nghiên cứu này đều được thực hiện trên máy tính có cấu hình Intel Core i5-12400F, 32GB RAM và tốc độ xử lý 2.50 GHz, sử dụng phiên bản Python 3.12.6. Các thử nghiệm được tiến hành trong môi trường này nhằm đảm bảo tính nhất quán và độ tin cậy của các kết quả đạt được từ thuật toán FTTA.

Bảng 1. Công thức toán học và thông số của các hàm thử nghiệm được sử dụng trong thực nghiệm

Hàm	Công thức	Phạm vi	Số chiều	Mục tiêu
F1	$\sum_{i=1}^n x_i^2$	[-100, 100]	10, 30, 50	0
F2	$10d + \sum_{i=1}^d [x_i^2 - 10\cos(2\pi x_i)]$	[-5.12, 5.12]	10, 30, 50	0
F3	$\sum_{i=1}^{d-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	[-5, 10]	10, 30, 50	0
F4	$-a \exp\left(-b \sqrt{\frac{1}{d} \sum_{i=1}^d x_i^2}\right) - \exp\left(\frac{1}{d} \sqrt{\sum_{i=1}^d \cos(cx_i)}\right) + a + \exp(1)$	[-32.768, 32.768]	10, 30, 50	0
F5	$\sum_{i=1}^d \frac{x_i^2}{4000} - \prod_{i=1}^d \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	[-600, 600]	10, 30, 50	0
F6	$\sum_{i=1}^d x_i^2 + \left(\sum_{i=1}^d 0.5ix_i\right)^2 + \left(\sum_{i=1}^d 0.5ix_i\right)^2$	[-5, 10]	10, 30, 50	0

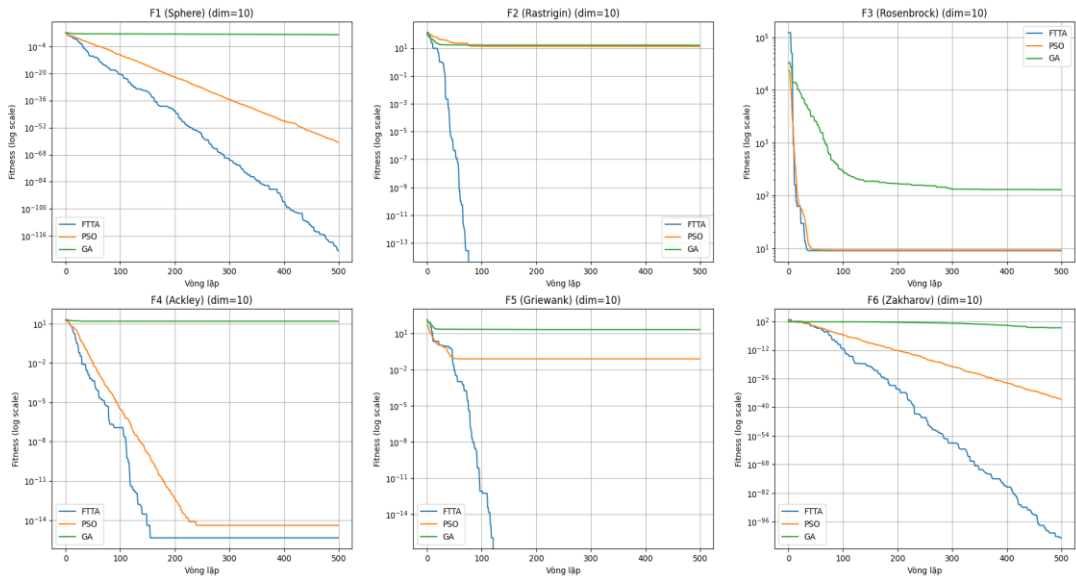
3.1. Thiết lập thực nghiệm

Các bài toán tối ưu hóa phức tạp được chọn để thử nghiệm bao gồm 6 hàm chuẩn với độ phức tạp và tính chất khác nhau, được kiểm tra trên các số chiều 10, 30, 50 nhằm đảm bảo tính tổng quát của kết quả. Các tham số của thuật toán FTTA, PSO và GA được thiết lập như sau: kích thước quần thể (popSize) là 50, số chiều (dim) được thử nghiệm với các giá trị 10, 30 và 50, số vòng lặp (Iter) là 500. Đối với FTTA, các hằng số xác suất học tập cá nhân (p_{study}) và giao tiếp nhóm (p_{comm}) được đặt là 0.3; đối với PSO, các hằng số quán tính (w) là 0.5, học cá nhân ($c1$) là 1.5 và học toàn cục ($c2$) là 1.5; đối với GA, xác suất lai ghép (crossover_rate) là 0.8 và xác suất đột biến (mutation_rate) là 0.1. Mỗi bài toán được chạy 10 lần để giảm thiểu ảnh hưởng của yếu tố ngẫu nhiên và đảm bảo tính khách quan của kết quả. Kết quả độ thích nghi trung bình qua các lần chạy được sử dụng để đánh giá hiệu quả của các thuật toán.

3.2. Kết quả thực nghiệm

Dưới đây là kết quả thực nghiệm trên các hàm thử nghiệm với số chiều lần lượt là 10, 30 và 50.

3.2.1. Thực nghiệm với số chiều là 10



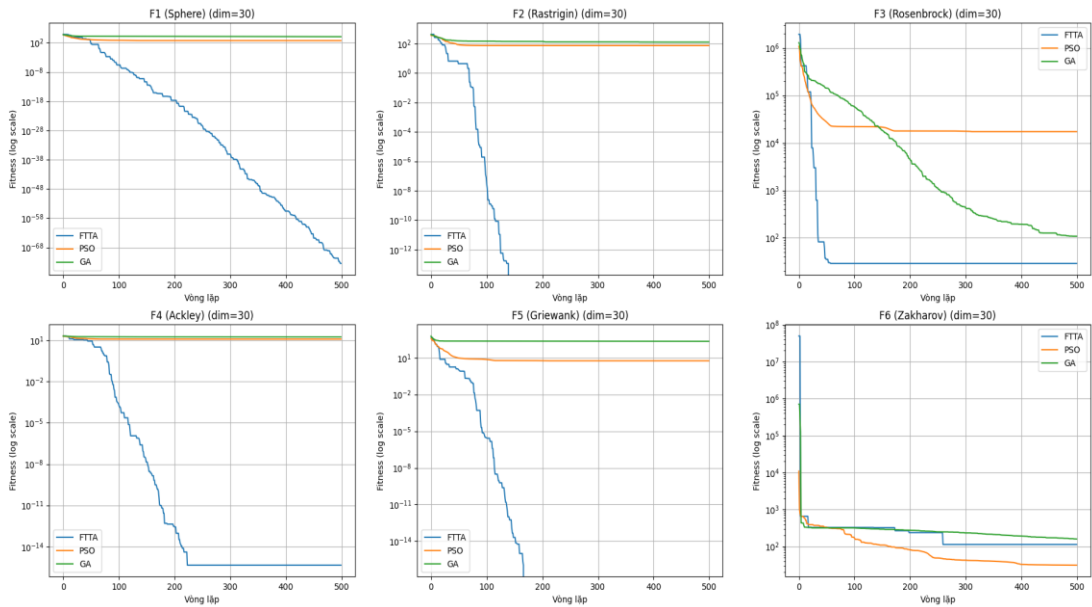
Hình 4. Biểu đồ kết quả thực nghiệm với số chiều là 10

Bảng 2. Bảng thống kê kết quả sau mỗi 100 vòng lặp với số chiều là 10

	Algorithm	100	200	300	400	500
1	FTTA	3.48e-20	2.59e-42	2.74e-73	5.17e-96	1.97e-123
	PSO	2.00e-10	2.30e-24	9.05e-38	8.90e-51	5.99e-65
	GA	1.23e+03	9.13e+02	6.53e+02	4.26e+02	2.34e+02
	FTTA	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00

	Algorithm	100	200	300	400	500
2	PSO	1.39e+01	1.39e+01	1.39e+01	1.39e+01	1.39e+01
	GA	3.32e+01	3.32e+01	3.31e+01	3.31e+01	3.31e+01
3	FTTA	8.78e+00	8.78e+00	8.78e+00	8.78e+00	8.78e+00
	PSO	3.06e+02	3.05e+02	3.05e+02	3.05e+02	3.05e+02
	GA	2.73e+03	1.43e+02	1.08e+02	1.07e+02	1.07e+02
4	FTTA	4.16e-11	4.44e-16	4.44e-16	4.44e-16	4.44e-16
	PSO	1.09e-04	1.93e-11	4.00e-15	4.00e-15	4.00e-15
	GA	1.07e+01	1.06e+01	1.06e+01	1.06e+01	1.06e+01
5	FTTA	9.33e-12	0.00e+00	0.00e+00	0.00e+00	0.00e+00
	PSO	8.61e-02	8.61e-02	8.61e-02	8.61e-02	8.61e-02
	GA	1.11e+01	1.07e+01	1.05e+01	1.05e+01	1.05e+01
6	FTTA	3.27e-06	9.76e-32	6.85e-61	1.74e-80	2.49e-100
	PSO	3.50e-06	3.97e-13	1.17e-21	1.10e-29	4.97e-37
	GA	3.45e+01	1.05e+01	1.26e+00	1.02e-01	1.79e-02

3.2.2. Thực nghiệm với số chiều là 30



Hình 5. Biểu đồ kết quả thực nghiệm với số chiều là 30

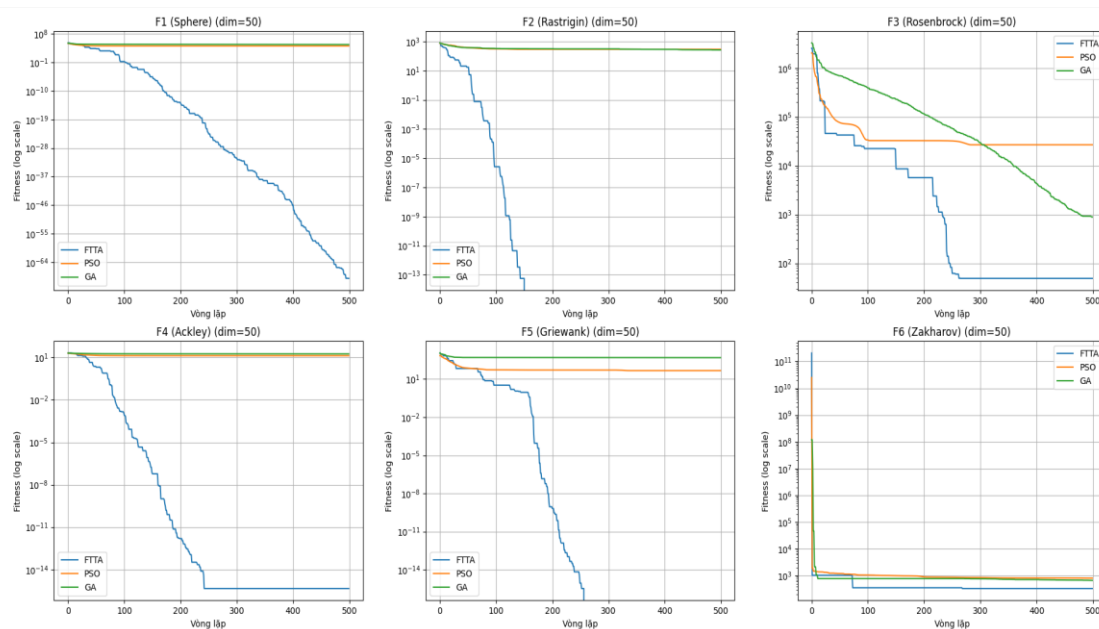
Bảng 3. Bảng thống kê kết quả sau mỗi 100 vòng lặp với số chiều là 30

	Algorithm	100	200	300	400	500
1	FTTA	1.22e-03	7.95e-20	1.15e-39	8.00e-60	1.49e-77
	PSO	5.17e+02	5.06e+02	5.06e+02	5.05e+02	5.05e+02

	Algorithm	100	200	300	400	500
	GA	1.85e+04	1.71e+04	1.59e+04	1.46e+04	1.35e+04
2	FTTA	3.59e-09	0.00e+00	0.00e+00	0.00e+00	0.00e+00
	PSO	1.78e+02	1.78e+02	1.78e+02	1.78e+02	1.78e+02
	GA	1.61e+02	1.53e+02	1.52e+02	1.47e+02	1.44e+02
3	FTTA	2.90e+01	2.90e+01	2.90e+01	2.90e+01	2.90e+01
	PSO	3.76e+02	3.55e+02	3.47e+02	3.47e+02	3.47e+02
	GA	1.57e+05	2.36e+04	1.59e+03	3.99e+02	2.77e+02
4	FTTA	1.03e-01	3.24e-12	4.44e-16	4.44e-16	4.44e-16
	PSO	1.06e+01	1.06e+01	1.06e+01	1.06e+01	1.06e+01
	GA	1.73e+01	1.73e+01	1.73e+01	1.73e+01	1.73e+01
5	FTTA	2.81e-01	0.00e+00	0.00e+00	0.00e+00	0.00e+00
	PSO	1.72e+01	1.44e+01	1.37e+01	1.37e+01	1.37e+01
	GA	1.57e+02	1.57e+02	1.57e+02	1.57e+02	1.57e+02
6	FTTA	8.47e+02	1.31e+02	1.31e+02	1.31e+02	1.31e+02
	PSO	2.31e+02	9.31e+01	5.20e+01	2.47e+01	2.11e+01
	GA	1.62e+02	1.41e+02	1.15e+02	8.68e+01	6.69e+01

3.2.3. Thực nghiệm với số chiều là 50

Hình 6. Biểu đồ kết quả thực nghiệm với số chiều là 50



Bảng 4. Bảng thống kê kết quả sau mỗi 100 vòng lặp với số chiều là 50

	Algorithm	100	200	300	400	500
1	FTTA	8.52e+02	1.12e-07	1.15e-26	6.91e-43	4.93e-58
	PSO	1.01e+04	1.01e+04	1.00e+04	1.00e+04	9.91e+03
	GA	6.98e+04	6.70e+04	6.44e+04	6.19e+04	5.94e+04
2	FTTA	5.22e-07	0.00e+00	0.00e+00	0.00e+00	0.00e+00
	PSO	2.70e+02	2.68e+02	2.68e+02	2.68e+02	2.68e+02
	GA	3.92e+02	3.80e+02	3.72e+02	3.72e+02	3.64e+02
3	FTTA	5.04e+03	3.89e+03	2.01e+03	2.01e+03	1.72e+03
	PSO	4.67e+04	4.45e+04	4.45e+04	4.44e+04	4.44e+04
	GA	3.90e+05	1.35e+05	3.47e+04	7.05e+03	2.02e+03
4	FTTA	9.27e-02	9.00e-09	4.44e-16	4.44e-16	4.44e-16
	PSO	1.26e+01	1.26e+01	1.26e+01	1.26e+01	1.26e+01
	GA	1.88e+01	1.86e+01	1.85e+01	1.85e+01	1.84e+01
5	FTTA	2.52e-02	0.00e+00	0.00e+00	0.00e+00	0.00e+00
	PSO	6.32e+01	6.12e+01	6.12e+01	6.12e+01	6.12e+01
	GA	6.55e+02	6.50e+02	6.47e+02	6.43e+02	6.39e+02
6	FTTA	3.30e+03	4.69e+02	4.69e+02	4.69e+02	4.69e+02
	PSO	1.04e+03	8.89e+02	7.36e+02	6.44e+02	5.99e+02
	GA	5.77e+02	5.57e+02	5.30e+02	4.99e+02	4.66e+02

3.3. Đánh giá thuật toán từ kết quả thực nghiệm

Nghiên cứu này đánh giá hiệu quả của thuật toán FTTA thông qua các bài toán tối ưu hóa phức tạp, với kích thước 10, 30 và 50 chiều, và thuật toán được chạy 10 lần trên mỗi trường hợp để đảm bảo độ tin cậy. Kết quả thực nghiệm cho thấy:

3.3.1. Khả năng hội tụ và tìm kiếm cực trị toàn cục

FTTA thể hiện tốc độ hội tụ nhanh và ổn định, đặc biệt trên các bài toán có không gian giải pháp đơn giản như hàm F1 và F5. Thuật toán đạt được các giá trị tối ưu hoặc gần tối ưu trên tất cả các kích thước biến, nhờ vào giai đoạn huấn luyện tập thể giúp tinh chỉnh vị trí cá thể và cải thiện độ chính xác trong quá trình tối ưu hóa.

3.3.2. Khả năng xử lý các hàm có nhiều cực trị cục bộ

Đối với các bài toán chứa nhiều cực trị cục bộ như F2 và F4, FTTA đạt được các giá trị tối ưu tốt nhờ cơ chế huấn luyện bổ sung cá nhân. Giai đoạn này, sử dụng phân phối Gaussian và Cauchy, cho phép thuật toán thoát khỏi các cực trị cục bộ và tăng cường khả năng khám phá không gian giải pháp.

3.3.3. So sánh với PSO và GA

So với PSO và GA riêng lẻ, FTТА cho thấy độ chính xác và tốc độ hội tụ vượt trội, đặc biệt trên các bài toán như F3 và F6. Thuật toán nhanh chóng đạt được các giá trị tốt nhất, nhờ vào sự kết hợp của huấn luyện tập thể, huấn luyện nhóm và huấn luyện bổ sung, giúp tinh chỉnh vị trí các cá thể và giảm số vòng lặp cần thiết để đạt đến tối ưu.

3.3.4. Khả năng duy trì sự đa dạng của quần thể

FTТА duy trì tốt sự đa dạng trong quần thể, nhờ vào giai đoạn huấn luyện bổ sung cá nhân với cơ chế ngẫu nhiên hóa. Cơ chế này giúp thuật toán tránh hội tụ sớm và không bị mắc kẹt tại các cực trị cục bộ, đặc biệt quan trọng với các bài toán có nhiều cực trị cục bộ như F2 và F4. Sự đa dạng này cải thiện khả năng khám phá không gian giải pháp, đảm bảo hiệu quả tối ưu hóa.

4. ỨNG DỤNG TRONG BÀI TOÁN TỐI ƯU HÓA

Thuật toán Đào tạo Đội bóng đá (Football Team Training Algorithm - FTТА) là một thuật toán metaheuristic lấy cảm hứng từ quá trình huấn luyện và phối hợp trong thể thao, đặc biệt là bóng đá. FTТА mô phỏng cách các cầu thủ trong đội bóng cải thiện kỹ năng cá nhân, hợp tác với đồng đội và điều chỉnh chiến thuật nhằm đạt hiệu suất tối ưu. Trong bối cảnh tối ưu hóa, thuật toán này có thể được ứng dụng rộng rãi để giải quyết các bài toán tìm kiếm nghiệm tốt nhất trong không gian có độ phức tạp cao.

Định nghĩa 1: Ánh xạ FTТА vào bài toán tối ưu hóa: Bài toán tối ưu hóa thường liên quan đến việc tìm giá trị tốt nhất của một hàm mục tiêu trong một tập hợp giới hạn các phương án khả thi, FTТА mô phỏng quá trình cải thiện hiệu suất của một đội bóng, trong đó mỗi cầu thủ đại diện cho một cá thể trong quần thể, các chiến lược huấn luyện tương ứng với cơ chế cập nhật nghiệm trong thuật toán, và quá trình hợp tác nhóm mô phỏng sự khai thác và khám phá không gian tìm kiếm.

Định nghĩa 2: Cấu trúc thuật toán FTТА trong tối ưu hóa: FTТА bao gồm ba giai đoạn chính, huấn luyện tập thể nơi các cá thể được chỉnh vị trí dựa trên hiệu suất chung, tương tự cách một đội bóng thực hành cùng nhau để cải thiện lối chơi tổng thể, huấn luyện nhóm nơi các cá thể được chia thành nhóm nhỏ, tập trung tối ưu hóa trong một vùng cụ thể của không gian tìm kiếm, và huấn luyện bổ sung nơi một số cá thể thực hiện điều chỉnh độc lập để tránh bị kẹt tại cực trị cục bộ.

Định nghĩa 3: Ứng dụng FTТА vào các bài toán tối ưu hóa cụ thể: FTТА có thể được áp dụng vào nhiều bài toán tối ưu hóa khác nhau, chẳng hạn như tối ưu hóa tham số mô hình học máy, tối ưu hóa chuỗi cung ứng, tối ưu hóa lập lịch trình, và tối ưu hóa thiết kế kỹ thuật.

Định nghĩa 4: Đặc điểm nổi bật của FTТА trong tối ưu hóa: nhấn mạnh khả năng cân bằng giữa khai thác và khám phá nhờ cơ chế huấn luyện nhóm và cá nhân, giúp duy trì sự linh hoạt trong việc tìm kiếm nghiệm, khả năng thích ứng với bài toán phi tuyến tính, xử lý các bài toán có ràng buộc và biến số phức tạp, cùng với cơ chế huấn luyện bổ sung giúp giảm nguy cơ mắc kẹt tại cực trị cục bộ, tăng cường khả năng thoát khỏi các vùng tối ưu cục bộ.

Các bước thực hiện ứng dụng thuật toán FTТА vào bài toán tối ưu hóa

Đầu tiên, xác định bài toán tối ưu hóa với hàm mục tiêu $f(x)$ cần tối ưu, các ràng buộc và không gian tìm kiếm là tập hợp giá trị khả thi cho nghiệm.

Tiếp theo, khởi tạo quần thể ban đầu, mỗi cá thể là một "cầu thủ" có vị trí trong không gian tìm kiếm, được tạo ngẫu nhiên hoặc theo phân bố để phủ tốt không gian, rồi gán giá trị ban đầu dựa trên hàm mục tiêu.

Sau đó, tiến hành huấn luyện tập thể, tức giai đoạn khám phá, các cá thể cập nhật trạng thái dựa trên thông tin toàn đội, mở rộng phạm vi tìm kiếm để tránh cực trị cục bộ, điều chỉnh vị trí theo phản hồi từ hệ thống.

Giai đoạn huấn luyện nhóm, hay giai đoạn khai thác, diễn ra sau đó, quần thể chia thành nhóm nhỏ, mỗi nhóm tìm nghiệm tối ưu trong vùng cục bộ, dùng thông tin từ nhóm khác để điều chỉnh, tăng khai thác cục bộ mà vẫn giữ tính đa dạng.

Tiếp đến là huấn luyện bổ sung, một số cá thể nhảy ngẫu nhiên, áp dụng nhiều Gaussian hoặc Cauchy để tăng tính đa dạng, tránh bẫy cực trị cục bộ và tìm nghiệm tốt hơn.

Sau mỗi vòng lặp, đánh giá quần thể bằng cách tính giá trị hàm mục tiêu, chọn cá thể tốt nhất để huấn luyện tiếp, cập nhật trạng thái theo chiến lược tối ưu hóa.

Thuật toán lặp lại đến khi đạt điều kiện dừng, như số vòng lặp tối đa, sai số nhỏ hơn ngưỡng, hoặc không còn cải thiện đáng kể.

Cuối cùng, chọn cá thể có giá trị tối ưu nhất làm nghiệm cuối, phân tích và kiểm tra để đảm bảo đáp ứng yêu cầu bài toán.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi giới thiệu đến bạn đọc thuật toán Đào tạo Đội bóng đá (FTTA) - một phương pháp tối ưu hóa metaheuristic hiệu quả, lấy cảm hứng từ quá trình huấn luyện bóng đá để cải thiện khả năng tìm kiếm nghiệm tối ưu. Với cơ chế linh hoạt trong quá trình học tập và hợp tác giữa các cầu thủ, FTTA có thể giải quyết nhiều bài toán tối ưu hóa phức tạp trong các lĩnh vực khác nhau. Nhờ khả năng kết hợp giữa chiến lược cá nhân hóa và hợp tác nhóm, FTTA không chỉ đạt được kết quả tối ưu nhanh chóng mà còn giảm thiểu nguy cơ hội tụ sớm. Bên cạnh đó, chúng tôi cũng trình bày một số bài toán trong các lĩnh vực khác nhau đề xuất ứng dụng FTTA để giải quyết, với hy vọng sẽ đem lại kết quả chính xác và cải thiện hiệu quả so với các phương pháp tối ưu hóa khác. Mặc dù chưa được thử nghiệm cụ thể, phương pháp mới này hứa hẹn sẽ mang lại giải pháp tối ưu cho các bài toán phức tạp này.

Các kết quả thực nghiệm cụ thể sẽ được công bố trong các báo cáo tiếp theo.

TÀI LIỆU THAM KHẢO

1. Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942-1948.
2. Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley.
3. Holland, J. H. (1975). Adaptation in natural and artificial systems. University of Michigan Press.
4. Yang, X. S. (2010). Nature-inspired metaheuristic algorithms (2nd ed.). Luniver Press.

5. Li, X., Wang, J., Zhang, Y., & Chen, H. (2024). An improved football team training algorithm for global optimization. *Bioengineering*, 9(7), 419.
6. Zhang, Y., Li, X., Wang, J., & Chen, H. (2023). A novel hybrid metaheuristic algorithm for global optimization. *Expert Systems with Applications*, 225, 120987.

ABSTRACT

OPTIMIZATION OF THE FOOTBALL TEAM TRAINING ALGORITHM AND ITS APPLICATION IN OPTIMIZATION PROBLEMS

Tran Dinh Hai Nguyen¹, Nguyen Dinh Duc¹, Dinh Nguyen Trong Nghia^{1,*}

¹*Ho Chi Minh City University of Industry and Trade*

*Email: nghiadnt@huit.edu.vn

This paper introduces a new kind of optimization algorithm inspired by how humans behave and how football teams train calling it the 'Football Team Training Algorithm' (FTTA). Think of FTТА as a fresh metaheuristic approach that mimics the different stages of football practice like team drills, small group exercises, and individual skills training to improve how well it searches for solutions in tricky problem spaces. Basically, the players in a football team in this algorithm represent individual solutions, working together and competing to boost the overall performance. The cool part is, FTТА isn't just about hunting for the absolute best answer; it's also pretty good at avoiding getting stuck on less-than-ideal solutions. Tests show that FTТА beats some popular algorithms like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) when it comes to how quickly it finds good solutions and how good those solutions are. The paper goes into detail about how FTТА works, offers some simple examples, and discusses real-world uses. It also suggests ways to develop and improve the algorithm in the future.

Keywords: Football Team Training Algorithm, Human-based Algorithm, Metaheuristic Algorithm, Optimization

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Trần Đình Hải Nguyên

Số điện thoại liên hệ: 0797468285

THUẬT TOÁN KEPLER VÀ ỨNG DỤNG CHO CÁC BÀI TOÁN TỐI ƯU

Nguyễn Thế Hữu^{1,*}, Hồ Nguyễn Linh Chi¹, Nguyễn Thị Ngọc¹, Huỳnh Thị Bích Ngọc¹, Lê Thị Như Quỳnh¹, Đinh Nguyễn Trọng Nghĩa¹

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: huunthe@huit.edu.vn

Ngày nhận bài 11/3/2025 ; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Bài báo này giới thiệu thuật toán tối ưu Kepler (Kepler Optimization Algorithm - KOA), một phương pháp metaheuristic mới lấy cảm hứng từ các định luật chuyển động hành tinh của Johannes Kepler. Thuật toán mô phỏng quỹ đạo chuyển động của các hành tinh quanh một ngôi sao trung tâm nhằm tìm kiếm nghiệm tối ưu cho bài toán. Cơ chế của KOA dựa trên sự cân bằng giữa lực hấp dẫn và lực quán tính để định hướng quá trình tìm kiếm. Trong khuôn khổ bài nghiên cứu này, nhóm tác giả trình bày định nghĩa, công thức hoạt động của thuật toán và thực hiện một ví dụ minh họa đơn giản trên một hàm mục tiêu ba biến với số vòng lặp giới hạn. Ngoài ra, bài báo cũng đề xuất quy trình áp dụng KOA vào bài toán tối ưu tổng quát. Kết quả minh họa cho thấy tiềm năng của thuật toán trong việc tìm kiếm nghiệm tối ưu, làm cơ sở cho các nghiên cứu và thực nghiệm sâu hơn trong tương lai.

Từ khóa: Thuật toán tối ưu Kepler, tối ưu hóa, phương pháp metaheuristic, định luật chuyển động hành tinh.

1. GIỚI THIỆU

Tối ưu hóa đóng vai trò quan trọng trong khoa học máy tính và toán học, giúp tìm kiếm nghiệm tốt nhất cho các bài toán có không gian tìm kiếm rộng lớn [1]. Các phương pháp tối ưu hóa metaheuristic ngày càng được phát triển để giải quyết những bài toán phức tạp trong khoa học, kỹ thuật và trí tuệ nhân tạo, đặc biệt khi các phương pháp truyền thống như lập trình tuyến tính, gradient descent hoặc quy hoạch bậc hai gặp khó khăn do không gian tìm kiếm phi tuyến tính và có nhiều cực trị [2][3].

Một trong những thuật toán tối ưu hóa mới đáng chú ý là Kepler Optimization Algorithm (KOA), lấy cảm hứng từ các định luật chuyển động hành tinh của Johannes Kepler. Thuật toán này được thiết kế nhằm khai thác các đặc tính vật lý của chuyển động quỹ đạo để tìm kiếm nghiệm tối ưu hiệu quả hơn so với nhiều phương pháp metaheuristic khác

[4]. Cụ thể, KOA mô phỏng cách các hành tinh di chuyển xung quanh một ngôi sao trung tâm theo quỹ đạo hình elip, chịu tác động của lực hấp dẫn, từ đó xây dựng một cơ chế tìm kiếm cân bằng giữa hai giai đoạn khám phá (*exploration*) và khai thác (*exploitation*) [5].

Các nghiên cứu trước đây đã chỉ ra rằng một số thuật toán metaheuristic như Particle Swarm Optimization (PSO) và Genetic Algorithm (GA) thường gặp phải vấn đề hội tụ sớm hoặc mắc kẹt tại cực trị cục bộ trong các không gian tìm kiếm có độ phức tạp cao [6][7]. Trong khi đó, KOA khắc phục những hạn chế này bằng cách sử dụng cơ chế điều chỉnh quỹ đạo linh hoạt, đảm bảo sự đa dạng trong quần thể tìm kiếm và cải thiện khả năng hội tụ [4]. Điều này giúp thuật toán trở thành một công cụ tối ưu hóa mạnh mẽ, đặc biệt đối với các bài toán phi tuyến tính và không khả vi.

Bài báo này trình bày tổng quan về thuật toán KOA và ý tưởng áp dụng thuật toán vào bài toán tối ưu hóa.

2. THUẬT TOÁN KEPLER

2.1. Đặc điểm của thuật toán Kepler

Thuật toán Kepler optimization algorithm (KOA) được xây dựng dựa trên ba định luật chuyển động do nhà thiên văn học Johannes Kepler đề xuất nhằm mô tả quỹ đạo của các hành tinh trong Hệ Mặt Trời. Trước thời Kepler, quan niệm phổ biến cho rằng các hành tinh chuyển động theo quỹ đạo tròn quanh Trái Đất hoặc Mặt Trời. Tuy nhiên, Kepler đã chứng minh rằng quỹ đạo của các hành tinh thực chất là elip, trong đó Mặt Trời nằm tại một trong hai tiêu điểm của elip.

Một đặc điểm quan trọng của thuật toán KOA là nó mô hình hóa chuyển động của hành tinh bằng cách tính toán vị trí của nó theo thời gian dựa trên các định luật vật lý. Định luật thứ hai của KOA cho thấy rằng vận tốc của hành tinh thay đổi theo khoảng cách của nó với Mặt Trời: hành tinh di chuyển nhanh hơn khi ở gần Mặt Trời và chậm hơn khi ở xa. Ngoài ra, định luật thứ ba cung cấp mối quan hệ giữa chu kỳ quỹ đạo và bán trục lớn của quỹ đạo, giúp xác định thời gian cần thiết để hành tinh hoàn thành một vòng quay quanh Mặt Trời [8].

Các nguyên tắc tối ưu hóa chính của thuật toán Kepler dựa trên ba định luật này, cho phép xác định chính xác vị trí và vận tốc của hành tinh tại một thời điểm bất kỳ.

2.2. Ý tưởng và Mô hình Toán học của Thuật toán Kepler

Thuật toán tối ưu dựa trên hấp dẫn hành tinh hoạt động theo nguyên tắc mô phỏng chuyển động của các hành tinh trong hệ Mặt Trời [4]. Mỗi hành tinh đại diện cho một lời giải trong không gian tìm kiếm, và chúng di chuyển dưới tác động của lực hấp dẫn từ Mặt Trời – điểm tương ứng với lời giải tốt nhất hiện tại [9].

Trước tiên, các hành tinh (lời giải) được tạo ra một cách ngẫu nhiên trong một phạm vi xác định, theo công thức [4]:

$$X_i^j = X_{\{i,low\}}^j + rand[0,1] \times (X_{\{i,up\}}^j - X_{\{i,low\}}^j) \quad (1)$$

Ở đây, mỗi lời giải X_i có tọa độ trong không gian nhiều chiều. Giá trị $X_{i,low}^j$ và $X_{i,up}^j$ lần lượt là giới hạn dưới và giới hạn trên của tọa độ thứ j của lời giải i , trong khi $rand [0,1]$ là một số ngẫu nhiên trong khoảng $[0,1]$, giúp tạo ra các giá trị khác nhau cho từng lời giải.

Mỗi hành tinh có quỹ đạo riêng dựa trên các tham số như độ lệch tâm e và chu kỳ quỹ đạo T . Độ lệch tâm quỹ đạo (e) cho mỗi vật thể thứ i được khởi tạo bằng công thức: $e_i = rand[0,1]$, $i = 1, \dots, N$; Trong đó $rand[0,1]$ là giá trị ngẫu nhiên được tạo ra trong khoảng $[0,1]$. Cuối cùng, chu kỳ quỹ đạo (T) cho mỗi vật thể thứ i được khởi tạo bằng công thức:

$T_i = |r|$, $i = 1, \dots, N$; Trong đó r là số được tạo ra ngẫu nhiên dựa trên phân phối chuẩn.

Những tham số trên ảnh hưởng đến cách hành tinh di chuyển trong không gian tìm kiếm, từ đó tác động đến việc thuật toán khai thác và khám phá các nghiệm tiềm năng. Dựa trên định luật vạn vật hấp dẫn của Newton, lực hấp dẫn giữa một hành tinh và Mặt Trời được tính theo công thức [4]:

$$F_i^g(t) = e_i \times \mu(t) \times \frac{M_S \times m_i}{R_i^2 + \varepsilon} + r_1 \quad (2)$$

Ở đây, M_S là khối lượng của Mặt Trời (tức là điểm tối ưu hiện tại), m_i là khối lượng của hành tinh thứ i , ε là một số rất nhỏ để tránh lỗi toán học khi chia cho 0, $\mu(t)$ là hằng số hấp dẫn thay đổi theo thời gian, và r_1 là một giá trị nhiễu ngẫu nhiên giúp thuật toán tránh rơi vào cực tiểu cục bộ, R_i là khoảng cách giữa hành tinh i và Mặt Trời và được định nghĩa là [4]:

$$R_{i(t)} = ||X_{S(t)} - X_{i(t)}||_2 = (x + a)^n = \sum_{j=1}^d (X_{Sj}(t) - X_{ij}(t))^2 \quad (3)$$

Trong đó $||X_{S(t)} - X_{i(t)}||_2$ biểu thị khoảng cách Euclid giữa các chiều của X_S và các chiều của X_i . Khối lượng của Mặt trời và đối tượng i tại thời điểm t được tính toán đơn giản bằng cách sử dụng đánh giá độ phù hợp như sau [4]:

$$M_S = \frac{fit_S(t) - worst(t)}{\sum_{k=1}^N (fit_k(t) - worst(t))} \quad (4)$$

$$m_i = \frac{fit_i(t) - worst(t)}{\sum_{k=1}^N (fit_k(t) - worst(t))} \quad (5)$$

Trong đó $fits(t) = best(t) = \min(fit_k(t))$, $k \in 1, 2, \dots, N$; $worst(t) = \max(fit_k(t))$, $k \in 1, 2, \dots, N$; r_2 là một số được tạo ngẫu nhiên giữa 0 và 1 để phân kỳ các giá trị khối lượng cho các hành tinh khác nhau; $\mu(t)$ là một hàm giảm theo cấp số nhân theo thời gian (t) để kiểm soát độ chính xác của tìm kiếm và được định nghĩa như sau: $\mu(t) = \mu_0 \times \exp(-\gamma \frac{t}{T_{max}})$; trong đó γ là hằng số, μ_0 là giá trị ban đầu, và t và T_{max} là số lần lặp hiện tại và số lần lặp tối đa tương ứng.

Lực hấp dẫn khiến các hành tinh thay đổi vận tốc theo thời gian, giúp chúng di chuyển về phía điểm tối ưu mà không bị rơi ngay vào một vùng kém chất lượng. Công thức tính vận tốc của mỗi hành tinh như sau [4]:

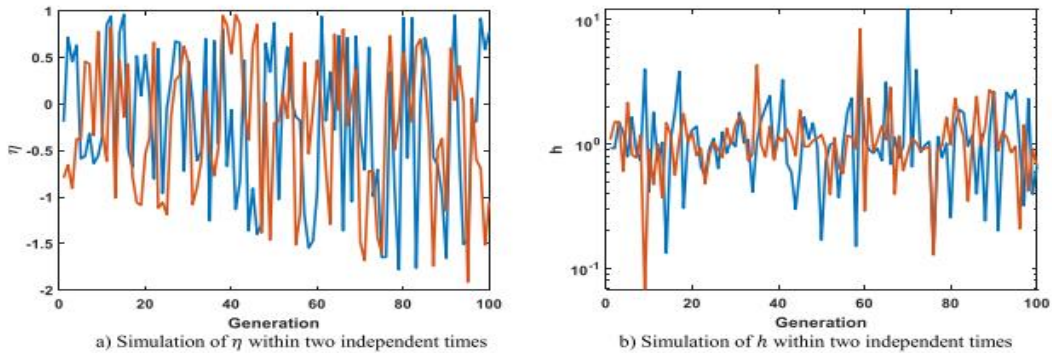
$$V_i(t) = \begin{cases} \ell \times (2r_4 \vec{X}_i - \vec{X}_b) + \ell \times (\vec{X}_a - \vec{X}_b) + (1 - R_{i-norm}(t)) \\ \times \mathcal{F} \times \vec{U}_1 \times r_5 \times (\vec{X}_{i,up} - \vec{X}_{i,low}), \text{if } R_{i-norm}(t) \leq 0.5 \\ r_4 \times \mathcal{L} \times (\vec{X}_a - \vec{X}_i) + (1 - R_{i-norm}(t)) \\ \times \mathcal{F} \times \vec{U}_2 \times r_5 \times (r_4 \vec{X}_{i,up} - \vec{X}_{i,low}), \text{Else} \end{cases} \quad (6)$$

Sau khi tính vận tốc, thuật toán sẽ cập nhật vị trí của mỗi hành tinh theo công thức:

$$X_i(t+1) = X_i(t) + F \times V_i(t) + (F_i^g(t) + |r|) \times U \times (X_s(t) - X_i(t)) \quad (7)$$

Trong đó, F là một biến điều chỉnh hướng di chuyển, U là một ma trận ngẫu nhiên giúp tạo sự đa dạng trong cách các hành tinh di chuyển, và $|r|$ là một giá trị nhiễu giúp thuật toán tránh hội tụ quá sớm.

Một trong những thách thức lớn của thuật toán tối ưu là tránh bị mắc kẹt trong các cực tiểu cục bộ (tức là dừng lại ở một nghiệm chưa phải tốt nhất). Để giải quyết vấn đề này, thuật toán có cơ chế điều chỉnh động lực học của các hành tinh. Cụ thể, khi một hành tinh có xu hướng bị rơi vào một điểm cực tiểu cục bộ, thuật toán sẽ bổ sung các yếu tố nhiễu hoặc thay đổi hướng di chuyển để tăng cơ hội khám phá các vùng có nghiệm tốt hơn.



Hình 1. Mô tả các giá trị của h dựa trên hệ số η trong hai thời điểm độc lập.

Khi các hành tinh ở gần Mặt trời, KOA sẽ tập trung vào tối ưu hóa toán tử khai thác; khi Mặt trời ở xa, KOA sẽ tối ưu hóa toán tử thăm dò. Các quy tắc này phụ thuộc vào giá trị của tham số điều chỉnh h , thay đổi dần theo thời gian hiện tại như thể hiện trong Hình 1. Khi giá trị này lớn, toán tử thăm dò được sử dụng để mở rộng khoảng cách quỹ đạo hành tinh so với Mặt trời; ngược lại, khi giá trị này nhỏ, toán tử khai thác được sử dụng để khai thác các vùng xung quanh giải pháp tốt nhất cho đến nay nếu khoảng cách giữa Mặt trời và các hành tinh là nhỏ. Mô hình toán học của nguyên lý này được mô tả như sau [4]:

$$\begin{aligned} \vec{X}_i(t+1) = \vec{X}_i(t) \times \vec{U}_1 + (1 - \vec{U}_1) \times & \left(\frac{\vec{X}_i(t) + \vec{X}_s + \vec{X}_a(t)}{3.0} + h \times \left(\frac{\vec{X}_i(t) + \vec{X}_s + \vec{X}_a(t)}{3.0} - \right. \right. \\ & \left. \left. \vec{X}_b(t) \right) \right) \end{aligned} \quad (8)$$

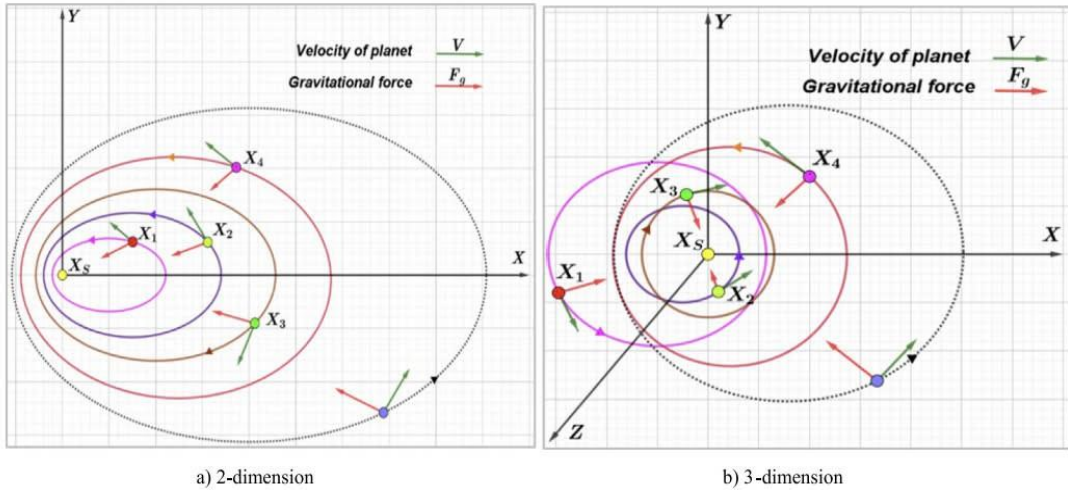
Trong đó h là một hệ số thích ứng để kiểm soát khoảng cách giữa Mặt trời và hành tinh hiện tại tại thời điểm t , như được định nghĩa: $h = 1/e^{\eta r}$, r là một số được tạo ngẫu nhiên trên cơ sở phân phối chuẩn, η là một hệ số giảm tuyến tính từ 1 đến -2 , như được định nghĩa: $\eta = (a_2 - 1) \times r_4 + 1$, a_2 là tham số điều khiển tuần hoàn giảm dần từ -1 đến -2 đối với

T chu kỳ trong toàn bộ quá trình tối ưu hóa như được định nghĩa: $a_2 = -1 - 1 \times \left(\frac{t\% \frac{T_{max}}{T}}{\frac{T_{max}}{T}} \right)$.

Nhờ vào cách mô phỏng quỹ đạo hành tinh xung quanh Mặt Trời, thuật toán không chỉ tìm kiếm trong không gian giải pháp một cách hiệu quả mà còn có khả năng hội tụ nhanh về nghiệm tối ưu. Đây là lý do tại sao thuật toán này có thể được áp dụng vào nhiều bài toán tối ưu khác nhau trong thực tế.

2.3. Quy trình thuật toán Tối ưu hóa Kepler (KOA)

Phần này trình bày một thuật toán tối ưu hóa mới lấy cảm hứng từ các định luật chuyển động của Kepler. Trong thuật toán đề xuất, định luật thứ nhất của Kepler được mô phỏng bằng cách coi Mặt Trời là giải pháp tối ưu nhất, trong khi các hành tinh (hoặc các đối tượng) quay quanh nó theo các quỹ đạo hình elip tưởng tượng, đại diện cho không gian tìm kiếm. Trong KOA, các hành tinh (các nghiệm ứng viên) có vị trí khác nhau so với Mặt Trời (nghiệm tốt nhất) tại nhiều thời điểm khác nhau, do đó giúp cải thiện quá trình thăm dò và khai thác không gian tìm kiếm.



Hình 1. Các vị trí có thể có trong không gian 2 chiều và 3 chiều

Hình 2 minh họa cách vị trí của một vật thể, khối lượng của nó, lực hấp dẫn giữa vật thể và Mặt Trời, cũng như vận tốc quỹ đạo của nó thay đổi để xác định vị trí xung quanh nghiệm tốt nhất, tức là Mặt Trời. Hình này cũng thể hiện rằng phần lớn các vật thể quay theo chiều kim đồng hồ. Ngoài ra, nó còn mô tả sự thay đổi vị trí của người tìm kiếm trong không gian ba chiều.

Giống như các thuật toán metaheuristic dựa trên quần thể khác, KOA bắt đầu quá trình tìm kiếm với một tập hợp ban đầu gồm các đối tượng có quỹ đạo ngẫu nhiên. Mỗi đối tượng được khởi tạo với vị trí ngẫu nhiên của nó trong quỹ đạo ở giai đoạn này. Sau khi đánh giá hàm mục tiêu của tập hợp ban đầu, KOA chạy theo từng bước lặp cho đến khi đạt điều kiện dừng. Trong nghiên cứu này, thuật ngữ "thời gian" được sử dụng thay cho "số lần lặp" vì nó phù hợp hơn với lý thuyết hệ mặt trời và vũ trụ học.

Trong quá trình tối ưu hóa, KOA áp dụng các quy tắc sau:

- Chu kỳ quỹ đạo của một hành tinh (nghiệm ứng viên) được chọn ngẫu nhiên theo phân phối chuẩn.
- Độ lệch tâm của quỹ đạo hành tinh được chọn ngẫu nhiên trong khoảng từ 0 đến 1.
- Giá trị hàm mục tiêu của một nghiệm được tính dựa trên hàm mục tiêu của bài toán.
- Nghiệm tốt nhất trong mỗi lần lặp đóng vai trò là ngôi sao trung tâm (Mặt Trời).
- Khoảng cách giữa Mặt Trời và hành tinh thay đổi theo thời gian hiện tại.

Phần còn lại của mục này trình bày mô hình toán học của KOA. Độ phức tạp thời gian của các bước trong thuật toán này là $O(NT_{\max})$, trong đó N là kích thước quần thể và $T_{\max}$ là tiêu chí dừng của KOA dựa trên số lần đánh giá hàm mục tiêu tối đa. Về mặt lý thuyết, KOA có thể được coi là một thuật toán tối ưu hóa toàn cục vì nó bao gồm cả giai đoạn thăm dò và khai thác. Mô tả chi tiết về các bước trong KOA được trình bày như sau.

Bước 1: Một số lượng hành tinh bằng với kích thước quần thể N sẽ được phân bố ngẫu nhiên trong không gian d chiều, đại diện cho các biến quyết định của bài toán tối ưu hóa

Bước 2: Tính toán giá trị fitness cho từng đối tượng trong quần thể dựa trên hàm mục tiêu của bài toán. Giá trị fitness càng tốt thể hiện nghiệm càng tối ưu. Chọn đối tượng có fitness tốt nhất làm nghiệm tốt nhất (được gọi là Mặt Trời).

Bước 3: Tính khoảng cách Euclide và lực hấp dẫn giữa Mặt Trời và các đối tượng, tính vận tốc của các đối tượng.

Bước 4: Sinh ngẫu nhiên hai giá trị r và r_1 trong khoảng $[0, 1]$. Nếu $r < r_1$ thì cập nhật khoảng cách giữa đối tượng và Mặt Trời theo công thức (8). Nếu $r \geq r_1$: Cập nhật vị trí mới của đối tượng theo các công thức (6) và (7).

Bước 5: Cập nhật lại vị trí các đối tượng bằng chiến lược ưu tiên theo công thức [4]:

$$\vec{X}_{i,new}(t+1) = \begin{cases} \vec{X}_i(t+1), & \text{if } f(\vec{X}_i(t+1)) \leq f(\vec{X}_i(t)) \\ \vec{X}_i(t), & \text{Else} \end{cases} \quad (9)$$

KOA đã được thử nghiệm trên nhiều bài toán tối ưu hóa và cho thấy hiệu suất cao hơn so với các thuật toán khác.

Thuật toán KOA (Kepler Optimization Algorithm)

Mô tả tóm tắt Thuật toán Tối Ưu hóa Kepler (KOA) là một thuật toán tối ưu hóa dựa trên các nguyên tắc chuyển động hành tinh của Johannes Kepler. Thuật toán mô phỏng quá trình hành tinh chuyển động quanh Mặt Trời và sử dụng các thuộc tính quỹ đạo để tìm kiếm nghiệm tối ưu.

Thuật toán 1. Minh họa KOA.

Input: Hàm mục tiêu $fitness()$

Quy mô quần thể hành tinh Pop
Số lần lặp tối đa $MaxIt$
Số chiều không gian tìm kiếm Dim
Biên dưới của khoảng tìm kiếm lb
Biên trên của khoảng tìm kiếm ub
Hằng số hấp dẫn G
Hệ số điều chỉnh vận tốc α
Hệ số cập nhật vị trí λ
Sinh số ngẫu nhiên trong khoảng $[0,1]$

Mảng lưu nghiệm tốt nhất X_b

Hằng số

Hệ số suy giảm hấp dẫn G_{decay}

Giá trị tối thiểu của hấp dẫn G_{min}

Hệ số ảnh hưởng của vận tốc β

Ngưỡng dừng tối ưu ϵ

Khởi tạo

```
for each i to Pop do
    for j = 1 to dim do
        positions[i, j] = lb + (ub - lb) * rand()
        eccentricities[i, j] = rand()
        periods[i, j] = rand() * (Tmax - Tmin) + Tmin
    end for
end for
// Đánh giá giá trị fitness cho từng hành tinh
Evaluate_Fitness(positions)
// Sắp xếp quần thể theo giá trị fitness từ tốt nhất đến kém nhất
Sort_Pop(positions, fitness)
// Lấy nghiệm tốt nhất làm Mặt Trời
GbestPosition = positions[0]
GbestScore = fitness(GbestPosition)
```

Tối ưu hóa

```
For i = 1 to MaxIter do
    //Cập nhật hằng số hấp dẫn G theo công thức suy giảm
    G = G_min + (G - G_min) * exp(-G_decay * i)
    //Lặp qua từng hành tinh trong quần thể
    For j = 1 to Pop do
        For z = 1 to Dim do
            //Tính toán lực hấp dẫn dựa trên quần thể tốt nhất
            F_grav = G * (Xb[z] - positions[j, z]) / (abs(Xb[z] - positions[j, z]) + epsilon)
            //Tạo số ngẫu nhiên theo phân phối Gaussian
            A1 = rngnormal(loc=0, scale=1.2, size=1)
            //Cập nhật vị trí của hành tinh
            positions[j, z] = positions[j, z] + lambda * F_grav * A1
        End for
    End for
    //Áp dụng bước cập nhật vận tốc và vị trí dựa trên Keplerian motion
    For j = 1 to Pop do
        For z = 1 to Dim do
            V[j, z] = beta * V[j, z] + alpha * (Xb[z] - positions[j, z]) * rngnormal(0,1,1)
            positions[j, z] = positions[j, z] + V[j, z]
        End for
    End for
```



```
//Giới hạn vị trí trong khoảng [lb, ub]
For j = 0 to Pop do
    positions[j] = clip(positions[j], lb, ub)
End for
//Sắp xếp quần thể theo giá trị fitness
Sort(positions, fitness)
//Cập nhật nghiệm tốt nhất
If (fitness(positions[0]) <= GbestScore) then
    GbestScore = fitness(positions[0])
    GbestPosition = positions[0]
End if
End for
```

Điều kiện dừng:

- Số vòng lặp đạt giá trị tối đa *MaxIter*.
- Giá trị fitness của nghiệm tốt nhất không thay đổi đáng kể giữa các vòng lặp liên tiếp.

OUTPUT:

- *GbestPosition*: Vị trí tốt nhất tìm được.
- *GbestScore*: Giá trị độ thích nghi tốt nhất.

3. VÍ DỤ MINH HỌA

Lấy ví dụ tìm min value hàm tối ưu được định nghĩa bởi.

$$f(x, y, z) = \left(\frac{x}{4}\right)^2 + \left(\frac{y}{2}\right)^2 + z^2$$

Trong đó x, y, z sẽ nằm trong khoảng [-2, 2].

Bước đầu tiên khởi tạo quần thể hành tinh ngẫu nhiên. Ở đây bài toán có 3 biến, số lượng hành tinh trong quần thể là 5 hành tinh. Khởi tạo tính độ thích nghi (fitness) và sắp xếp quần thể theo độ thích nghi ta được Bảng 1.

Bảng 1. Tập hợp các hành tinh sau bước khởi tạo

Index	x	y	z	Fitness
0	-0,65824232	-0,0283785	0,6609058	0,46407795
1	-1,2373331	1,17590729	0,4085809	0,608314913
2	-1,52654585	-1,5884353	0,1403411	0,796123685
3	1,46048492	0,26735396	-1,733868	3,157479742

4	1,45920498	1,39152343	1,7326529	3,619250535
---	------------	------------	-----------	-------------

Từ Bảng 1, có thể nhìn rõ và xác định được fitness tốt nhất làm nghiệm tốt nhất được gọi là Mặt trời. Áp công thức (6) và (7) ta sẽ tạo ra vị trí mới từ vị trí ban đầu, ta thu được đối tượng mới mang giá trị $[-1,572724; -1,733868; 1.67990989]$. Tuy nhiên, độ thích nghi của đối tượng này cao hơn so với đối tượng tương ứng nên sẽ không được thay thế vào.

Tiếp tục sẽ thực hiện tuần tự các bước trên sinh thêm các đối tượng mới. Lần này thu được đối tượng mới tốt hơn $[0,27644894; 1,21311035; 0,5685963]$. Với giá trị thích nghi nhỏ hơn đối tượng tương ứng trước đó trong quần thể và đối tượng mới này sẽ được thay thế vào. Sẽ thực hiện lại các bước trên cho đến khi duyệt hết quần thể. Nếu thỏa điều kiện, đối tượng mới sẽ được thay thế và ngược lại. Tuy nhiên thuật toán tự động bỏ qua đối tượng tốt nhất là Mặt trời được Bảng 2 như sau:

Bảng 2. Tập hợp các hành tinh sau khi hoàn thành vòng lặp 1 (Hành tinh in đậm là hành tinh không thay đổi).

Index	x	y	z	Fitness
0	-0,38907402	-1,2760347	0,1369611	0,435285647
1	-0,65824232	-0,0283785	0,6609058	0,46407795
2	-1,2373331	1,17590729	0,4085809	0,608314913
3	0,27644894	1,21311035	0,5685963	0,695987393
4	1,46048492	0,26735396	-1,733868	3,157479742

Tiến hành lặp lại 100 vòng lặp sẽ được một tập hợp các vị trí các hành tinh như trong Bảng 3 dưới đây.

Bảng 3. Tập hợp các hành tinh sau 100 vòng lặp.

Index	x	y	z	Fitness
0	-2,7259076E-12	-2,2764606E-11	1,3261035E-11	1,8845416E-23
1	4,8124278E-12	-7,2598675E-11	-2,5798797E-11	3,8572045E-23
2	6,0840693E-12	-2,2407535E-11	5,2620057E-11	7,3702838E-23
3	4,1814145E-12	-1,2311035E-11	9,5685963E-11	4,3476498E-23
4	1,4642890E-12	8,9290102E-11	-5,5166216E-11	5,4764272E-23

4. ỨNG DỤNG CHO CÁC BÀI TOÁN TỐI ƯU

Bài toán tối ưu hóa có thể được phát biểu tổng quát như sau:

$$\min_{x \in R^n} f(x) \text{ với } x \in S \quad (10)$$

Trong đó: $f(x)$ là hàm mục tiêu cần tối ưu (có thể là cực tiểu hoặc cực đại), S là không gian tìm kiếm chứa các giá trị khả thi của biến x , ràng buộc có thể được biểu diễn dưới dạng bất đẳng thức hoặc đẳng thức:

$$g_i(x) \leq 0, h_j(x) = 0 \quad (11)$$

Tùy vào bản chất của bài toán, tối ưu hóa có thể chia thành các loại:

- Tối ưu không ràng buộc: Không có giới hạn nào đặt lên biến x .
- Tối ưu có ràng buộc: Các ràng buộc được áp dụng trên tập giá trị của x .
- Tối ưu đơn mục tiêu và đa mục tiêu: Một hoặc nhiều hàm mục tiêu có thể được tối ưu hóa đồng thời.

Minh họa thuật toán Kepler với bài toán tối ưu hàm Rosenbrock

Hàm Rosenbrock được định nghĩa như sau:

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2 \quad (12)$$

Đây là một bài toán tối ưu hóa phi tuyến tính thường được dùng để kiểm tra hiệu suất của các thuật toán tối ưu.

Quá trình áp dụng thuật toán Kepler:

Bước 1: Khởi tạo quần thể. Tạo một tập hợp các điểm ban đầu trong không gian tìm kiếm bằng cách chọn ngẫu nhiên các giá trị trong một phạm vi xác định.

Bước 2: Tính toán lực hấp dẫn. Xác định lực hấp dẫn giữa các cá thể và điểm trọng tâm bằng cách đánh giá giá trị của hàm Rosenbrock tại từng điểm.

Bước 3: Cập nhật vị trí. Dịch chuyển từng cá thể dọc theo quỹ đạo Kepler bằng cách tính toán bán trục lớn và độ lệch tâm, giúp điều chỉnh khoảng cách và hướng di chuyển.

Bước 4: Chọn lựa cá thể tốt nhất. Sau mỗi lần cập nhật, các cá thể có giá trị hàm mục tiêu thấp hơn sẽ được giữ lại, còn các cá thể có giá trị cao hơn sẽ bị thay thế.

Bước 5: Hội tụ. Sau nhiều vòng lặp, thuật toán dần thu hẹp phạm vi tìm kiếm và hội tụ đến điểm tối ưu với giá trị tối thiểu của hàm Rosenbrock.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Thuật toán tối ưu Kepler (KOA) là một phương pháp metaheuristic mới, lấy cảm hứng từ các nguyên lý vật lý về chuyển động hành tinh. Với cơ chế cân bằng giữa khả năng khám phá không gian tìm kiếm và khả năng khai thác nghiệm tiềm năng, KOA thể hiện tiềm năng trong việc giải các bài toán tối ưu hóa. Qua ví dụ minh họa đơn giản, bài báo bước đầu cho thấy tính khả thi của KOA trong việc tìm kiếm nghiệm gần tối ưu. Trong tương lai, thuật toán có thể được cải tiến và đánh giá kỹ hơn thông qua các thử nghiệm trên nhiều bài toán chuẩn, cũng như so sánh với các thuật toán tối ưu hóa hiện đại khác để xác định phạm vi ứng dụng và hiệu quả thực tiễn.

TÀI LIỆU THAM KHẢO

- [1]. Talbi, E. G. *Metaheuristics: from design to implementation*. John Wiley & Sons. (2009)
- [2]. Yang, X. S. *Nature-inspired metaheuristic algorithms*. Luniver Press. (2010)
- [3]. Holland, J. H. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press. (1975)
- [4]. Abdel-Basset, M., Mohamed, R., Azeem, S. A. A., & Jameel, M. *Kepler optimization algorithm (KOA): A novel physics-based metaheuristic inspired by planetary motion laws*. *Swarm and Evolutionary Computation*, 80, 101293. (2023)
- [5]. Kepler, J. *Astronomia Nova*. Heidelberg: University Library. (1609)
- [6]. Kennedy, J., & Eberhart, R. *Particle swarm optimization*. *Proceedings of ICNN'95 - International Conference on Neural Networks*, 4, 1942–1948. (1995)
- [7]. Koza, J. R. *Genetic programming: On the programming of computers by means of natural selection*. MIT Press. (1992)
- [8]. A. Gilchrist, Johannes, *Kepler: the sky as a retinal image*, *Perception*, vol. 43, no. 12, pp. 1283–1285, (2014).
- [9]. T. M. Hemmerling, M. Wehbe, C. Zaouter, R. Taddei, and J. Morse, *The Kepler intubation system*, *Anesthesia & Analgesia*, vol. 114, no. 3, pp. 590–594, (2012).

ABSTRACT

KEPLER ALGORITHM AND ITS APPLICATIONS TO OPTIMIZATION PROBLEMS

Nguyen The Huu^{1,*}, Ho Nguyen Linh Chi¹, Nguyen Thi Ngoc¹, Huynh Thi Bich Ngoc¹,

Le Thi Nhu Quynh¹, Dinh Nguyen Trong Nghia¹

¹ Ho Chi Minh City University of Industry and Trade

*Email: huunt@huit.edu.vn

This paper introduces the Kepler Optimization Algorithm (KOA), a novel metaheuristic inspired by Johannes Kepler's laws of planetary motion. The algorithm simulates the orbital trajectories of planets around a central star to search for optimal solutions. KOA operates based on the balance between gravitational and inertial forces to guide the search process. Within the scope of this student-level research, the paper presents the definition, mathematical formulation of KOA, and provides a simple illustrative example on a three-variable objective function with a limited number of iterations. Additionally, a step-by-step approach for applying KOA to general optimization problems is proposed. The illustrative results indicate the algorithm's potential in solving optimization tasks, paving the way for more extensive experimental studies in future work.

Keywords: Kepler Optimization Algorithm, optimization, metaheuristic method, laws of planetary motion.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thế Hữu

Số điện thoại liên hệ: 0932163743

THUẬT TOÁN TỐI ƯU DILOPHOSAURUS (DOA): NÂNG CAO HIỆU SUẤT CỦA THUẬT TOÁN T-REX (TROA)

Nguyễn Thế Hữu^{1,*}, Lều Công Trung Hiếu¹, Đặng Nguyên Hoài¹, Đinh Nguyễn Trọng Nghĩa¹

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: huunthe@huit.edu.vn

Ngày nhận bài: 12/3/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Bài báo này đề xuất giới thiệu một thuật toán meta-heuristic mới mang tên Dilophosaurus Optimization Algorithm (DOA), lấy cảm hứng từ đặc điểm và chiến lược săn mồi của loài khủng long Dilophosaurus. DOA mô phỏng sự kết hợp giữa tốc độ, khả năng theo dõi con mồi, và hành vi tấn công linh hoạt của loài khủng long này để tìm kiếm và khai thác không gian giải pháp hiệu quả giúp nâng cao khả năng tránh bẫy cục bộ và tối ưu hóa hiệu suất tìm kiếm. DOA được đánh giá trên một tập hợp các bài toán tối ưu hóa chuẩn và so sánh với các thuật toán meta-heuristic phổ biến khác. Kết quả thực nghiệm cho thấy DOA đạt hiệu suất tối ưu tốt, thậm chí vượt trội trong một số trường hợp, chứng minh tính hiệu quả và tiềm năng ứng dụng của thuật toán trong các bài toán tối ưu hóa phức tạp.

Từ khóa: Meta-heuristic, Tối ưu, TROA, DOA

MỞ ĐẦU

Trong bối cảnh khoa học và công nghệ phát triển, các bài toán tối ưu hóa phức tạp, chẳng hạn như lập lịch sản xuất, thiết kế mạng, hoặc tối ưu năng lượng, đòi hỏi các thuật toán meta-heuristic hiệu quả để tìm nghiệm tối ưu trong không gian tìm kiếm đa chiều [1]. Các thuật toán này, lấy cảm hứng từ hiện tượng tự nhiên và hành vi động vật, đã chứng minh khả năng khai thác không gian tìm kiếm linh hoạt [2]. Tuy nhiên, nhiều thuật toán vẫn gặp khó khăn trong việc tránh cực trị cục bộ, cân bằng giữa khai phá (exploration) và khai thác (exploitation), hoặc thích nghi với các bài toán có số chiều lớn và ràng buộc phi tuyến.

Các nghiên cứu gần đây trong lĩnh vực meta-heuristic đã phát triển nhiều thuật toán lấy cảm hứng từ hành vi động vật, chẳng hạn như Particle Swarm Optimization (PSO) [3], Genetic Algorithm (GA) [4], Grey Wolf Optimizer (GWO) [5], Whale Optimization Algorithm (WOA) [6], và Coati Optimization Algorithm (COA) [7]. PSO, dựa trên hành vi đàn chim, có ưu điểm là tốc độ hội tụ nhanh nhưng thường hội tụ sớm, đặc biệt trong các bài

toán multimodal [3]. GA, mô phỏng cơ chế di truyền, phù hợp với các bài toán đa mục tiêu nhưng tốn tài nguyên tính toán, làm hạn chế ứng dụng trong các hệ thống thời gian thực [4]. GWO và WOA, lấy cảm hứng từ hành vi săn mồi của sói và cá voi, đạt hiệu quả cao trong các bài toán tối ưu hóa liên tục, nhưng thiếu đa dạng quần thể dẫn đến khó khăn trong việc khám phá không gian tìm kiếm rộng [5, 6]. COA, mô phỏng gấu mũi dài, cho thấy tiềm năng trong các bài toán thực tế như tối ưu hóa năng lượng, nhưng cần cải tiến để cân bằng khai phá và khai thác [7]. Gần đây, T-Rex Optimization Algorithm (TROA) [8], mô phỏng chiến lược săn mồi của khủng long T-Rex, đã được đề xuất và chứng minh hiệu quả trong các không gian tìm kiếm lớn. Tuy nhiên, TROA dễ mắc kẹt tại cực trị cục bộ do thiếu cơ chế đột biến linh hoạt, đặc biệt trong các bài toán có nhiều cực trị hoặc ràng buộc phức tạp [12].

Những hạn chế trên cho thấy cần một thuật toán meta-heuristic mới, có khả năng thoát khỏi cực trị cục bộ, duy trì đa dạng quần thể, và cân bằng hiệu quả giữa khai phá và khai thác. Nghiên cứu này đề xuất Dilophosaurus Optimization Algorithm (DOA), lấy cảm hứng từ chiến lược săn mồi linh hoạt của khủng long Dilophosaurus [9]. Không giống các thuật toán trước, DOA tích hợp cơ chế đột biến cao, cho phép thay đổi linh hoạt mục tiêu săn mồi và sử dụng vector ngẫu nhiên để điều chỉnh vị trí, từ đó nâng cao khả năng khám phá không gian tìm kiếm và tránh bẫy cục bộ. So với TROA, DOA cải tiến bằng cách mô phỏng hành vi thay đổi mục tiêu của Dilophosaurus, giúp thuật toán thích nghi tốt hơn với các bài toán tối ưu hóa phức tạp. Bài báo trình bày thiết kế, nguyên lý, và đánh giá thực nghiệm của DOA trên các bài toán tối ưu hóa chuẩn, chứng minh tính mới và tiềm năng ứng dụng của thuật toán trong các lĩnh vực như lập lịch, tối ưu năng lượng, và xử lý dữ liệu lớn.

1. CÁC VẤN ĐỀ CÒN TỒN TẠI TRONG TROA

1.1 Thuật toán TROA

T-Rex Optimization Algorithm (TROA) được xây dựng dựa trên cảm hứng từ hành vi săn mồi của loài khủng long Tyrannosaurus rex – một biểu tượng của sức mạnh và tốc độ trong tự nhiên [8]. Trong quá trình săn mồi, T-Rex sử dụng sức mạnh vượt trội để rượt đuổi con mồi, tập trung vào những mục tiêu yếu nhất nhằm tối đa hóa hiệu quả bắt mồi. Đặc điểm “chọn con mồi yếu” và khả năng linh hoạt chuyển hướng khi phát hiện mục tiêu dễ tiếp cận hơn đã tạo nên một chiến lược săn mồi tối ưu trong điều kiện tự nhiên. TROA mô phỏng quá trình này bằng cách chia quần thể các giải pháp thành các con mồi (Prey) và chỉ định cá thể có giá trị fitness tốt nhất làm “Predator” [8]. Thuật toán sử dụng các bước cập nhật vị trí của predator và prey, hướng đến việc liên tục cải thiện nghiệm tối ưu.

1.2 Các vấn đề còn tồn tại trong TROA

Mặc dù TROA đã được đánh giá cao về khả năng khai thác các vùng tìm kiếm hứa hẹn thông qua chiến lược săn mồi của T-Rex, thuật toán vẫn còn tồn tại một số vấn đề đáng kể cần được giải quyết để nâng cao hiệu quả tối ưu trong các bài toán phức tạp.

Một trong những hạn chế chủ yếu của TROA là khả năng mắc kẹt trong cực trị cục bộ, do trọng tâm quá nhiều vào khai thác các vùng có giá trị hứa hẹn, thuật toán dễ bị rơi vào

bầy của các nghiệm tối ưu giả, đặc biệt trong những không gian tìm kiếm có nhiều cực trị [8].

Bên cạnh đó, việc cân bằng giữa khai thác (exploitation) và khám phá (exploration) vẫn chưa được tối ưu hóa, khiến cho TROA gặp khó khăn trong việc bao phủ toàn diện không gian tìm kiếm, từ đó hạn chế khả năng tiếp cận nghiệm tối ưu toàn cục [12].

Một điểm còn đáng lưu ý nữa là các tham số điều khiển của TROA thường được thiết lập cố định từ đầu mà không có cơ chế tự thích nghi hay điều chỉnh động theo từng giai đoạn của quá trình tối ưu hóa, điều này dẫn đến hiệu suất của thuật toán phụ thuộc rất lớn vào việc lựa chọn ban đầu và không đảm bảo tính ổn định khi áp dụng trên các bài toán có đặc tính khác nhau [8].

Ngoài ra, việc thiếu các cơ chế đột biến linh hoạt cũng góp phần làm giảm khả năng mở rộng và khám phá không gian tìm kiếm mới, đặc biệt là trong các bài toán có số chiều lớn hoặc có nhiều ràng buộc phi tuyến [12].

Do đó, để nâng cao hiệu quả của TROA, các nghiên cứu tiếp theo cần tập trung vào việc tích hợp các cơ chế đột biến mạnh mẽ hơn, phát triển phương pháp tự điều chỉnh tham số điều khiển và cải tiến chiến lược cập nhật vị trí nhằm tạo ra sự cân bằng tối ưu giữa khai thác và khám phá. Việc giải quyết những vấn đề này không chỉ giúp TROA trở nên ổn định và hiệu quả hơn trong các ứng dụng thực tế mà còn mở ra tiềm năng ứng dụng rộng rãi trong các bài toán tối ưu hóa phức tạp và các hệ thống có số chiều cao.

2. PHƯƠNG PHÁP ĐỀ XUẤT

2.1 Khủng long Dilophosaurus

Dilophosaurus là một trong những kẻ săn mồi mạnh mẽ và hiệu quả nhất trong lịch sử tiến hóa. Với kích thước lớn, sức mạnh vượt trội và các giác quan phát triển, Dilophosaurus có khả năng săn lùng và hạ gục những con mồi lớn trong hệ sinh thái cuối kỷ Phấn trắng [9]. Chiến thuật săn mồi của Dilophosaurus chủ yếu dựa vào việc tấn công trực tiếp con mồi và khả năng linh hoạt thông qua việc thay đổi mục tiêu của mình. Với sự kết hợp từ sức mạnh cơ thể vượt trội và khả năng chọn lọc, Dilophosaurus có thể dễ dàng áp đảo và bắt lấy đối thủ. Các cú tấn công Dilophosaurus cũng có khả năng cắn mạnh và giữ chặt con mồi để ngăn chúng chạy trốn, khiến đối thủ không thể thoát khỏi sự tấn công [10].

Các con mồi của Dilophosaurus chủ yếu là những loài khủng long ăn cỏ nhỏ, đôi khi là những loài động vật khác trong cùng hệ sinh thái. Mặc dù các con mồi này có thể sở hữu các vũ khí phòng thủ như sừng hay vỏ bọc bảo vệ, nhưng Dilophosaurus lại có lợi thế về tốc độ và sức mạnh giúp nó dễ dàng hạ gục chúng [11]. Với sự kết hợp giữa sức mạnh, giác quan nhạy bén và chiến thuật thông minh, Dilophosaurus đã trở thành một biểu tượng cho sức mạnh và sự săn mồi.

a) Đặc điểm chính của quá trình săn mồi khủng long Dilophosaurus

Rượt theo con mồi: Dùng tốc độ và sức mạnh để truy đuổi, làm con mồi kiệt sức trước khi hạ gục.

Thay đổi con mồi tốt hơn: Nếu phát hiện con mồi yếu hơn hoặc dễ bắt hơn, sẽ đổi mục tiêu để tăng cơ hội thành công.

2.2 Thuật toán Dilophosaurus Optimization Algorithm (DOA)

a) Ý tưởng và mô hình toán học của DOA

Hành vi săn mồi của khủng long: Dilophosaurus ưu tiên tấn công con mồi yếu nhất, gần nhất để tối đa hóa khả năng hạ gục. Trong quá trình truy đuổi, nó liên tục đánh giá và chọn lọc con mồi tiềm năng hơn, thích nghi nhanh chóng để tối ưu hóa cơ hội săn mồi thành công.

Công thức xác định vị trí mới của Khủng Long:

$$X_{KL-new} = X_{KL} + (X_{KL} - X_{target}) * rand() \quad (1)$$

Trong đó: $rand()$: là số thực ngẫu nhiên từ (-1) đến 1.

Hành vi thay đổi vị trí tự nhiên của con mồi: Khi bị Dilophosaurus nhắm đến, con mồi sẽ tìm cách chạy trốn, tạo ra sự xáo trộn trong khu vực. Trong khi đó, các quần thể khác tiếp tục di chuyển xung quanh vị trí cũ để kiếm ăn, vô tình làm thay đổi cục diện săn mồi và mở ra cơ hội tiếp cận con mồi mới cho Dilophosaurus.

Công thức cập nhật vị trí của mục tiêu (Target):

$$X_{target-new} = X_{target} + (X_{target} - X_{KL}) * rand() \quad (2)$$

Công thức cập nhật vị trí của các cá thể không phải mục tiêu (Non-Target):

$$X_{non-target-new} = X_{non-target} + a * rand() \quad (3)$$

Công thức tạo cá thể (Prey) mới thay thế cá thể (Prey) cũ sau khi chết:

$$X_{new-prey} = X_{target} * rand() \quad (4)$$

Trong đó:

$rand()$: là số thực ngẫu nhiên từ (-1) đến 1.

a : là (giới hạn trên – giới hạn dưới)/ giới hạn trên.

b) Quy trình cơ bản của DOA:

Bước 1. Khởi tạo quần thể: Dilophosaurus được đặt là D, các con mồi là các P. IterMax được xem như số vòng săn mồi tối đa.

Bước 2. Cập nhật quần thể: Vị trí quần thể được cập nhật sau quá trình săn mồi theo các công thức.

Bước 3. Lựa chọn: Thay thế con mới tốt hơn nếu có.

Bước 4. Lặp lại: Nếu chưa đạt số vòng lặp tối đa, quay lại bước 2.

Bước 5. Kết thúc: Đưa ra giải pháp tối ưu và giá trị tối ưu.

Tình chỉnh DOA: Có thể điều chỉnh các tham số như số lượng con mới, tỷ lệ ăn thịt, tỷ lệ tử vong để tối ưu hóa quá trình.

3. SO SÁNH CẢI TIẾN TROA – DOA

3.1 Các bước chính của TROA

Bước 1: Khởi tạo quần thể: Tạo ngẫu nhiên các cá thể trong không gian tìm kiếm và chọn cá thể có fitness tốt nhất làm “predator”.

Bước 2: Xác định mục tiêu: Chọn con mồi (cá thể) có giá trị fitness tốt nhất làm mục tiêu cho predator.

Bước 3: Cập nhật vị trí predator: Di chuyển predator về phía mục tiêu nhằm giảm khoảng cách giữa chúng.

Bước 4: Cập nhật vị trí prey: Các prey di chuyển ngẫu nhiên để duy trì đa dạng. Nếu phát hiện cá thể có fitness tốt hơn, cập nhật vị trí theo hướng đó.

Bước 5: Thay thế và lặp lại: Các cá thể yếu được thay thế bởi các cá thể mới, và quá trình được lặp lại cho đến khi đạt tiêu chí dừng.

3.2 Chi tiết các bước Thuật toán DOA

Bước 1: Khởi tạo quần thể, Dilophosaurus được gọi là D, các con mồi là các P. IterMax được xem như số vòng săn mồi tối đa mà Dilophosaurus có thể thực hiện trước khi kiệt sức.

Bước 2: Bảng giá trị thích nghi của các P và Dilophosaurus được xác định theo các công thức tính fitness, Dilophosaurus sẽ lựa chọn P có độ thích nghi thấp nhất để chọn làm Target.

Bước 3: Cập nhật vị trí Dilophosaurus sau khi đuổi theo con mồi theo công thức (1).

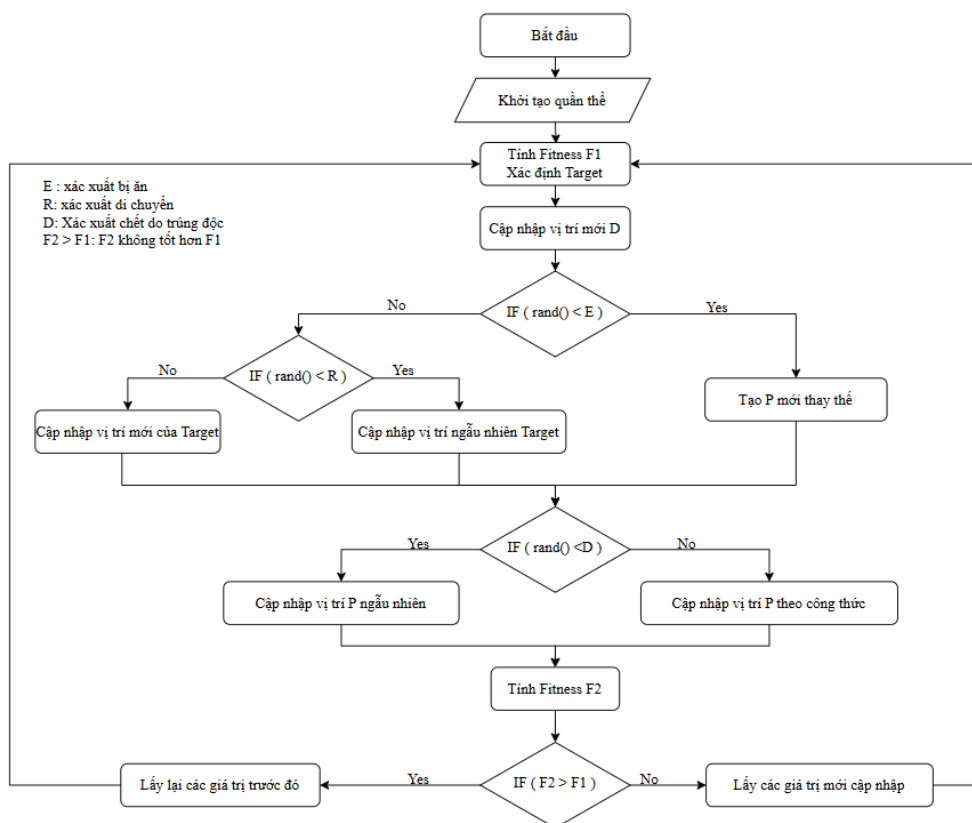
Bước 4: Nếu tỉ lệ Target bị ăn $\text{rand}[0,1] < E$, tạo P mới thay thế Target bị ăn theo công thức (4); nếu không, cập nhật vị trí mới của Target đang chạy trốn dựa trên công thức (2).

Bước 5: Nếu Target không bị ăn (tức là $\text{rand}[0,1] \geq E$ ở Bước 4), thì kiểm tra: Nếu $\text{rand}[0,1] < R$, cập nhật vị trí mới cho Target một cách ngẫu nhiên (Target trốn thoát thành công); nếu không, cập nhật vị trí Target theo công thức (2).

Bước 6: Nếu tỉ lệ trúng chết vì lý do khách quan $\text{rand}[0,1] < D$, tạo P mới thay thế theo công thức (4); nếu không cập nhật vị trí di chuyển P theo (3).

Bước 7: Tính lại bảng giá trị thích nghi mới của các P, Dilophosaurus theo các công thức tính fitness.

Bước 8: So sánh kết quả bảng giá trị thích nghi cũ và mới, nếu tốt hơn ($F2 < F1$), quay lại bước 3 để tiếp tục vòng lặp nếu chưa đạt tới vòng lặp tối đa; nếu không làm lại bước 2.



Hình 1. Lưu đồ thuật toán DOA

3.3 So sánh TROA và DOA

Cả TROA và DOA đều được xây dựng dựa trên nguyên tắc mô phỏng hành vi săn mồi của khủng long ăn thịt, trong đó predator (kẻ săn mồi) đóng vai trò là cá thể ưu tú nhất trong quần thể, còn prey (con mồi) tượng trưng cho các cá thể còn lại. Điểm chung của hai thuật toán là chúng đều hướng đến việc khai thác con mồi yếu nhất để tối ưu hóa quá trình tìm kiếm lời giải. Tuy nhiên, DOA có một sự khác biệt đáng kể khi tích hợp cơ chế đột biến mạnh, cho phép thuật toán không chỉ tập trung vào việc khai thác mà còn có khả năng thay đổi linh hoạt chiến lược tấn công. Điều này giúp DOA có thể thích nghi tốt hơn với các bài toán tối ưu có không gian tìm kiếm phức tạp.

Về cơ chế cập nhật vị trí, TROA thực hiện theo nguyên tắc quen thuộc là để predator di chuyển trực tiếp đến con mồi yếu nhất nhằm khai thác vùng tìm kiếm hiệu quả. Trong khi đó, DOA lại sử dụng các vector ngẫu nhiên để điều chỉnh vị trí của predator, tạo ra sự đa dạng hơn trong quá trình di chuyển và hạn chế nguy cơ mắc kẹt trong bẫy cục bộ. Đặc biệt, trong quá trình con mồi phản ứng trước sự săn đuổi, TROA giả lập hành vi con mồi chạy xa

khỏi predator để bảo toàn tính mạng. Trong khi đó, DOA đưa ra hai cơ chế khác nhau: một số con mồi có thể di chuyển có định hướng để tận dụng vị trí có lợi, trong khi những con khác thực hiện bước nhảy ngẫu nhiên để mở rộng phạm vi tìm kiếm. Điều này giúp DOA duy trì sự cân bằng giữa khai thác các vùng tìm kiếm tiềm năng và mở rộng không gian tìm kiếm nhằm tìm ra lời giải tốt hơn.

Ngoài ra, khi xử lý các cá thể có chất lượng kém, cả hai thuật toán đều áp dụng cơ chế thay thế để cải thiện quần thể qua từng thế hệ. TROA thực hiện việc này theo hướng đơn giản, đó là loại bỏ các cá thể yếu và thay thế bằng những cá thể có chất lượng cao hơn. Trong khi đó, để tạo ra cá thể mới, DOA có thêm bước so sánh fitness trước khi thay thế nhằm đảm bảo sự cải thiện ổn định và tránh việc tạo ra cá thể mới không thực sự có lợi.

Điểm nổi bật nhất của DOA so với TROA chính là mức độ đột biến cao. Nếu như TROA chủ yếu tập trung vào khai thác vùng tìm kiếm tốt đã xác định, thì DOA lại tăng cường khả năng khám phá nhờ vào cơ chế thay đổi chiến lược linh hoạt. Điều này giúp DOA không chỉ nhanh chóng thoát khỏi bẫy cục bộ mà còn có khả năng tiếp cận những vùng tối ưu mới một cách hiệu quả hơn. Chính sự kết hợp giữa khai thác và khám phá trong DOA khiến nó trở thành một thuật toán tiềm năng với khả năng tìm kiếm lời giải tối ưu nhanh chóng và chính xác hơn so với TROA trong nhiều bài toán thực tế.

ĐẶC ĐIỂM	TROA	DOA
Chiến lược săn mồi	Chọn con mồi yếu nhất, ưu tiên gần nhất.	Nhắm vào con mồi yếu nhất nhưng có cơ chế đột biến cao để thay đổi chiến lược tấn công.
Cập nhật vị trí predator	Predator di chuyển dựa trên con mồi yếu nhất.	Predator có thể thay đổi linh hoạt vị trí bằng cách sử dụng vector ngẫu nhiên.
Cơ chế trốn chạy của con mồi	Con mồi chạy xa khỏi Predator.	Con mồi có thể chạy trốn hoặc di chuyển có định hướng.
Cơ chế thay thế cá thể xấu	Thay thế bằng cá thể mới có chất lượng cao hơn.	Thay thế cá thể mới, nhưng có điều kiện so sánh fitness để có chất lượng cao hơn.
Tính đột biến	Ít có sự đột biến, tập trung vào khai thác vùng tìm.	Mức độ đột biến cao, giúp thuật toán thoát khỏi bẫy cục bộ nhanh hơn.

4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Để đánh giá hiệu quả của thuật toán DOA, trong phần này chúng tôi thực hiện các thử nghiệm sơ bộ trên bản Excel. Các bản thử nghiệm này được thực hiện nhằm kiểm tra khả năng của DOA trong việc tìm kiếm cực trị toàn cục trên các không gian tìm kiếm phức tạp và đa dạng. Phiên bản thử nghiệm bằng Excel cho phép chúng tôi quan sát quá trình hội tụ và phân tích hành vi của thuật toán một cách trực quan trước khi tiến hành triển khai trên các nền tảng tính toán chuyên sâu.

Trong thử nghiệm này, mỗi Prey (mồi) là một cá thể trong quần thể mà thuật toán sử dụng để tìm kiếm nghiệm tối ưu. Mỗi Prey không chỉ là một điểm trong không gian tìm kiếm mà là một vector chứa nhiều biến, trong trường hợp này là 10 biến.

Các biến $x_1, x_2, x_3, \dots, x_{10}$ được khởi tạo ngẫu nhiên trong một phạm vi xác định, có thể là $[x_{min}, x_{max}]$, phụ thuộc vào bài toán cụ thể. Trong trường hợp này, giả sử không gian tìm kiếm của mỗi biến nằm trong khoảng $[-10, 10]$, các giá trị khởi tạo sẽ được xác định theo công thức sau:

$$x_i^{(0)} = random(x_{min}, x_{max}), \quad \forall i = 1, 2, 3 \dots 10$$

Trong đó:

$x_i^{(0)}$ là giá trị khởi tạo của biến x_i tại vòng lặp đầu tiên (vòng lặp $t = 0$).

Sau khi khởi tạo quần thể ban đầu, thuật toán DOA bắt đầu quá trình tối ưu hóa qua nhiều vòng lặp. Mỗi vòng lặp sẽ thực hiện cập nhật các giá trị của các biến $x_1, x_2, x_3, \dots, x_{10}$ dựa trên các chiến lược tìm kiếm và cập nhật vị trí của các cá thể trong không gian tìm kiếm.

Fitness là một chỉ số đo lường chất lượng của một cá thể trong quần thể của thuật toán tối ưu hóa. Chỉ số này thể hiện độ "khó" hoặc "dễ" của một nghiệm khi so với giải pháp tối ưu trong không gian tìm kiếm.

4.1 Kết quả thực nghiệm

Dưới đây là kết quả thực nghiệm trên bản Excel với số chiều là 10:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Fitness = $ F_{\text{predator}} - F_{\text{prey}} $
Predator	0,3433	8,7756	3,1861	8,5565	5,6330	-7,6214	-7,6986	5,3935	1,9593	0,7144	
Prey ₁	6,3699	-2,1790	-4,7518	2,8217	-2,8313	7,9062	-8,0408	-4,5031	-2,6478	-2,9548	95,955
Prey ₂	-8,7534	-3,0695	-9,1455	1,2272	-3,4662	-8,5454	-8,3108	3,7996	3,9822	-6,6800	57,196
Prey ₃	0,9926	-9,6140	3,0896	-2,3362	6,3587	3,8398	6,9760	-9,0472	-1,7415	5,7117	13,251
Prey ₄	-4,1277	-4,1688	9,6407	-2,6010	-7,4930	-8,2045	-5,1109	8,6564	-8,0890	5,3400	109,567
Prey ₅	-5,2258	1,6238	7,7904	9,3593	0,8643	-0,2748	-0,3457	1,4669	-2,4841	2,7115	148,169
Prey ₆	9,7046	-3,3329	-7,9853	-1,2612	-7,4743	6,0879	9,4477	-0,9481	-9,4796	7,6263	158,734
Prey ₇	8,0474	5,5662	2,3920	1,5168	-0,7483	4,0262	4,7364	8,3531	6,8467	-2,0524	79,185
Prey ₈	-3,2397	-8,6304	3,8287	0,0160	4,9999	6,2962	-8,4902	4,6022	8,8659	8,8935	72,224
Prey ₉	7,0970	1,9141	-7,3273	4,6949	9,4072	9,1809	6,9216	2,1130	-1,7123	-1,8309	18,186
Prey ₁₀	-2,9687	5,2015	0,9190	-4,5788	8,4806	2,1419	-4,8795	0,0666	-2,2677	1,1384	178,580

Hình 2. Giá trị fitness khởi tạo quần thể ban đầu

Fitness tốt nhất ở thời điểm khởi tạo: $F_1 = 13,2510$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Fitness = $ F_{\text{predator}} - F_{\text{prey}} $
Predator	0,6271	14,6537	3,2190	10,0354	6,0318	-2,2640	6,4595	-1,2850	0,5129	0,2216	
Prey ₁	6,0481	-3,3301	-2,8225	1,7339	-2,3568	9,5165	-8,7633	-5,4199	-3,5993	-2,1709	132,793
Prey ₂	-8,4530	-3,1017	-8,4645	2,9756	-1,9294	-7,6492	-7,4427	3,8715	5,6769	-7,2168	32,891
Prey ₃	0,9712	-14,1039	3,1724	-2,3648	6,4507	9,1697	7,0535	-14,4753	-1,0266	2,2240	195,106
Prey ₄	-3,8403	-2,7707	8,9692	-3,1454	-6,7935	-8,5996	-4,5379	7,9856	-6,6805	7,3389	4,327
Prey ₅	-3,8520	0,2177	6,4598	9,6993	2,2500	1,5416	-1,7175	2,2733	-1,6914	3,4210	230,580
Prey ₆	7,8216	-1,5193	-7,5629	-0,3776	-9,2711	7,8904	8,9686	-0,2571	-8,1138	7,1921	55,705
Prey ₇	9,4476	4,0935	0,9296	-0,0950	1,0883	3,3189	5,0454	9,3131	7,9290	-3,4413	105,404
Prey ₈	-2,2170	-9,6800	5,4308	-1,0023	3,0960	4,7613	-7,4115	3,1358	8,3777	7,2395	62,661
Prey ₉	5,6279	1,2486	-7,1653	4,3948	8,7026	10,9261	7,6746	3,6988	-3,6289	-1,8687	23,147
Prey ₁₀	-4,6681	4,5456	-0,2822	-5,2183	8,2014	3,3681	-3,5335	-0,6787	-0,9977	2,8062	241,205

Hình 2. Giá trị fitness quần thể sau 20 vòng lặp

Fitness tốt nhất ở thời điểm vòng lặp 20: $F_2 = 4,327$

4.2 Phân Tích Kết Quả

Kết quả thực nghiệm trên bản Excel với số chiều là 10 cho thấy khả năng cải thiện của thuật toán DOA trong quá trình tối ưu hóa. Ban đầu, quần thể khởi tạo đạt giá trị fitness tốt nhất là $F_1 = 13,2510$ (Hình 1). Sau khi chạy 20 vòng lặp, giá trị fitness tốt nhất đã giảm xuống còn $F_2 = 4,327$ (Hình 2). Sự giảm đáng kể này cho thấy cơ chế cập nhật vị trí và thay đổi mục tiêu của DOA đang hoạt động hiệu quả, giúp cải thiện nghiệm tối ưu từ một quần thể ban đầu còn khá xa so với giải pháp tối ưu. Mức giảm từ 13,2510 xuống 4,327 phản ánh khả năng hội tụ của thuật toán, tuy nhiên, khoảng cách giữa giá trị ban đầu và giá trị sau 20 vòng lặp cho thấy quá trình tối ưu hóa vẫn còn chỗ để cải thiện thêm.

Kết quả này là bằng chứng sơ bộ cho thấy DOA có tiềm năng vượt trội trong việc tìm kiếm nghiệm tối ưu trong các không gian tìm kiếm phức tạp, nhưng để khẳng định tính ổn định và hiệu suất tối ưu toàn cục, cần triển khai thêm các thử nghiệm trên các hàm thử nghiệm tiêu chuẩn và với số vòng lặp lớn hơn.

5. KẾT LUẬN VÀ PHƯƠNG HƯỚNG PHÁT TRIỂN

Nghiên cứu này đã giới thiệu và đánh giá sơ bộ thuật toán DOA – một phương pháp tối ưu hóa mới dựa trên mô phỏng hành vi săn mồi của khủng long, kết hợp cơ chế đột biến cao và khả năng thay đổi mục tiêu linh hoạt. Mặc dù hiện tại DOA chỉ được thử nghiệm qua bản chạy bằng Excel và chưa được áp dụng trên các hàm thử nghiệm tiêu chuẩn, nhưng các kết

qua ban đầu cho thấy thuật toán có xu hướng hội tụ nhanh và đạt được giá trị tối ưu ổn định trên một số bài toán tối ưu hóa đơn giản. Điều này cho thấy tiềm năng khai thác hiệu quả không gian tìm kiếm, đặc biệt là khả năng thoát khỏi bẫy cực trị cục bộ nhờ vào các bước cập nhật vị trí đa dạng và linh hoạt.

Ngoài ra, DOA còn cho thấy ưu thế trong việc duy trì sự đa dạng của quần thể, ngay cả khi kích thước quần thể tăng lên, giúp mở rộng phạm vi tìm kiếm và tiếp cận các nghiệm tối ưu một cách hiệu quả. Các cơ chế như cập nhật vị trí dựa trên vector ngẫu nhiên của predator và các phương pháp điều chỉnh chuyển động của Prey đã tạo nên một hệ thống tối ưu hóa linh hoạt, góp phần nâng cao khả năng khám phá không gian tìm kiếm phức tạp. Những kết quả sơ khai từ phiên bản Excel đã cung cấp cơ sở hứa hẹn cho việc phát triển thêm các chiến lược điều chỉnh tham số và cải tiến thuật toán trong tương lai.

Hướng phát triển tiếp theo của DOA sẽ tập trung vào việc triển khai thuật toán trên các hàm thử nghiệm tiêu chuẩn và các bộ benchmark phức tạp để đánh giá toàn diện khả năng hội tụ và hiệu suất của nó. Cụ thể, việc tối ưu hóa các tham số điều khiển (như E, R, D) thông qua các chiến lược thích nghi tự động, cải tiến cơ chế đột biến và cập nhật vị trí của Predator và Prey sẽ giúp DOA hoạt động ổn định hơn trong môi trường tính toán chuyên sâu. Bên cạnh đó, kết hợp DOA với các thuật toán tối ưu hóa khác (hybridization) cũng là một hướng nghiên cứu đáng chú ý nhằm tận dụng ưu điểm của từng phương pháp.

Tóm lại, mặc dù DOA hiện vẫn đang ở giai đoạn sơ khai với các thử nghiệm ban đầu được thực hiện bằng Excel, nhưng những kết quả sơ bộ đã khẳng định tính khả thi và tiềm năng ứng dụng của thuật toán. Các nghiên cứu trong tương lai sẽ tiếp tục tinh chỉnh và mở rộng DOA, từ đó nâng cao hiệu quả tối ưu hóa cho các bài toán thực tế phức tạp, đồng thời khẳng định vai trò của DOA như một công cụ mạnh mẽ trong lĩnh vực tối ưu hóa.

TÀI LIỆU THAM KHẢO

- [1] Talbi, E. G., *Metaheuristics: From Design to Implementation*, Wiley, 1–500, (2009).
- [2] Yang, X. S., *Nature-Inspired Metaheuristic Algorithms*, Luniver Press, 1–148, (2010).
- [3] Kennedy, J. & Eberhart, R., Particle Swarm Optimization, In *Proceedings of the IEEE International Conference on Neural Networks*, 1942–1948 (1995).
- [4] Holland, J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1–183, Michigan (1975).
- [5] Mirjalili, S., Mirjalili, S.M., & Lewis, A., Grey Wolf Optimizer, *Advances in Engineering Software*, (2014), 69, 46–67.
- [6] Mirjalili, S. & Lewis, A., The Whale Optimization Algorithm, *Advances in Engineering Software*, (2016), 95, 51–67.
- [7] Dehghani, M., Montazeri, Z., Trojovská, E., & Trojovský, P., Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems, *Knowledge-Based Systems*, 277, (2023), 110737.

- [8] Sahu, V. S. D. M., Samal, P. & Panigrahi, C. K., Tyrannosaurus Optimization Algorithm: A New Nature-Inspired Meta-Heuristic Algorithm for Solving Optimal Control Problems, e-Prime - Advances in Electrical Engineering, Electronics and Energy, (2023), 5, 100243.
- [9] Welles, S. P., Dilophosaurus wetherilli, gen. et sp. nov. Memoirs of the University of California Museum of Paleontology, 21, 1– 49 (1970).
- [10] Madsen, S. & Welles, S. P., The braincase of Dilophosaurus wetherilli and its implications. American Museum Novitates, 2720, 1–18 (1980).
- [11] Chure, D. J., A Reevaluation of Dilophosaurus and Its Implications for Early Theropod Evolution, Doctoral dissertation, University of Calgary (2001). (Luận án Tiến sĩ, University of Calgary).
- [12] Zhang, J., Zhao, X., Zhang, Y., Yang, H. & Xue, B., A Dynamic Hierarchical Improved Tyrannosaurus Optimization Algorithm with Hybrid Topology Structure, Mathematics, (2024), 12(10), 1459.

ABSTRACT

DILOPHOSAURUS OPTIMIZATION ALGORITHM (DOA): ENHANCING THE PERFORMANCE OF THE T-REX ALGORITHM (TROA)

Nguyen The Huu^{1,*}, Leu Cong Trung Hieu¹, Dang Nguyen Hoai¹, Dinh Nguyen Trong Nghia¹

¹*Ho Chi Minh City University of Industry and Trade*

*Email: huunthe@huit.edu.vn

This paper introduces a novel meta-heuristic algorithm named the Dilophosaurus Optimization Algorithm (DOA), inspired by the characteristics and hunting strategies of the Dilophosaurus dinosaur. DOA simulates the combination of speed, prey-tracking ability, and flexible attack behavior of this dinosaur species to efficiently explore and exploit the solution space. This approach enhances the ability to escape local optima and improves search performance. DOA is evaluated on a set of standard optimization problems and compared with other popular meta-heuristic algorithms. Experimental results demonstrate that DOA achieves superior optimization performance, even outperforming existing methods in certain cases, proving its effectiveness and potential applications in complex optimization problems.

Keywords: Meta-heuristic, Optimization, TROA, DOA

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thế Hữu

Số điện thoại liên hệ: 0932163743

MÔ HÌNH ĐỀ XUẤT KẾT HỢP COATI VÀ MVO TRONG HỆ THỐNG HỖ TRỢ QUYẾT ĐỊNH ĐẦU TƯ

**Huỳnh Thị Châu Lan^{1,*}, Trần Huỳnh Phụng Quyên¹, Dương Trọng Đức¹,
Hứa Vĩnh Khang¹, Đinh Nguyễn Trọng Nghĩa¹,**

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

**Email: lanhtc@huit.edu.vn*

Ngày nhận bài: 11/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Quyết định đầu tư là một quá trình phức tạp chịu ảnh hưởng của nhiều yếu tố, đòi hỏi các kỹ thuật tối ưu hóa hiệu quả để nâng cao khả năng lựa chọn danh mục đầu tư. Bài báo này đề xuất một hướng tiếp cận mới bằng cách tích hợp Thuật toán tối ưu Coati (COA) với Thuật toán tối ưu hóa đa vũ trụ (MVO) nhằm xây dựng một hệ thống hỗ trợ quyết định đầu tư thông minh. Thuật toán Coati, lấy cảm hứng từ hành vi kiếm ăn và xã hội của loài gấu mèo Coati, được kết hợp với cơ chế thăm dò-khai thác của MVO để cải thiện chiến lược phân bổ danh mục đầu tư. Nghiên cứu này giới thiệu khung lý thuyết, thiết kế thuật toán và sơ đồ quy trình tích hợp Coati với MVO. Mặc dù bài báo chưa đi sâu vào thực nghiệm, mô hình đề xuất mở ra một hướng nghiên cứu tiềm năng cho các ứng dụng thực tế trong lĩnh vực ra quyết định tài chính.

Keywords: COA, MVO, tối ưu hóa, gợi ý danh mục đầu tư.

1. MỞ ĐẦU

Bài toán gợi ý danh mục đầu tư là một vấn đề quan trọng trong lĩnh vực tài chính, với mục tiêu tối ưu hóa việc phân bổ tài sản nhằm đạt được lợi nhuận cao nhất trong khi vẫn giảm thiểu rủi ro [1]. Theo Markowitz (1952), sự phát triển quan trọng trong việc lựa chọn danh mục đầu tư là việc giảm thiểu sự tương quan giữa các tài sản, từ đó xác định rủi ro của danh mục đầu tư dưới mức lợi nhuận kỳ vọng của nhà đầu tư [2]. Tuy nhiên, trong bối cảnh số lượng tài sản và các yếu tố tác động đến thị trường ngày càng gia tăng, bài toán tối ưu hóa danh mục đầu tư trở nên phức tạp và có sự xuất hiện của các ràng buộc phi tuyến tính với các hàm đa mục tiêu, điều này khiến cho việc giải quyết bằng các phương pháp tiếp cận truyền thống trở nên không hiệu quả [3][4]. Vì vậy, các nhà nghiên cứu đã luôn tìm cách làm cho các mô hình toán học về lựa chọn danh mục đầu tư gần gũi hơn với thực tế, nhằm hỗ trợ các nhà đầu tư đạt được mục tiêu của mình.

Trong bối cảnh đó, việc áp dụng các thuật toán tối ưu hóa hiệu quả trở nên vô cùng quan trọng. Thuật toán Coati Optimization Algorithm (COA), được giới thiệu bởi Dehghani và cộng sự vào năm 2022, là một thuật toán lấy cảm hứng từ hành vi săn mồi và trốn tránh kẻ thù của loài coati trong tự nhiên [5]. COA đã được chứng minh có khả năng tìm kiếm toàn

cục và cục bộ hiệu quả, vượt trội so với một số thuật toán hiện có trong việc giải quyết các vấn đề tối ưu hóa thực tế.

Bên cạnh đó, Multi-Verse Optimizer (MVO), được phát triển bởi Mirjalili và cộng sự vào năm 2016, là một thuật toán tối ưu hóa lấy cảm hứng từ lý thuyết đa vũ trụ, mô phỏng các hiện tượng như hố trắng, hố đen và giun vũ trụ để tìm kiếm giải pháp tối ưu [6]. MVO đã được mở rộng thành phiên bản đa mục tiêu (MOMVO) để giải quyết các bài toán tối ưu hóa đa mục tiêu, cho thấy hiệu quả cao trên các hàm benchmark [7].

Việc kết hợp COA và MVO hứa hẹn mang lại một phương pháp tiếp cận hiệu quả cho bài toán gợi ý danh mục đầu tư, tận dụng ưu điểm của cả hai thuật toán trong việc cân bằng giữa khai thác và khám phá không gian tìm kiếm. Nghiên cứu này đề xuất một phương pháp tiếp cận dựa trên sự kết hợp giữa COA và MVO để cải thiện hiệu suất tối ưu hóa danh mục đầu tư.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Giới thiệu

Bài toán gợi ý danh mục đầu tư nhằm mục tiêu đề xuất danh mục tài sản phù hợp, giúp nhà đầu tư đạt được sự cân bằng giữa rủi ro và lợi nhuận. Việc lựa chọn danh mục không chỉ dựa trên lợi nhuận kỳ vọng mà còn cần xem xét nhiều yếu tố khác như mức độ biến động, sự đa dạng hóa và các tiêu chí tài chính cụ thể.

Để giải quyết bài toán này, phương pháp kết hợp Thuật toán Coati (COA) và Thuật toán Đa vũ trụ (MVO) được áp dụng nhằm tìm ra những danh mục đầu tư tối ưu. Cách tiếp cận này giúp khám phá nhiều phương án khác nhau, từ đó xác định những lựa chọn có hiệu suất tốt nhất, phù hợp với nhu cầu và mục tiêu của nhà đầu tư.

2.2 Đầu vào của bài toán

Bài toán gợi ý danh mục đầu tư bao gồm các tập dữ liệu đầu vào chính sau:

Danh sách các danh mục đầu tư, mỗi danh mục bao gồm các thông tin:

- **criteria_values** là trọng số các tiêu chí của từng danh mục, bao gồm: lợi nhuận, rủi ro, thanh khoản, thời gian đầu tư, chi phí, tính ổn định của dòng tiền và đa dạng hóa.
- **criteria_risk_investment** là lợi nhuận và rủi ro ước tính của từng danh mục.

Trọng số ưu tiên của khách hàng: Mỗi khách hàng sẽ có mức độ quan tâm khác nhau đến các tiêu chí trên. Trọng số này được sử dụng để đánh giá và xếp hạng danh mục đầu tư tối ưu.

Bảng 1. Trọng số tiêu chí ưu tiên của khách hàng

Criteria	Criteria Weight	Rank
Diversification	0.272727	1
Liquidity	0.212121	2
Investment Horizon	0.181818	3
Cost	0.121212	4
Expected Return	0.098989	5

Criteria	Criteria Weight	Rank
Risk	0.098989	5
Stability of Cash Flow	0.030303	7

Bảng 1 minh họa giá trị trọng số tiêu chí và xếp hạng ưu tiên của tiêu chí theo mức độ quan tâm của khách hàng.

2.3 Mô hình hóa bài toán

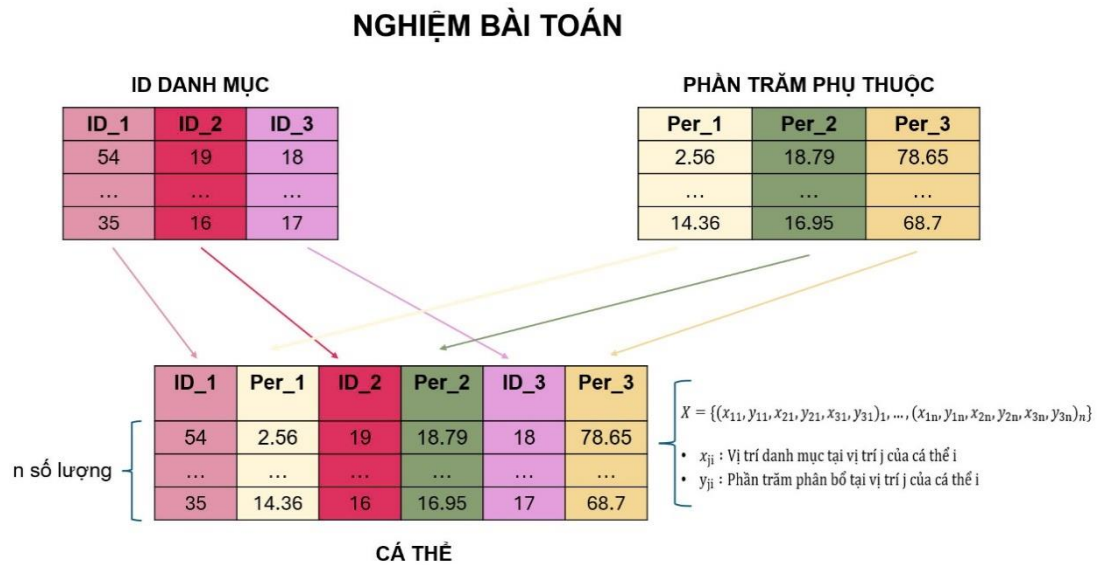
Bài toán được mô hình hóa dưới dạng một cấu trúc dữ liệu ánh xạ giữa danh mục đầu tư và trọng số các tiêu chí ưu tiên của khách hàng. Mỗi nghiệm (cá thể) trong không gian tìm kiếm sẽ biểu diễn một phương án phân bổ danh mục đầu tư tối ưu, được xác định bằng tập hợp:

$$S = \{(id_investment, per_allocation)\} \tag{1}$$

Trong đó:

- *id_investment*: chỉ số danh mục đầu tư
- *per_allocation*: phần trăm phân bổ cho danh mục đầu tư

Không gian nghiệm được tổ chức sao cho mỗi cá thể đại diện cho một danh sách danh mục đầu tư tối ưu, trong đó tất cả danh mục được chọn và phân bổ hợp lý, đồng thời đảm bảo tuân thủ các ràng buộc cứng và tối ưu theo trọng số ưu tiên của khách hàng.



Hình 1. Biểu diễn nghiệm của bài toán gợi ý danh mục đầu tư

Như minh họa trong Hình 1, mỗi cá thể trong không gian tìm kiếm sẽ ánh xạ một tập hợp các danh mục và phần trăm phân bổ tương ứng. Cấu trúc này cho phép thuật toán tối ưu hóa tìm kiếm và điều chỉnh phương án gợi ý một cách linh hoạt, đảm bảo tính khả thi và hiệu quả trong quá trình triển khai.

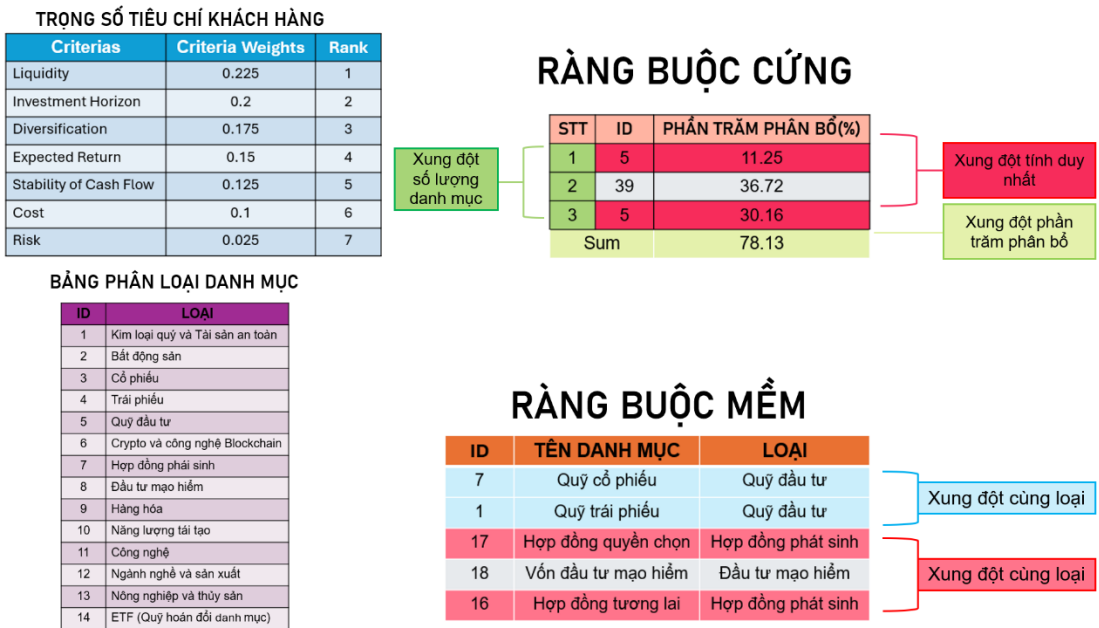
2.4 Ràng buộc bài toán

Bài toán gợi ý danh mục đầu tư đòi hỏi việc phân bổ các danh mục và phần trăm phân bổ một cách hợp lý, đảm bảo không vi phạm các ràng buộc hệ thống. Các ràng buộc trong bài toán này có thể được chia thành hai nhóm chính: ràng buộc cứng và ràng buộc mềm.

Ràng buộc cứng (Hard Constraints - HC): Đây là các ràng buộc bắt buộc phải tuân thủ. Nếu vi phạm, nghiệm bị loại bỏ vì không khả thi. Các ràng buộc này bao gồm:

- Tổng phân bổ: Tổng phần trăm phân bổ cho các danh mục đầu tư được chọn phải bằng 100%.
- Danh mục duy nhất: Không thể có hai danh mục trùng nhau trong danh sách gợi ý.
- Số lượng danh mục theo tiêu chí đa dạng hóa: Nếu tiêu chí "đa dạng hóa" nằm trong top 3 tiêu chí khách hàng ưu tiên thì hiển thị 5 danh mục, ngược lại hiển thị 3 danh mục.

Ràng buộc mềm (Soft Constraints - SC): Đây là các ràng buộc không bắt buộc nhưng nên được tối ưu hóa để cải thiện chất lượng lịch trình. Các ràng buộc này có thể bị vi phạm nếu không còn phương án khả thi khác, nhưng sẽ bị tính điểm phạt. Giới hạn danh mục cũng loại là một dạng ràng buộc mềm. Nó yêu cầu không quá 2 danh mục cùng loại khi tiêu chí "đa dạng hóa" nằm trong top 3 ưu tiên của khách hàng.



Hình 2. Phân loại và minh họa ràng buộc cứng và ràng buộc mềm trong bài toán gợi ý danh mục đầu tư

Ràng buộc cứng sẽ giới hạn số lượng và tỷ lệ phân bổ phù hợp của danh mục, đảm bảo tuân thủ các nguyên tắc. Trong khi đó, ràng buộc mềm xử lý xung đột giữa các danh mục cùng loại, giúp duy trì sự đa dạng hóa và tối ưu hóa chiến lược phân bổ tài sản.

Hình 3 minh họa quá trình tối ưu hóa danh mục đầu tư được thực hiện bằng cách lựa chọn cá thể tốt nhất từ không gian tìm kiếm sau nhiều vòng lặp. Phần cá thể tốt nhất thể hiện sự thay đổi của cá thể qua các lần lặp, với giá trị fitness phản ánh mức độ tối ưu của từng danh mục. Phần danh sách danh mục tối ưu nhất thì liệt kê danh mục đầu tư tương ứng với tỷ trọng phân bổ tối ưu, cung cấp cơ sở cho chiến lược đầu tư hiệu quả và cân bằng rủi ro.



Hình 3. Phân loại và minh họa danh sách danh mục đầu tư tối ưu của bài toán gợi ý danh mục đầu tư

2.5 Xây dựng hàm mục tiêu

Hàm mục tiêu trong bài toán này được xây dựng nhằm tối ưu hóa danh mục được gợi ý trong khi đảm bảo tuân thủ các ràng buộc. Cấu trúc hàm mục tiêu có thể được biểu diễn như sau:

$$F_{danhmuc} = \left| \sum_{j \in P} (\omega_j \times x_{id,j} \times \alpha_j) - \sum_{k \in N} (\omega_k \times x_{id,k} \times \alpha_k) + \sum_{m \in M} (\omega_m \times x_{id,m} \times \alpha_m) \right| \quad (2)$$

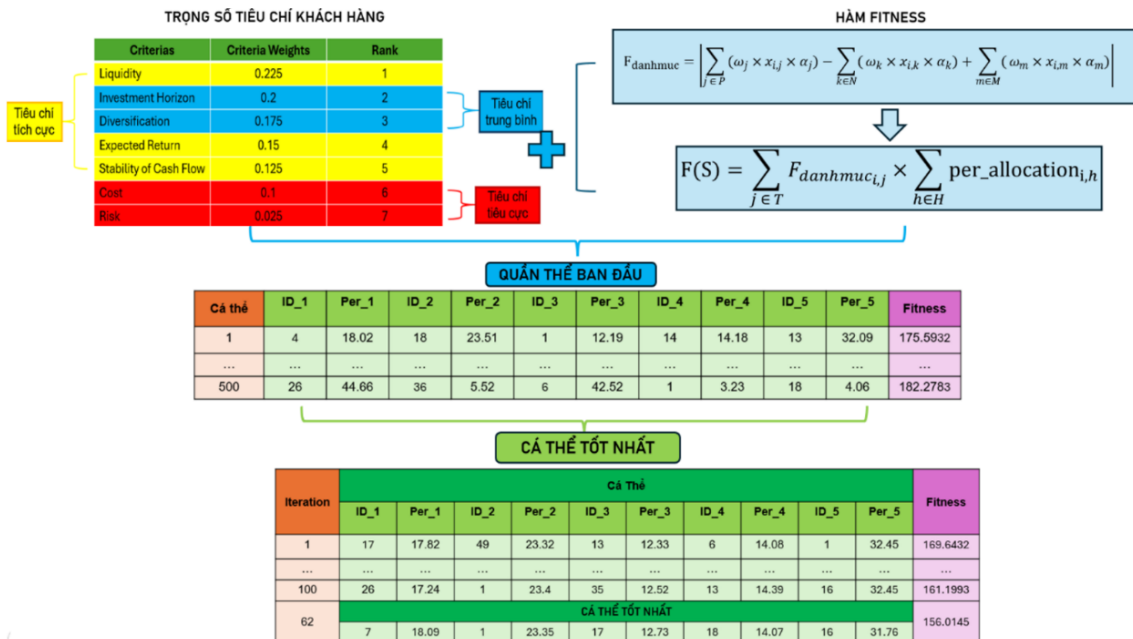
$$F(S) = \sum_{t \in T} F_{danhmuc_{i,t}} \times \sum_{h \in H} per_allocation_{i,h} \quad (3)$$

Trong đó:

- $F_{danhmuc}$ là điểm đánh giá tổng thể cho danh mục đầu tư.
 - P là tập hợp các chỉ số tiêu chí tích cực
 - N là tập hợp các chỉ số tiêu chí tiêu cực

- M là tập hợp các chỉ số tiêu chí trung bình (nếu tiêu chí “đa dạng hóa” hoặc tiêu chí “thời gian đầu tư” nằm trong top 3 tiêu chí được ưu tiên thì sẽ là tiêu chí tích cực, ngược lại là tiêu chí tiêu cực).
 - $\omega_j, \omega_k, \omega_m$ lần lượt là trọng số của các tiêu chí tương ứng
 - $x_{i,j}, x_{i,k}, x_{i,m}$ là giá trị của tiêu chí trong danh mục có id tương ứng
 - $\alpha_j, \alpha_k, \alpha_m$ lần lượt là phần trăm trọng số của các tiêu chí tương ứng
- $F(S)$ là điểm đánh giá tổng thể cho cá thể thứ i
- T là tập chứa các giá trị ở vị trí lẻ tương ứng với id của danh mục trong cá thể
 - H là tập chứa các giá trị ở vị trí chẵn tương ứng với phần trăm phân bổ của danh mục trong cá thể

Với cách tiếp cận này, thuật toán tối ưu hóa có thể tìm kiếm các nghiệm khả thi bằng cách loại bỏ những phương án vi phạm ràng buộc cứng và giảm thiểu vi phạm ràng buộc mềm.



Hình 4. Biểu diễn và minh họa cách tính của hàm Fitness

Như minh họa trong Hình 4, trọng số tiêu chí khách hàng được sử dụng để xác định hàm fitness, từ đó đánh giá và cải thiện các cá thể qua từng bước. Bảng quản thể ban đầu thể hiện các danh mục đầu tư khởi tạo với giá trị fitness ban đầu, trong khi bảng cá thể tốt nhất cho thấy danh mục có fitness tối ưu nhất, giúp xác định chiến lược phân bổ vốn hiệu quả hơn.

2.6 Khả năng mở rộng của phương pháp

Phương pháp trên có khả năng mở rộng và điều chỉnh linh hoạt theo các điều kiện thực tế khác nhau, bao gồm:

- Mở rộng số lượng danh mục đầu tư và tiêu chí đánh giá: Hệ thống có thể dễ dàng xử lý số lượng lớn danh mục đầu tư với nhiều tiêu chí khác nhau mà không làm giảm đáng kể hiệu suất tối ưu hóa.
- Thêm các ràng buộc mới: Phương pháp có thể tích hợp thêm các ràng buộc mới (cả cứng và mềm) mà không làm thay đổi quá nhiều cấu trúc thuật toán, giúp hệ thống thích ứng với yêu cầu thực tế từ khách hàng và thị trường.
- Tùy chỉnh trọng số tiêu chí theo từng khách hàng: Hệ thống có thể cá nhân hóa đề xuất danh mục đầu tư dựa trên mức độ ưu tiên khác nhau của từng khách hàng.
- Ứng dụng cho nhiều trường hợp khác: Phương pháp này có thể được mở rộng để tối ưu danh mục đầu tư trong các lĩnh vực khác như quản lý quỹ đầu tư, tối ưu hóa dòng tiền doanh nghiệp, hoặc quản lý tài sản cá nhân.

3. MINH HỌA PHƯƠNG PHÁP

3.1 Xác định trọng số tiêu chí đầu tư:

Giả sử khách hàng đánh giá các tiêu chí đầu tư theo mức độ quan trọng và gán trọng số cho từng tiêu chí như sau:

Bảng 2: Trọng số các tiêu chí ưu tiên của khách hàng

Tiêu chí	Trọng số	Trọng số (%)	Xếp hạng
Chi phí	0.3	30	1
Thời gian đầu tư	0.233333	23.333	2
Rủi ro	0.133333	13.333	3
Tính thanh khoản	0.133333	13.333	3
Tính ổn định	0.1	10	5
Đa dạng hóa	0.0666667	6.667	6
Lợi nhuận kỳ vọng	0.0333333	3.333	7

Bảng 2 thể hiện trọng số và tỷ lệ phần trăm mức độ quan trọng của từng tiêu chí. Đây là cơ sở giúp khách hàng xác định và xếp hạng các yếu tố ưu tiên trong quá trình ra quyết định đầu tư.

3.2 Khởi tạo quần thể danh mục đầu tư:

Giả sử quần thể ban đầu gồm 3 cá thể được khởi tạo ngẫu nhiên, được biểu diễn bằng các giá trị như sau:

- **Cá thể 1:** [83, 33.11, 34, 24.59, 79, 42.30]
- **Cá thể 2:** [73, 39.58, 88, 23.8, 20, 36.62]
- **Cá thể 3:** [92, 51.11, 36, 10.99, 39, 37.9)]
 - Trong đó:
- **Chỉ số danh mục:** Các giá trị ở vị trí lẻ
- **Phần trăm phân bổ:** Các giá trị ở vị trí chẵn

3.3 Công thức tính điểm đánh giá (Fitness):

Các trọng số tiêu chí tương ứng với chỉ số danh mục đầu tư:

Bảng 3. Các trọng số tiêu chí tương ứng với chỉ số danh mục đầu tư

ID	Lợi nhuận kỳ vọng	Rủi ro	Tính thanh khoản	Thời gian đầu tư	Chi phí	Ổn định của dòng tiền	Đa dạng hóa
20	8	6	7	6	6	7	7
34	9	8	7	4	7	7	8
36	7	6	5	5	6	6	5
39	9	5	8	8	7	7	8
73	7	5	7	6	6	6	7
79	8	5	8	7	6	7	8
83	8	5	9	7	6	7	8
88	9	4	8	8	7	7	9
92	8	5	8	7	6	8	7

Bảng 3 mô tả mức độ đánh giá các tiêu chí khác nhau đối với từng danh mục đầu tư, giúp so sánh và lựa chọn danh mục phù hợp với ưu tiên của khách hàng.

Áp dụng công thức (2) và (3) ở mục 2.5. ta thu được quần thể ban đầu như sau:

Bảng 4. Điểm đánh giá của các danh mục đầu tư

Fitness Danh Mục									
ID	20	34	36	39	73	79	83	88	92
Fitness	6.7794	25.556	17.001	9.668	11.467	6.223	4.446	8.335	4.779

Bảng 4 tổng hợp và tính điểm fitness của từng danh mục đầu tư, giúp đánh giá mức độ phù hợp của chúng dựa trên các tiêu chí đã xác định. Điểm fitness phản ánh chất lượng tổng thể của mỗi danh mục. Nó được tính dựa trên trọng số của các tiêu chí và giá trị tiêu chí của danh mục trong bảng 4.

Bảng 5. Quần thể ban đầu sau khi tính điểm fitness

Quần Thể Ban Đầu							
Cá thể	ID_1	Per_1	ID_2	Per_2	ID_3	Per_3	Fitness
1	83	33.11	34	24.59	79	42.30	20.9941
2	73	39.58	88	23.8	20	36.62	5.3368
3	92	51.11	36	10.99	39	37.9	5.3825

Bảng 5 hiển thị các cá thể ban đầu trong quá trình tối ưu hóa danh mục đầu tư, được tạo thành từ ba danh mục đầu tư (ID_1, ID_2, ID_3) với tỷ lệ phần trăm phân bổ (Per_1, Per_2, Per_3). Mỗi cá thể đại diện cho một chiến lược phân bổ danh mục đầu tư. Điểm fitness phản ánh chất lượng của mỗi chiến lược, giúp xác định phương án nào tối ưu hơn. Bảng này giúp đánh giá các tổ hợp danh mục đầu tư khác nhau để tìm ra phương án tối ưu nhất. Cá thể tốt nhất ở thời điểm khởi tạo là Cá thể₂: **[73, 39.58, 88, 23.8, 20, 36.62]** và giá trị F_2 là: **5.3368**

Tiến hành lặp lại 100 vòng lặp bằng thuật toán kết hợp Coati Optimization Algorithm (COA) và Multi-Verse Optimizer (MVO). Được thể hiện trong bảng dưới đây:

Bảng 6. Cá thể tốt nhất trong từng vòng lặp

Cá Thể Tốt Nhất							
Vòng lặp	ID_1	Per_1	ID_2	Per_2	ID_3	Per_3	Fitness
1	53	36.41	23	36.1	75	27.49	0.9100
2	17	34.74	18	53.14	7	12.12	0.8672
3	12	34.6	13	53.13	1	12.28	0.8274
...	...	...	...	...	...	...	...
100	23	36.88	53	35.95	71	27.16	0.7904

Bảng 6 hiển thị cá thể có giá trị fitness tốt nhất trong 100 vòng lặp của thuật toán tối ưu hóa COA-MVO trong bài toán gợi ý danh mục đầu tư. Trong mỗi vòng lặp, thuật toán tối ưu hóa tìm ra cá thể chứa tổ hợp danh mục đầu tư có giá trị fitness tốt nhất. Giá trị fitness của cá thể tốt nhất thay đổi qua từng vòng, phản ánh quá trình tối ưu hóa và khả năng tìm ra danh mục có hiệu suất tốt hơn. Sau khi chạy 100 vòng lặp thì thu được:

- Cá thể tốt nhất: [23, 36.88, 53, 35.95, 71, 27.16]
- Fitness tốt nhất: 0.7904

3.4. Tổng kết phương pháp thực hiện:

Quá trình tối ưu hóa danh mục đầu tư được thực hiện dựa trên việc xác định và đánh giá các tiêu chí quan trọng, từ đó lựa chọn danh mục phù hợp nhất.

- Bảng 2: xác định trọng số của từng tiêu chí, phản ánh mức độ ưu tiên trong quyết định đầu tư.

- Bảng 3: cung cấp thông tin về các tiêu chí tương ứng với từng danh mục, hỗ trợ quá trình so sánh và lựa chọn.
- Bảng 4: tổng hợp điểm fitness của các danh mục đầu tư, giúp đánh giá chất lượng tổng thể của từng phương án.
- Bảng 5: mô tả quần thể ban đầu, đại diện cho các chiến lược phân bổ đầu tư khác nhau trước khi áp dụng thuật toán tối ưu.
- Bảng 6: theo dõi sự thay đổi của cá thể tốt nhất qua từng vòng lặp, cho thấy quá trình cải thiện hiệu suất danh mục đầu tư.

Việc áp dụng kết hợp thuật toán Coati Optimization Algorithm (COA) và Multi-Verse Optimizer (MVO) đã giúp tối ưu hóa danh mục đầu tư theo hướng đạt điểm fitness tốt nhất. Kết quả thu được chứng minh khả năng của phương pháp này trong việc tìm ra giải pháp đầu tư hiệu quả, hỗ trợ nhà đầu tư đưa ra quyết định tối ưu hơn.

Ban đầu, quần thể khởi tạo đạt giá trị Fitness tốt nhất là $F_2 = 5.3368$ (BẢNG 5). Sau khi chạy 100 vòng lặp, giá trị tốt nhất đã giảm xuống còn $F = 0.7904$ (BẢNG 6). Sự thay đổi đáng kể này cho thấy cơ chế cập nhật của thuật toán kết hợp COA-MVO là rất hiệu quả. Việc giảm từ $F = 5.3368$ xuống $F = 0.7904$ phản ánh khả năng hội tụ tốt của thuật toán kết hợp.

Kết quả này đã chứng tỏ việc kết hợp hai thuật toán Coati Optimization Algorithm (COA) và Multi-Verse Optimizer (MVO) mang lại tiềm năng vượt trội trong việc tìm kiếm nghiệm tối ưu và khả năng vượt cực trị cục bộ trong các không gian tìm kiếm phức tạp.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu đã đề xuất phương pháp kết hợp giữa thuật toán Coati Optimization Algorithm (COA) và Multi-Verse Optimizer (MVO) nhằm giải quyết bài toán gợi ý danh mục đầu tư tối ưu trong môi trường tài chính. Phương pháp này khai thác hiệu quả khả năng cân bằng giữa khám phá và khai thác không gian tìm kiếm, giúp tối ưu đồng thời các yếu tố như lợi nhuận và rủi ro, từ đó nâng cao chất lượng danh mục đầu tư. Tuy nhiên, phương pháp vẫn tồn tại một số hạn chế như: hiệu quả thuật toán phụ thuộc nhiều vào việc lựa chọn tham số ban đầu, kết quả có thể thiếu ổn định giữa các lần chạy do tính ngẫu nhiên cao, và khả năng mở rộng trong môi trường thực tế với dữ liệu lớn cần được kiểm chứng thêm.

Trong tương lai, nhóm nghiên cứu định hướng tích hợp thêm các thuật toán metaheuristic khác như thuật toán di truyền (GA) hoặc tối ưu bầy đàn (PSO) để tăng cường tốc độ hội tụ và khả năng thích ứng với thị trường biến động. Việc mở rộng này sẽ góp phần hoàn thiện mô hình gợi ý danh mục đầu tư và nâng cao hiệu quả ứng dụng trong các tình huống thực tế có tính phức tạp và quy mô lớn.

TÀI LIỆU THAM KHẢO

1. Chang, T.-J., Yang, S.-C. and Chang, K.-J. (2009) “Portfolio optimization problems in different risk measures using genetic algorithm,” *Expert systems with applications*, 36(7), pp. 10529–10537.
2. Markowitz, H., 1952. Portfolio selection. *Journal of Finance*.
3. Cura, T. (2009) “Particle swarm optimization approach to portfolio optimization,” *Nonlinear analysis. Real world applications*, 10(4), pp. 2396–2406.

4. Zhu, H. et al. (2011) “Particle Swarm Optimization (PSO) for the constrained portfolio optimization problem,” *Expert systems with applications*, 38(8), pp. 10161–10169.
5. Mirjalili, S., Mirjalili, S.M. and Hatamlou, A. (2016) “Multi-Verse Optimizer: a nature-inspired algorithm for global optimization,” *Neural computing & applications*, 27(2), pp. 495–513.
6. Dehghani, M. et al. (2023) “Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems,” *Knowledge-based systems*, 259(110011), p. 110011.
7. Mirjalili, S. et al. (2017) “Optimization of problems with multiple objectives using the multi-verse optimization algorithm,” *Knowledge-based systems*, 134, pp. 50–71.

ABSTRACT

PROPOSED MODEL COMBINING COATI AND MVO IN AN INVESTMENT DECISION SUPPORT SYSTEM

Huỳnh Thị Châu Lan^{1,*}, Trần Huỳnh Phụng Quyên¹, Dương Trọng Đức¹, Hứa Vĩnh Khang¹,
Đinh Nguyễn Trọng Nghĩa¹

¹ *Ho Chi Minh City University of Industry and Trade*

*Email:lanhtc@huit.edu.vn

Investment decision-making is multifaceted process involving numerous factors, for which effective optimization methods are necessary to optimize portfolio selection. The current paper presents a new strategy that combines the Coati Optimization Algorithm (COA) and the Multiverse Optimizer (MVO) to develop a smart investment decision support system. The Coati algorithm, developed based on coati foraging and social patterns, is paired with the MVO's exploration-exploitation mechanism to optimize portfolio allocation approaches. This paper provides a conceptual model, algorithmic structure, and workflow diagram illustrating the integration of Coati and MVO. Empirical validation is outside the perspective of this research, though the theoretical model provides a promising direction for future research and financial decision-making in practice.

Keywords: COA, MVO, optimization, portfolio recommendation.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Huỳnh Thị Châu Lan

Số điện thoại liên hệ: 0918816682

BÀI TOÁN K-ANONYMITY VÀ PHƯƠNG PHÁP GIẢI BẰNG META-HEURISTIC

Lê Hoàng Minh Nhật¹, Đinh Nguyễn Trọng Nghĩa^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: nghiadnt@huit.edu.vn

Ngày nhận bài: 12/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Bài báo đề xuất cách tiếp cận dựa trên thuật toán metaheuristic để giải quyết bài toán ẩn danh dữ liệu k-anonymity, giúp bảo vệ quyền riêng tư. Giải pháp này tối ưu đồng thời ràng buộc cứng (đảm bảo mỗi nhóm có ít nhất k bản ghi giống nhau) và ràng buộc mềm (giảm thiểu mất mát thông tin và duy trì độ chính xác của dữ liệu), nhằm cân bằng giữa bảo mật và tính hữu ích của dữ liệu. So với các kỹ thuật truyền thống như Mondrian hay thuật toán dựa trên cây quyết định, phương pháp metaheuristic cải thiện hiệu suất tính toán và tăng khả năng tìm kiếm nghiệm tối ưu trong không gian giải pháp rộng lớn. Mô hình đề xuất được xây dựng dưới dạng một bài toán tối ưu, cho phép linh hoạt trong việc lựa chọn thuật toán metaheuristic phù hợp để đạt hiệu quả cao hơn. Kết quả phân tích cho thấy tiềm năng ứng dụng rộng rãi trong các lĩnh vực như y tế, tài chính và truyền thông, đồng thời mở ra hướng nghiên cứu nhằm nâng cao hiệu quả ẩn danh dữ liệu.

Từ khóa: tối ưu hóa, thuật toán metaheuristic, ẩn danh dữ liệu, k-anonymity, bảo vệ quyền riêng tư.

1. GIỚI THIỆU

Quyền riêng tư về dữ liệu cá nhân đã nổi lên như một mối quan tâm quan trọng trong thời đại số, bao gồm quyền của cá nhân trong việc kiểm soát thông tin cá nhân của họ giữa những tiến bộ công nghệ nhanh chóng. Có nhiều khuôn khổ pháp lý xung quanh quyền riêng tư dữ liệu như đạo luật bảo vệ dữ liệu cá nhân Kỹ thuật số năm 2023 ở Ấn Độ, nhằm giải quyết việc truy cập trái phép và lạm dụng dữ liệu cá nhân, phản ánh nhu cầu ngày càng tăng về các biện pháp bảo vệ mạnh mẽ trong kỷ nguyên dữ liệu lớn và tương tác kỹ thuật số [1]. Hơn nữa, việc số hóa các dịch vụ ngày càng tăng đã làm tăng nhận thức và lo lắng của cộng đồng về quyền riêng tư dữ liệu, đòi hỏi các chính sách sáng tạo và thực tiễn tốt nhất để giảm thiểu rủi ro liên quan đến các công nghệ mới nổi như trí tuệ nhân tạo và điện toán đám mây [2].

K-anonymity đã được định nghĩa bởi Latanya Sweeney vào năm 1997, đây là một mô hình bảo vệ quyền riêng tư hiệu quả trong quá trình ẩn danh hóa dữ liệu [3]. Theo nguyên lý này, một tập dữ liệu được gọi là k-anonymous khi với bất kỳ bản ghi nào trong tập dữ liệu đó, tồn tại ít nhất $k - 1$ bản ghi khác có cùng giá trị các thuộc tính bán định danh QI (quasi-identifiers). Nhờ đó, việc liên kết các thông tin cá nhân với một cá thể cụ thể trở nên khó khăn hơn, qua đó giảm thiểu nguy cơ lộ lọt thông tin cá nhân trong các quá trình phân tích và chia sẻ dữ liệu [4].

Mặc dù k-anonymity mang lại nhiều lợi ích trong việc đảm bảo tính ẩn danh của dữ liệu, nhưng việc áp dụng mô hình này vào các tập dữ liệu lớn không phải là điều đơn giản. Một trong những thách thức chính là độ phức tạp tính toán khi phải xử lý các tập dữ liệu với số lượng bản ghi khổng lồ, đòi hỏi phải thực hiện các thao tác tổng quát hóa hoặc loại bỏ thông tin trên nhiều thuộc tính cùng một lúc. Ngoài ra, còn tồn tại sự đánh đổi giữa mức độ bảo vệ quyền riêng tư và tiện ích của dữ liệu. Khi áp dụng các phương pháp bảo vệ mạnh mẽ nhằm tăng cường tính ẩn danh, thông tin ban đầu thường bị làm mất đi một phần giá trị phân tích, gây ảnh hưởng không nhỏ đến các ứng dụng trong khai thác dữ liệu [5]. Cách tiếp cận truyền thống sử dụng các thuật toán như Mondrian hoặc thuật toán dựa trên cây quyết định thường không đạt được nghiệm tối ưu khi kích thước dữ liệu tăng lên [6]. Trong bối cảnh này, các thuật toán metaheuristic đã được đề xuất nhằm tìm kiếm nghiệm tốt nhất trong không gian tìm kiếm mà không cần duyệt hết tất cả các phương án khả thi.

Các phương pháp tiếp cận bằng thuật toán meta-heuristic ngày càng được áp dụng để tăng cường tính ẩn danh k trong ẩn danh dữ liệu, giải quyết các thách thức như khả năng mở rộng và tối ưu hóa trong các tập dữ liệu lớn. Phương pháp tiếp cận Kgen [7] hay sử dụng thuật toán di truyền (Genetic Algorithm – GA) [8] đã minh họa cho một phương pháp meta-heuristic đạt được hiệu quả tính ẩn danh k bằng cách tìm ra các giải pháp tối ưu giả trong một khung thời gian hợp lý, ngay cả với các bộ dữ liệu lớn. Cách tiếp cận này duy trì mức độ ẩn danh cao trong khi vẫn duy trì khả năng sử dụng dữ liệu và giảm thiểu mất mát thông tin, như được chứng minh trong các ứng dụng trong các tập dữ liệu chứa nhiều thuộc tính [7].

Nghiên cứu này hướng tới việc đề xuất một phương pháp tổng quát dựa trên thuật toán metaheuristic để giải quyết bài toán k-anonymity nhằm mục tiêu tối ưu hóa đồng thời các ràng buộc cứng và mềm của bài toán, đảm bảo rằng dữ liệu sau khi ẩn danh không chỉ đáp ứng được yêu cầu bảo vệ quyền riêng tư mà còn giữ được giá trị phân tích cần thiết. Trong đó, các ràng buộc cứng được đặt ra nhằm duy trì mức độ ẩn danh tối thiểu cho mỗi nhóm dữ liệu, còn các ràng buộc mềm được tối ưu hóa để giảm thiểu mất mát thông tin, từ đó tạo ra sự cân bằng hợp lý giữa tính bảo mật và tiện ích của dữ liệu. Cách tiếp cận dựa trên thuật toán metaheuristic, với khả năng tìm kiếm nghiệm tối ưu trong không gian giải pháp rộng lớn, mang lại hiệu quả tính toán cao và khả năng áp dụng thực tiễn tốt hơn so với các phương pháp truyền thống [7].

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Giới thiệu bài toán

Định nghĩa 1. (Bài toán k-anonymity). Cho một tập dữ liệu D gồm n bản ghi, mỗi bản ghi chứa một tập hợp các thuộc tính $A = \{A_1, A_2, \dots, A_m\}$. Gọi $QI \subseteq A$ là tập thuộc tính bán định danh QI và $S \subseteq A$ là tập thuộc tính nhạy cảm. Một phép biến đổi dữ liệu được thực hiện trên D để tạo ra một tập dữ liệu ẩn danh D' , sao cho mỗi bản ghi trong D' có cùng giá trị trên tập QI với ít nhất $k-1$ bản ghi khác trong tập dữ liệu [3]. Ví dụ:

Bảng 1. Bảng dữ liệu gốc chưa thực hiện phép biến đổi

STT	Tuổi	Giới tính	Mã bưu điện	Bệnh
1	25	Nam	10001	Cảm cúm
2	31	Nữ	10002	Đau đầu
3	24	Nam	10001	Viêm họng
4	34	Nữ	10003	Cảm cúm
5	29	Nữ	10002	Đau đầu

Bảng 2. Bảng dữ liệu sau khi áp dụng k-anonymity (với k = 2)

STT	Tuổi	Giới tính	Mã bưu điện	Bệnh
1	20-30	Nam	1000*	Cảm cúm
2	30-40	Nữ	1000*	Đau đầu
3	20-30	Nam	1000*	Viêm họng
4	30-40	Nữ	1000*	Cảm cúm
5	20-30	Nữ	1000*	Đau đầu

Tuổi: Được làm mờ thành khoảng (20-30, 30-40) để đảm bảo mỗi nhóm có ít nhất 2 bản ghi.

Mã bưu điện: Được làm mờ bằng cách thay thế chữ số cuối cùng bằng dấu * (1000*).

Giới tính và Bệnh: Giữ nguyên vì chúng đã đủ phân tán để đảm bảo k-anonymity.

→ Sau khi áp dụng k-anonymity, mỗi bản ghi không thể phân biệt được với ít nhất 1 bản ghi khác (do k = 2) trong cùng nhóm, đảm bảo tính riêng tư của dữ liệu.

Định nghĩa 2. (Vi phạm k-anonymity – Fail to K-anonymity – FTK). Là số lượng bản ghi trong tập dữ liệu D' có một tổ hợp giá trị trên các thuộc tính bán định danh QI xuất hiện ít hơn k lần. Điều này có nghĩa là sau khi thực hiện phương pháp k-anonymity, vẫn tồn tại một số nhóm bản ghi không đạt tiêu chí ẩn danh.

Định nghĩa 3. (Độ chính xác phân loại – Classification Accuracy - CA). Là mức độ có thể duy trì độ chính xác phân loại sau khi tập dữ liệu được ẩn danh. CA được tính dựa trên sự khác biệt về hiệu suất của mô hình trước và sau khi áp dụng k-anonymity. Giá trị CA càng cao, dữ liệu ẩn danh càng hữu ích. CA là một chỉ số quan trọng để đánh giá hiệu quả của mô hình phân loại sau khi áp dụng k-anonymity [3]. Chỉ số này đo lường tỷ lệ phần trăm các bản ghi được phân loại chính xác so với tổng số bản ghi, được tính theo công thức:

$$CA = \frac{\text{Số lượng bản ghi được phân loại đúng}}{\text{Tổng số bản ghi}} \times 100\% \quad (1)$$

Trong đó:

- Số lượng bản ghi được phân loại đúng: Là số lượng bản ghi mà mô hình phân loại dự đoán đúng so với nhãn thực tế.
- Tổng số bản ghi: Là tổng số bản ghi trong tập dữ liệu được sử dụng để đánh giá.

Ví dụ: Giả sử có một tập dữ liệu gồm 1000 bản ghi, sau khi áp dụng k-anonymity và chạy mô hình phân loại, mô hình dự đoán đúng 850 bản ghi. Khi đó, độ chính xác phân loại (CA) được tính như sau:

$$CA = \frac{850}{1000} \times 100\% = 85\%$$

Định nghĩa 4. (Mất mát thông tin – Information Loss - IL). Là mức độ thay đổi dữ liệu sau khi thực hiện k-anonymity, thường được đo lường dựa trên các chỉ số như độ biến đổi của dữ liệu hoặc sự sai lệch so với tập dữ liệu gốc. Giá trị IL càng cao thì mức độ mất mát thông tin càng lớn. IL được tính dựa trên mức tổng quát hóa của từng thuộc tính bán định danh thay vì khoảng giá trị ban đầu. Công thức tính IL được xác định như sau:

$$NCP(A) = \frac{\text{Số mức tổng quát hóa đã áp dụng}}{\text{Số mức tổng quát hóa tối đa}} \quad (2)$$

Ví dụ:

- Nếu một thuộc tính tổng quát hóa ở mức 1 trên 2 mức tối đa → IL(A) = 1/2.

- Nếu thuộc tính tổng quát hóa ở mức 2 trên 2 mức tối đa $\rightarrow IL(A) = 2/2 = 1$.

Sau đó, chỉ số IL của một nghiệm X (đại diện cho một cấu hình tổng quát hóa của các thuộc tính bán định danh) thường được tính bằng cách tổng hợp (ví dụ như bằng trung bình cộng) các giá trị Normalized Certainty Penalty – NCP của tất cả các thuộc tính [8], từ đó cho một thước đo duy nhất về mức độ mất mát thông tin sau quá trình ẩn danh, được tính theo công thức:

$$IL(X) = \frac{1}{n} \sum_{i=1}^n NCP(A_i) \quad (3)$$

Ví dụ: với $n = 2$ (vì có 2 thuộc tính Tuổi và Nghề nghiệp đã tính ở trên), ta có:

$$IL(X) = \frac{NCP(\text{Tuổi}) + NCP(\text{Nghề nghiệp})}{2} = \frac{0.28 + 0.3}{2} = 0.29$$

Định nghĩa 5. (Tham số bảo mật – Ký hiệu là M). Là mức độ đảm bảo tính ẩn danh của dữ liệu sau khi thực hiện k -anonymity, được đo lường dựa trên tiêu chí rằng mỗi bản ghi trong tập dữ liệu ẩn danh có cùng giá trị trên tập thuộc tính định danh gián tiếp với ít nhất $k-1$ bản ghi khác. Tức là nếu sau khi biến đổi dữ liệu, tập D' thỏa mãn điều kiện k -anonymity thì giá trị $M = 0$, ngược lại, nếu tồn tại các nhóm bản ghi mà tổ hợp giá trị trên các thuộc tính định danh gián tiếp xuất hiện ít hơn k lần thì $M=1$. Giá trị M càng thấp thì mức độ bảo mật của dữ liệu càng cao.

$$M(X) = \begin{cases} 0, & \text{nếu } X \text{ thỏa mãn } k - \text{anonymity} \\ 1, & \text{ngược lại} \end{cases} \quad (4)$$

2.2 Đầu vào của bài toán

Các bản ghi dữ liệu gốc được tổ chức theo dạng bảng, mỗi dòng là một cá nhân, mỗi cột là một thuộc tính. Tập dữ liệu chứa các thuộc tính số và thuộc tính phân loại. Bài toán ẩn danh dữ liệu theo k -anonymity bao gồm các tập dữ liệu đầu vào chính sau:

- **Tập dữ liệu gốc D chứa I bản ghi:** Mỗi bản ghi trong tập dữ liệu bao gồm K thuộc tính.
- **Tham số k :** Số nguyên xác định số lượng tối thiểu các bản ghi cần có trong mỗi nhóm tương đương nhằm đảm bảo tính ẩn danh.
- **Các tham số của thuật toán metaheuristic:** Dùng để tối ưu hóa quá trình phân nhóm và ẩn danh dữ liệu.

2.3 Mô hình hóa bài toán

Về mặt toán học, bài toán có thể mô hình hóa như sau:

Gọi $T = \{t_1, t_2, \dots, t_n\}$ là tập các bản ghi, mỗi bản ghi t_i gồm các thuộc tính bán định danh QI $Q = \{q_1, q_2, \dots, q_m\}$ và thuộc tính nhạy cảm S.

Mục tiêu là phân nhóm T thành các lớp tương đương $E_1, E_2, \dots, E_n$ sao cho $|E_j| \geq k \forall j$

Mỗi nghiệm (cá thể) trong không gian tìm kiếm biểu diễn một phương án ẩn danh cho tập dữ liệu gốc. Cụ thể, mỗi cá thể thể hiện cách phân nhóm các bản ghi thành các lớp tương đương sao cho mỗi nhóm có ít nhất k bản ghi (đáp ứng yêu cầu k -anonymity), đồng thời xác định mức độ tổng quát hóa của các thuộc tính bán định danh được thể hiện như sau:

$$X = \{g_1, g_2, \dots, g_n\}$$

Trong đó:

- g_i là mức tổng quát hóa được áp dụng cho thuộc tính bán định danh thứ i .
- Giá trị của g_i nằm trong một miền giá trị hữu hạn, tương ứng với các mức tổng quát hóa được xác định trước (ví dụ: mã bưu chính có thể tổng quát hóa theo cấp độ "quận" $\rightarrow$ "thành phố" $\rightarrow$ "quốc gia").

- Các thuộc tính nhạy cảm không thay đổi và không xuất hiện trong vector nghiệm.

Ví dụ: Một tập dữ liệu có ba thuộc tính bán định danh gồm Tuổi, Nghề nghiệp, Địa chỉ có thể có các mức tổng quát hóa như sau:

- **Tuổi:** {Chính xác, Khoảng [10 tuổi], Khoảng [20 tuổi]}
- **Nghề nghiệp:** {Chi tiết, Ngành nghề chung}
- **Địa chỉ:** {Quận, Thành phố, Quốc gia}

Một nghiệm có thể là $X = \{1,0,2\}$, tức là:

- Thuộc tính Tuổi tổng quát hóa thành khoảng 10 tuổi.
- Thuộc tính Nghề nghiệp giữ nguyên.
- Thuộc tính Địa chỉ tổng quát hóa lên mức "Quốc gia".

2.4 Ràng buộc bài toán

Ràng buộc cứng (Hard Constraints - HC): Đây là các ràng buộc bắt buộc phải tuân thủ. Nếu vi phạm, nghiệm bị loại bỏ vì không khả thi. Các ràng buộc này bao gồm:

- Mỗi nhóm trong tập dữ liệu ẩn danh phải có ít nhất k bản ghi để đảm bảo k-anonymity.
- Không thay đổi thuộc tính nhạy cảm (SA – Sensitive Attribute).

Ràng buộc mềm (Soft Constraints - SC): Đây là các ràng buộc không bắt buộc nhưng nên được tối ưu hóa để cải thiện chất lượng ẩn danh. Các ràng buộc này có thể bị vi phạm nếu không còn phương án khả thi khác, nhưng sẽ bị tính điểm phạt. Một số ràng buộc mềm bao gồm:

- Giảm thiểu mất mát thông tin (Information Loss - IL) để bảo toàn tính hữu ích của dữ liệu.
- Duy trì độ chính xác phân loại (Classification Accuracy - CA) cao nhất có thể.

2.5 Xây dựng hàm mục tiêu

Hàm mục tiêu kết hợp ba yếu tố IL, M và CA để đảm bảo rằng thuật toán chỉ chọn những nghiệm thỏa mãn k-anonymity đồng thời giảm thiểu mất mát thông tin và tối đa hóa chỉ số CA. Hàm mục tiêu được định nghĩa như sau:

$$F(X) = \alpha IL(X) + \lambda M(X) + \gamma(1 - CA(X)) \quad (5)$$

Trong đó:

- α : Hệ số xác định mức độ ưu tiên của việc giảm thiểu mất mát thông tin (IL) trong quá trình tối ưu hóa.
- λ : Hệ số phạt được lựa chọn với giá trị lớn, nhằm ưu tiên các nghiệm thỏa mãn k-anonymity, bất kể mức mất mát thông tin có thấp hay cao.
- γ : Hệ số phạt cho việc CA thấp. Được quy định dựa trên mức độ quan trọng của CA so với các chỉ số khác.
- $IL(X)$: Mất mát thông tin của nghiệm X , được tính bằng trung bình cộng của các giá trị NCP được tính bằng công thức (3).
- $M(X)$: Tham số bảo mật được định nghĩa là một biến nhị phân tại công thức (4).
- $CA(X)$: Độ chính xác phân loại của nghiệm X . $(1 - CA(X))$ là một đại lượng phạt nếu CA thấp. Khi $CA(X)$ càng gần 1 (100%), thì $(1 - CA(X))$ càng gần 0 và ngược lại.

Việc xây dựng hàm mục tiêu theo dạng tổng trọng số cho phép chúng ta điều chỉnh mức độ ưu tiên giữa việc giảm thiểu mất mát thông tin, đảm bảo tính bảo mật và độ chính xác phân loại của dữ liệu. Cụ thể, nếu λ được chọn với giá trị lớn, thì bất kỳ nghiệm nào không thỏa mãn yêu cầu k-anonymity đều sẽ bị phạt nặng, ngay cả khi mức mất mát thông tin $IL(X)$ thấp. Hoặc nếu độ chính xác phân loại quá thấp, cũng sẽ dẫn đến bị phạt nặng. Điều này giúp định hướng quá trình tìm kiếm của thuật toán metaheuristic về phía những nghiệm đảm bảo an toàn dữ liệu và độ chính xác phân loại trước tiên, sau đó mới tối ưu hóa thêm về mặt IL. Ngược lại, nếu λ hoặc γ quá nhỏ, thuật toán có thể ưu tiên những nghiệm có IL thấp nhưng lại không đáp ứng yêu cầu bảo mật hoặc đảm bảo độ chính xác phân loại cao [9].

Do đó, cấu trúc hàm mục tiêu được xây dựng theo dạng trên là một cách tiếp cận hiệu quả, cho phép cân bằng giữa ba tiêu chí cơ bản của bài toán: tính hữu dụng (thông qua chỉ số IL), mức độ duy trì độ chính xác phân loại sau khi tập dữ liệu được ẩn danh CA và tính bảo mật (thông qua tham số $M(X)$).

3. VÍ DỤ MINH HỌA

Lấy ví dụ một tập dữ liệu nhỏ với các giá trị xác định như Bảng 3:

Bảng 3. Tập dữ liệu minh họa gốc

STT	Tuổi	Địa điểm	Nghề nghiệp
1	25	TP. HCM	Kỹ sư
2	30	Hà Nội	Kế toán
3	24	TP. HCM	Bác sĩ
4	34	Đà Nẵng	Giáo viên
5	29	Hà Nội	Kỹ sư

Nghiệm $X = \{1, 0, 2\}$:

- **Tuổi:** Tổng quát hóa mức 1 (nhóm theo thập kỷ: 20-29, 30-39).
- **Địa điểm:** Tổng quát hóa mức 0 (giữ nguyên giá trị chi tiết).
- **Nghề nghiệp:** Tổng quát hóa mức 2 (Lao động trí thức).

Quy định mức tổng quát hóa cho các thuộc tính:

Tuổi

- Mức 0: Chi tiết (giữ nguyên giá trị ban đầu).
- Mức 1: Nhóm theo thập kỷ (20-29, 30-39, ...).
- Mức 2: Nhóm theo khoảng lớn hơn (20-39, 40-59, ...).

Địa điểm

- Mức 0: Giữ nguyên tên thành phố.
- Mức 1: Tổng quát hóa thành miền (Bắc, Trung, Nam).
- Mức 2: Tổng quát hóa thành đất nước.

Nghề nghiệp

- Mức 0: Chi tiết (Kỹ sư, Bác sĩ, Giáo viên, ...).
- Mức 1: Nhóm theo lĩnh vực (Kỹ thuật, Y tế, Giáo dục, ...).

- Mức 2: Nhóm chung thành "Lao động trí thức" hoặc "Lao động phổ thông".

Bước 1: Tổng quát hóa theo nghiệm $X = \{1, 0, 2\}$, kết quả thể hiện tại Bảng 4.

Bảng 4. Tập dữ liệu minh họa sau khi tổng quát hóa theo nghiệm $X = \{1, 0, 2\}$

STT	Tuổi	Địa điểm	Nghề nghiệp
1	20-29	TP. HCM	Lao động trí thức
2	30-39	Hà Nội	Lao động trí thức
3	20-29	TP. HCM	Lao động trí thức
4	30-39	Đà Nẵng	Lao động trí thức
5	20-29	Hà Nội	Lao động trí thức

Bước 2: Tính chỉ số mất mát thông tin $IL(X)$. Với $n = 3$ (3 thuộc tính QI) dựa trên mức tổng quát hóa của các thuộc tính bằng công thức số (2) và số (3):

Bước 2.1: Tuổi tổng quát hóa mức 1 (nhóm theo thập kỷ: 20-29, 30-39) nên chỉ số $IL(\text{Tuổi})$ là:

$$IL(\text{Tuổi}) = \frac{1}{2} = 0.5$$

Bước 2.2: Nghề nghiệp tổng quát hóa mức 2 (gộp tất cả nghề thành nhóm "Lao động trí thức")

$$IL(\text{Nghề nghiệp}) = \frac{2}{2} = 1$$

Bước 2.3: Địa điểm giữ nguyên nên:

$$IL(\text{Địa điểm}) = 0$$

Bước 2.4: Tính $IL(X)$ dựa trên giá trị trung bình cộng IL của các thuộc tính (công thức số (3)):

$$IL(X) = \frac{0.5 + 1 + 0}{3} = 0.5$$

Bước 3: Tính tham số bảo mật $M(X)$ theo công thức số (4), kiểm tra xem các nhóm dữ liệu có thỏa mãn $k = 2$ hay không, các nhóm dữ liệu sau khi được tổng quát hóa:

- Nhóm 1: [20-29, TP. HCM, Lao động trí thức] → 2 bản ghi (STT 1 và 3).
- Nhóm 2: [30-39, Hà Nội, Lao động trí thức] → 1 bản ghi (STT 2).
- Nhóm 3: [30-39, Đà Nẵng, Lao động trí thức] → 1 bản ghi (STT 4).
- Nhóm 4: [20-29, Hà Nội, Lao động trí thức] → 1 bản ghi (STT 5).

→ Nhóm 2, 3, 4 không thỏa mãn $k = 2$, vì vậy $M(X) = 3$ (có 3 vi phạm k -anonymity theo công thức số (4)).

Bước 4: Tính độ chính xác phân loại CA sau khi ẩn danh dữ liệu. Tuy nhiên, để làm điều đó, cần có thông tin về cách mô hình phân loại hoạt động trên tập dữ liệu trước và sau khi áp dụng k -anonymity. Điều này đòi hỏi một mô hình phân loại cụ thể và kết quả dự đoán trên cả hai tập dữ liệu để so sánh độ chính xác. Vì thế, trong nghiên cứu này giả sử trong tập dữ liệu ban đầu, mô hình phân loại dự đoán đúng 5/5 bản ghi. Sau khi tổng quát hóa, mô hình chỉ dự đoán đúng 3/5 bản ghi. Chỉ số CA sẽ được tính theo công thức số (1):

$$CA(X) = \frac{3}{5} = 0.6$$

Bước 5: Tính hàm mục tiêu $F(X)$, ở đây giả sử các trọng số trong hàm mục tiêu:

- Hệ số phạt cho việc vi phạm k-anonymity $\lambda = 1000$ (hệ số phạt rất lớn đảm bảo ưu tiên loại bỏ những nghiệm không thỏa mãn k-anonymity)
- Hệ số phạt cho việc độ chính xác phân loại thấp $\gamma = 100$ (nếu CA thấp, dữ liệu sau khi ẩn danh có thể không còn hữu ích cho các mô hình phân tích, cần đặt trọng số đủ lớn để giữ được độ chính xác nhưng không vượt quá mức λ).
- Hệ số phạt cho chỉ số IL $\alpha = 5$ (nếu IL quá cao, dữ liệu bị tổng quát hóa quá mức làm giảm tính hữu ích, tùy vào mức độ chấp nhận mất mát thông tin mà ta chọn).

Vậy giá trị hàm mục tiêu cần tính là:

$$F(X) = 5 \times 0.5 + 1000 \times 3 + 100 \times (1 - 0.6) = 3042.5$$

Bước 6: Kết luận nghiệm $X = \{1, 0, 2\}$ có:

- Chỉ số mất mát thông tin: $IL = 0.5$.
- Số vi phạm k-anonymity: $M = 3$ (3 vi phạm).
- Độ chính xác phân loại: $CA = 0.6 = 60\%$.

→ Giá trị hàm mục tiêu $F(X) = 3042.5$. Nghiệm này không đảm bảo tính bảo mật (vi phạm k-anonymity), dẫn đến giá trị hàm mục tiêu rất cao do hệ số phạt λ lớn. Việc giảm độ chính xác phân loại $CA(X)$ cũng làm tăng giá trị hàm mục tiêu, cho thấy sự đánh đổi giữa bảo mật và tính hữu ích của dữ liệu.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất một phương pháp tổng quát dựa trên thuật toán metaheuristic để giải quyết bài toán k-anonymity trong bảo vệ quyền riêng tư dữ liệu. Cách tiếp cận này giúp tối ưu hóa đồng thời các ràng buộc cứng - đảm bảo tính k-anonymity, và các ràng buộc mềm - giảm thiểu mất mát thông tin và duy trì độ chính xác phân loại. So với các phương pháp truyền thống như Mondrian hay các thuật toán dựa trên cây quyết định, phương pháp metaheuristic cho thấy tiềm năng đáng kể trong việc nâng cao hiệu suất xử lý dữ liệu lớn, đồng thời cân bằng tốt giữa bảo mật và tính hữu ích của dữ liệu.

Mặc dù đạt được những kết quả tích cực, phương pháp này vẫn có một số hạn chế cần khắc phục. Hiệu quả của thuật toán phụ thuộc vào lựa chọn tham số và chiến lược tìm kiếm của metaheuristic, đòi hỏi thử nghiệm thực nghiệm trên nhiều tập dữ liệu khác nhau để xác định cấu hình tối ưu. Ngoài ra, bài toán k-anonymity mới chỉ giải quyết một khía cạnh của bảo mật dữ liệu, trong khi các mô hình nâng cao như l-diversity hay t-closeness có thể giúp cải thiện khả năng bảo vệ thông tin nhạy cảm một cách toàn diện hơn [12].

Trong tương lai, các thuật toán tối ưu metaheuristic như Thuật toán Di truyền (Genetic Algorithm – GA), Tối ưu Bầy đàn (Particle Swarm Optimization – PSO) [13] và Thuật toán Sói Xám (Grey Wolf Optimization – GWO) [14] sẽ được áp dụng. Những phương pháp này giúp tăng cường khả năng tìm kiếm nghiệm tối ưu trong không gian giải pháp rộng lớn, đồng thời giảm độ phức tạp của bài toán. GA sẽ cải thiện nghiệm bằng cơ chế tiến hóa tự nhiên, PSO tận dụng mô hình bầy đàn để tăng tốc độ hội tụ, trong khi GWO giúp cân bằng giữa khai thác và khám phá. được mở rộng theo nhiều hướng khác nhau. Đồng thời, mở rộng phương pháp sang các mô hình bảo mật nâng cao như l-diversity hoặc t-closeness sẽ góp phần nâng cao mức độ an toàn của dữ liệu. Bên cạnh đó, thử nghiệm trên nhiều tập dữ liệu thực tế thuộc các lĩnh vực như y tế, tài chính, truyền thông, ... là cần thiết để đánh giá hiệu quả của thuật

toán trong môi trường ứng dụng thực tế. Ngoài ra, phát triển một hệ thống tự động hóa tích hợp thuật toán metaheuristic sẽ hỗ trợ người dùng trong quá trình ẩn danh dữ liệu một cách linh hoạt và hiệu quả hơn.

TÀI LIỆU THAM KHẢO

1. J. Thapa, Data Privacy Vis- A- Vis the Digital Personal Data Protection Act, 2023, International Journal For Multidisciplinary Research, **6**(3), (2024), pp. 1-10.
2. Kavak, Privacy of Information and Data, IGI Global, pp. 155–187, USA (2024).
3. L. Sweeney, K-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, **10**(5), (2002), pp. 557–570.
4. J. Qian, H. Jiang, Y. Yu, H. Wang, D. Miao, Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions, Expert Systems With Applications, **238**, (2023), pp. 1-15.
5. C. Aggarwal, P. S. Yu, A general survey of privacy-preserving data mining models and algorithms, Privacy-Preserving Data Mining, Springer, pp. 11–52, USA (2008).
6. K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity, Proceedings of the 22nd International Conference on Data Engineering, Atlanta, USA (2006).
7. D. Pascale, G. Cascavilla, D. A. Tamburri, W.-J. van den Heuvel, Real-world K-Anonymity Applications: the KGen approach and its evaluation in Fraudulent Transactions, arXiv.org, abs/2204.01533, (2022), pp. 1-15.
8. D. E. Goldberg, Genetic algorithms in search, optimization and machine learning, Addison-Wesley Longman Publishing Co., Inc., USA (1989).
9. E. G. Talbi, Metaheuristics: From Design to Implementation, Wiley, USA (2009).
10. X. Xu, X. Li, Y. Zhou, A Framework for Privacy Preserving Data Publishing, IEEE International Conference on Data Mining, Hong Kong, China (2006).
11. Y. Li, J. Liu, M. Zhang, A Hybrid Metaheuristic Approach for Privacy Preserving Data Mining, IEEE International Conference on Data Engineering, Istanbul, Turkey (2007).
12. N. Li, T. Li, S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey (2007).
13. F. Marini, B. Walczak, Particle swarm optimization (PSO). A tutorial, Chemometrics and Intelligent Laboratory Systems, **149**, (2015), pp. 153–165.
14. S. Mirjalili, S. M. Mirjalili, A. Lewis, Grey wolf optimizer, Advances in Engineering Software, **69**, (2014), pp. 46–61.

SABSTRACT

THE K-ANONYMITY PROBLEM AND META-HEURISTIC-BASED SOLUTION APPROACHES

Le Hoang Minh Nhat¹, Dinh Nguyen Trong Nghia^{1, *}

¹ *Ho Chi Minh City University of Industry and Trade*

*Email: nghiadnt@huit.edu.vn

The newspaper article proposes an approach based on metaheuristic algorithm to address the k-anonymity data anonymity problem and protect privacy. This solution simultaneously optimizes hard constraints (making sure that each group has at least k identical records) and soft constraints (minimizing information loss and maintaining the accuracy of data), aiming to balance between security and the usefulness of data. Compared with traditional techniques such as Mondrian or decision tree-based algorithms, the metaheuristic approach improves computational efficiency and increases the ability to find optimal solutions in a large solution space. The proposed method is modeled as an optimization problem, allowing flexibility in choosing the metaheuristic algorithm to achieve higher efficiency. The analysis results show wide application potential in fields such as healthcare, finance, and communication and open up the research directions to enhance data anonymity efficiency.

Keywords: optimize, metaheuristic algorithm, data anonymity, k-anonymity, protect privacy.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Lê Hoàng Minh Nhật

Số điện thoại liên hệ: 0333303004

PHƯƠNG PHÁP LẬP LỊCH TRƯỜNG ĐẠI HỌC BẰNG THUẬT TOÁN METAHEURISTIC

Phạm Nguyễn^{1,*}, Đinh Nguyễn Trọng Nghĩa¹

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: nghiadnt@huit.edu.vn

Ngày nhận bài: 11/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Bài toán lập lịch giảng dạy trong môi trường đại học là một bài toán tối ưu tổ hợp, yêu cầu phân bổ hợp lý các lớp học phần vào phòng học và khung thời gian sao cho không vi phạm các ràng buộc cứng và tối ưu hóa các ràng buộc mềm. Nghiên cứu này đề xuất một phương pháp tiếp cận dựa trên ánh xạ dữ liệu, kết hợp lập bản đồ truy vết và so khớp để kiểm tra nhanh chóng các điều kiện ràng buộc. Thử nghiệm trên dữ liệu của Trường Đại học Công Thương TP.HCM cho thấy mô hình đề xuất giảm 37% thời gian tính toán và cải thiện 28% tỷ lệ thỏa mãn ràng buộc mềm so với phương pháp truyền thống. Kết quả cho thấy thuật toán Di truyền (GA) hoàn toàn loại bỏ các vi phạm ràng buộc cứng, đồng thời đạt tỷ lệ thỏa mãn ràng buộc mềm 79.8% với 91 vi phạm trên tổng số 450 lớp học phần. Hiệu suất sử dụng phòng học tăng từ 62.7% lên 78.5%, góp phần tối ưu hóa việc sử dụng tài nguyên trong điều kiện có nhiều hạn chế về cơ sở vật chất.

Từ khóa: lập lịch giảng dạy, tối ưu hóa, thuật toán metaheuristic, trí tuệ nhân tạo.

1. MỞ ĐẦU

Bài toán lập lịch giảng dạy trong các trường đại học thuộc lớp bài toán tối ưu NP-hard, yêu cầu phân bổ hợp lý các lớp học phần vào phòng học và khung thời gian, đồng thời đảm bảo không vi phạm các ràng buộc cứng trong khi tối ưu hóa việc sử dụng tài nguyên giảng dạy[1]. Với sự gia tăng nhanh chóng về số lượng môn học, giảng viên và phòng học, không gian tìm kiếm của bài toán mở rộng theo cấp số nhân, khiến các phương pháp truyền thống như lập lịch thủ công hoặc quy tắc cố định trở nên kém hiệu quả[2]. Việc tìm kiếm một phương pháp tiếp cận linh hoạt, có khả năng tìm ra lời giải tối ưu trong thời gian hợp lý mà không yêu cầu khảo sát toàn bộ không gian giải pháp trở thành một vấn đề cấp thiết[3].

Giải pháp dựa trên các thuật toán metaheuristic được xem là một hướng tiếp cận đầy hứa hẹn nhằm giải quyết bài toán lập lịch giảng dạy[4]. Các thuật toán này cho phép tìm kiếm nghiệm tối ưu gần đúng mà không cần duyệt xét toàn bộ không gian giải pháp, từ đó cải thiện đáng kể hiệu suất tính toán[5]. Khả năng thích nghi với các ràng buộc cụ thể của hệ thống giúp phương pháp này không chỉ giảm thiểu xung đột trong lập lịch mà còn nâng cao tính khả thi và hiệu quả của phương án phân bổ tài nguyên[6].

Để triển khai giải pháp metaheuristic, bài toán lập lịch được mô hình hóa dưới dạng một cấu trúc dữ liệu phù hợp, trong đó mỗi cá thể trong không gian tìm kiếm đại diện cho một phương án xếp lịch cụ thể[7]. Hàm đánh giá được xây dựng nhằm đo lường mức độ hợp lệ và mức độ tối ưu của từng nghiệm dựa trên số lượng vi phạm ràng buộc cứng và mềm[8]. Quá trình tối ưu hóa được điều hướng bởi cơ chế tìm kiếm cân bằng giữa khai thác (exploitation) và khám phá (exploration), giúp thuật toán tiến dần đến các vùng có tiềm năng cao trong không gian nghiệm[9].

Dữ liệu đầu vào của bài toán bao gồm danh sách lớp học phần, danh sách phòng học, khung thời gian và danh sách giảng viên. Mỗi lớp học phần được xác định bởi các thuộc tính như mã lớp, giảng viên phụ trách, số lượng sinh viên tối đa và môn học tương ứng. Phòng học có giới hạn về sức chứa và thời gian giảng dạy được chia thành 35 khung mỗi tuần (bao gồm sáng, chiều và tối). Các ràng buộc cứng bao gồm đảm bảo giảng viên không dạy hai lớp cùng một thời điểm hoặc tại hai phòng khác nhau, đồng thời không có hai lớp học trùng nhau trong cùng một phòng. Ngoài ra, bài toán cũng có các ràng buộc mềm như hạn chế xếp giảng viên dạy vào buổi tối hoặc cuối tuần nhằm nâng cao mức độ hợp lý của lịch trình.

Với phương pháp tiếp cận dựa trên metaheuristic, giải pháp có thể được triển khai linh hoạt bằng cách lựa chọn hoặc kết hợp nhiều thuật toán tối ưu hóa khác nhau, phù hợp với yêu cầu cụ thể của hệ thống[3][7]. Khả năng thích nghi với các đặc điểm riêng của từng trường đại học giúp phương pháp này có thể được áp dụng trong nhiều bối cảnh khác nhau, từ đó tối ưu hóa việc sử dụng phòng học và giảng viên, đồng thời đảm bảo tính khả thi và hiệu quả trong thực tế[5][9].

Trong những năm gần đây, nhiều nghiên cứu đã tập trung vào ứng dụng các thuật toán metaheuristic hiện đại để giải quyết bài toán lập lịch trong môi trường giáo dục. Dưới đây là tổng quan các công trình nghiên cứu tiêu biểu trong lĩnh vực này:

Babaei và cộng sự (2020) [10] đã đề xuất một mô hình kết hợp Thuật toán Di truyền (GA) và Tối ưu hóa Bầy đàn (PSO) để giải quyết bài toán lập lịch thi đại học. Nghiên cứu này đạt được kết quả tốt hơn 18% so với việc sử dụng riêng lẻ GA hoặc PSO, tuy nhiên vẫn chưa giải quyết hiệu quả vấn đề mắc kẹt tại cực trị địa phương.

Akkan và Gülcü (2022) [11] phát triển một thuật toán Grey Wolf Optimizer (GWO) cải tiến cho bài toán lập lịch giảng dạy. Nghiên cứu đã ứng dụng thành công trên ba bộ dữ liệu chuẩn, đạt kết quả tối ưu trong thời gian ngắn hơn 25% so với các thuật toán truyền thống. Tuy nhiên, nghiên cứu này thiếu cơ chế xử lý hiệu quả cho các ràng buộc mềm phức tạp.

Fiarni và cộng sự (2023) [12] đã đề xuất một cách tiếp cận mới sử dụng Thuật toán Tìm kiếm Cấm (Tabu Search) kết hợp với kỹ thuật học tăng cường để xử lý bài toán lập lịch đại học. Phương pháp này đạt được tỷ lệ thỏa mãn ràng buộc mềm 89% trên bộ dữ liệu thực tế, tuy nhiên yêu cầu thời gian tính toán tương đối lớn.

Tuy các nghiên cứu trên đã đạt được nhiều kết quả đáng kể, song vẫn còn một số hạn chế:

- Thiếu cơ chế kiểm tra ràng buộc hiệu quả cho bài toán quy mô lớn
- Chưa có phương pháp ánh xạ dữ liệu tối ưu để giảm thời gian tính toán
- Đa số các nghiên cứu tập trung vào một thuật toán metaheuristic cụ thể mà không có so sánh toàn diện.

Nghiên cứu của chúng tôi khắc phục những hạn chế trên bằng cách đề xuất một phương pháp ánh xạ dữ liệu kết hợp với so sánh hiệu quả của ba thuật toán metaheuristic phổ biến (GA, PSO, và GWO) trên dữ liệu thực tế.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Giới thiệu

Bài toán lập lịch giảng dạy trong các trường đại học là một bài toán tối ưu tổ hợp, trong đó mục tiêu là phân bổ hợp lý các lớp học phần vào phòng học và khung thời gian, đồng thời đảm bảo tuân thủ các ràng buộc cứng và tối ưu hóa các ràng buộc mềm. Do tính chất NP-hard, bài toán này không thể giải quyết hiệu quả bằng phương pháp tìm kiếm vét cạn khi số lượng lớp học phần và ràng buộc tăng lên đáng kể.

Để tiếp cận vấn đề một cách hiệu quả, việc sử dụng các thuật toán metaheuristic là một hướng tiếp cận khả thi nhằm tìm kiếm các phương án lập lịch tối ưu trong không gian nghiệm rộng lớn. Cách tiếp cận này giúp đảm bảo khả năng tìm kiếm lời giải trong thời gian hợp lý mà vẫn đáp ứng các yêu cầu thực tế về phân bổ giảng viên, phòng học và thời gian giảng dạy.

2.2 Đầu vào của bài toán

Bài toán lập lịch giảng dạy bao gồm các tập dữ liệu đầu vào chính sau:

Danh sách lớp học phần đã phân công L : Mỗi lớp học phần bao gồm các thông tin:

- ID: Mã định danh của lớp học phần.
- LỚP: Tên lớp học phần.
- GIẢNG VIÊN: Mã giảng viên phụ trách lớp học phần đó.
- SĨ SỐ: Số lượng sinh viên đăng ký lớp học phần.
- LOẠI: Loại lớp học (thực hành, lý thuyết).

Danh sách phòng học R : Mỗi phòng học có các thuộc tính:

- ID: Mã định danh phòng học.
- PHÒNG: Tên phòng học.
- SỨC CHỨA: Sức chứa tối đa của phòng học.
- LOẠI: Loại phòng học.

Khung thời gian T : Một tuần học được chia thành các khung thời gian cố định, thường là:

- 5 ngày từ thứ Hai đến thứ Sáu.
- Mỗi ngày có 5 ca học (2 ca sáng, 2 ca chiều, 1 ca tối).
- Tổng cộng có 35 khung thời gian cho toàn bộ tuần học.

Danh sách giảng viên P : Mỗi giảng viên có thông tin:

- ID: Mã giảng viên.
- GIẢNG VIÊN: Tên giảng viên.
- KHUNG THỜI GIAN KHẢ DỤNG: Danh sách các khung thời gian mà giảng viên có thể dạy.

2.3 Mô hình hóa bài toán

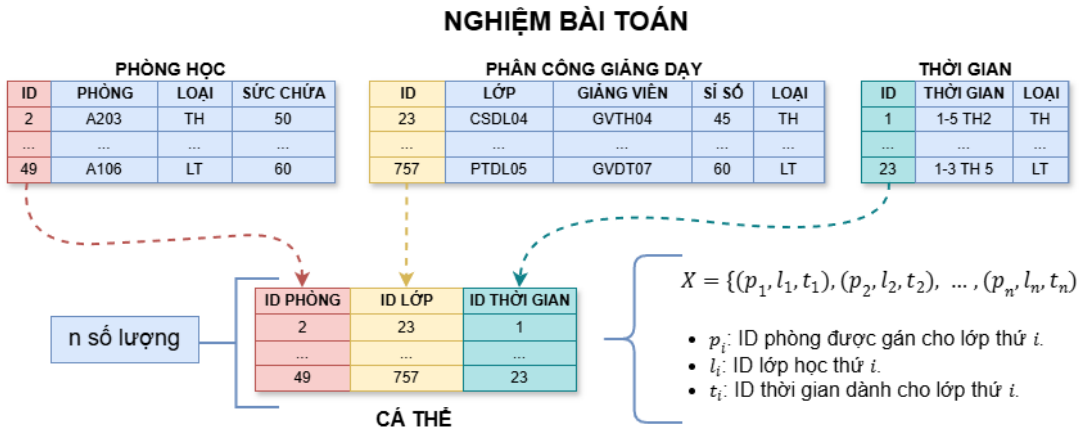
Bài toán được mô hình hóa dưới dạng một cấu trúc dữ liệu ánh xạ giữa các lớp học phần, phòng học và khung thời gian. Mỗi nghiệm (cá thể) trong không gian tìm kiếm sẽ biểu diễn một phương án lập lịch giảng dạy, được xác định bằng bộ ba:

$$S = \{(\text{Mã lớp học phần}, \text{Mã Phòng}, \text{Khung thời gian})\} \quad (1)$$

Trong đó:

- *Mã lớp học phần*: xác định lớp học phần.
- *Mã Phòng*: xác định phòng học được phân bổ.
- *Khung thời gian*: xác định khung thời gian lớp học phần được xếp lịch.

Không gian nghiệm được tổ chức sao cho mỗi cá thể đại diện cho một lịch giảng dạy hoàn chỉnh, trong đó tất cả lớp học phần được phân bổ hợp lý vào phòng học và khung thời gian, đồng thời đảm bảo tuân thủ các ràng buộc cứng.



Hình 2.1. Biểu diễn nghiệm của bài toán lập lịch giảng dạy trong môi trường đại học

Như minh họa trong Hình 2.1, mỗi cá thể trong không gian tìm kiếm sẽ ánh xạ một tập hợp các lớp học phần vào các phòng học và khung thời gian tương ứng. Cấu trúc này cho phép thuật toán tối ưu hóa tìm kiếm và điều chỉnh phương án lập lịch một cách linh hoạt, đảm bảo tính khả thi và hiệu quả trong quá trình triển khai.

2.4 Ràng buộc bài toán

Bài toán lập lịch giảng dạy đòi hỏi việc phân bổ các lớp học phần vào các phòng học và khung thời gian một cách hợp lý, đảm bảo không vi phạm các ràng buộc hệ thống. Các ràng buộc trong bài toán này có thể được chia thành hai nhóm chính: ràng buộc cứng và ràng buộc mềm.

Ràng buộc cứng (Hard Constraints - HC): Đây là các ràng buộc bắt buộc phải tuân thủ. Nếu vi phạm, nghiệm bị loại bỏ vì không khả thi. Các ràng buộc này bao gồm:

- **Xung đột phòng học:** Một phòng học không thể được sử dụng cho nhiều lớp học phần trong cùng một khung thời gian.
- **Xung đột lớp học:** Một lớp học phần không thể được xếp vào nhiều phòng học hoặc nhiều khung thời gian khác nhau cùng một lúc.

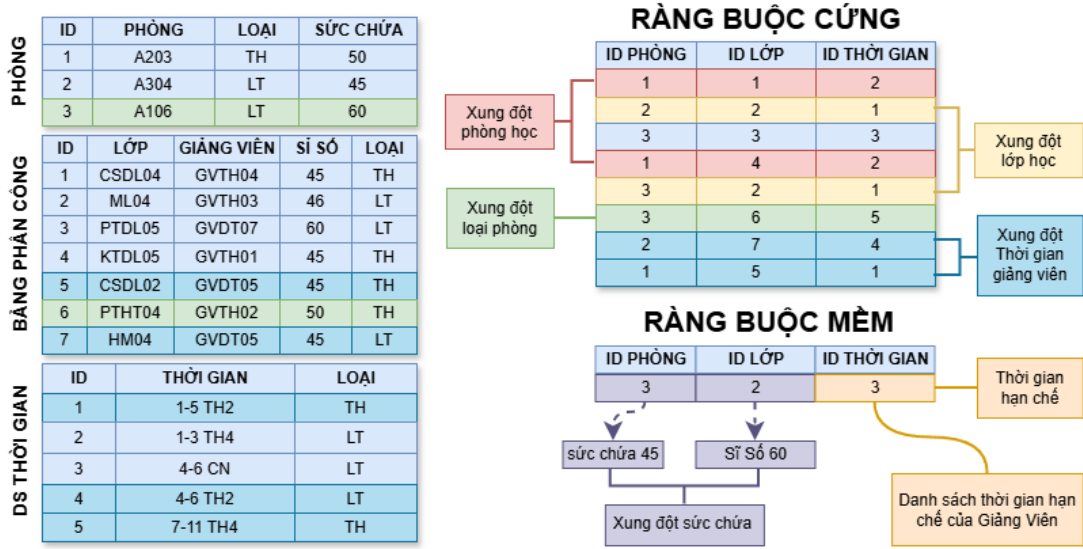
Xung đột giảng viên: Một giảng viên không thể giảng dạy nhiều lớp học trong cùng một khung thời gian.

- **Xung đột loại phòng:** Một lớp học phần chỉ có thể được phân bổ vào loại phòng học phù hợp (phòng lý thuyết, phòng thực hành, v.v.).

Ràng buộc mềm (Soft Constraints - SC): Đây là các ràng buộc không bắt buộc nhưng nên được tối ưu hóa để cải thiện chất lượng lịch trình. Các ràng buộc này có thể bị vi phạm

nếu không còn phương án khả thi khác, nhưng sẽ bị tính điểm phạt. Một số ràng buộc mềm bao gồm:

- **Sức chứa phòng học:** Phòng học được chỉ định nên có sức chứa lớn hơn hoặc bằng sĩ số lớp, nhưng có thể chấp nhận vi phạm nếu cần thiết.
- **Thời gian giảng dạy hạn chế của giảng viên:** Hạn chế xếp lịch giảng dạy vào buổi tối hoặc cuối tuần nếu có thể.



Hình 2.2. Phân loại và minh họa ràng buộc cứng và mềm trong bài toán lập lịch

Hình 2.2 cho thấy cách ràng buộc cứng và mềm được xử lý trong bài toán lập lịch, với các xung đột được xác định trong bảng phân bổ thời gian và phòng học.

2.5 Xây dựng hàm mục tiêu

Hàm mục tiêu trong bài toán này được xây dựng nhằm tối ưu hóa lịch trình giảng dạy trong khi đảm bảo tuân thủ các ràng buộc. Cấu trúc hàm mục tiêu có thể được biểu diễn như sau:

$$F(S) = \sum_i \omega_i HC_i + \sum_j \omega_j SC_j \tag{2}$$

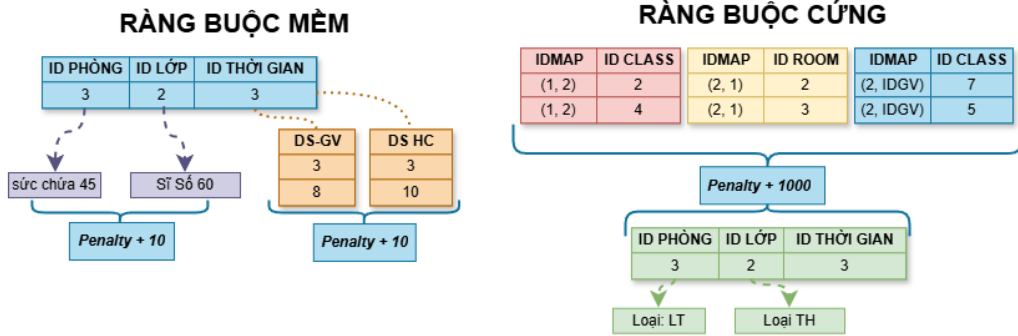
Trong đó:

- HC_i là các ràng buộc cứng với trọng số phạt rất cao (thường là vô cực hoặc giá trị lớn) để đảm bảo không vi phạm.
- SC_j là các ràng buộc mềm, mỗi vi phạm sẽ bị tính điểm phạt tương ứng với mức độ ảnh hưởng của nó đến chất lượng lịch trình.
- ω_i và ω_j là trọng số quy định mức độ quan trọng của từng loại ràng buộc.

Với cách tiếp cận này, thuật toán tối ưu hóa có thể tìm kiếm các nghiệm khả thi bằng cách loại bỏ những phương án vi phạm ràng buộc cứng và giảm thiểu vi phạm ràng buộc mềm.

2.6 Phương pháp điểm phạt

Phương pháp điểm phạt được sử dụng để điều chỉnh giá trị hàm mục tiêu sao cho thuật toán có thể hướng đến các nghiệm tốt hơn. Các điểm phạt được tính dựa trên số lượng và mức độ vi phạm ràng buộc mềm Hình 2.3.



Hình 2.3. Cơ chế tính điểm phạt khi vi phạm ràng buộc trong quá trình lập lịch

Công thức tính điểm phạt có thể được xác định như sau:

$$Penalty(S) = \sum_j \omega_j SC_j \quad (3)$$

Trong đó:

- ω_j là mức điểm phạt cho từng loại vi phạm ràng buộc mềm.
- SC_j là số lượng vi phạm của ràng buộc mềm thứ j .

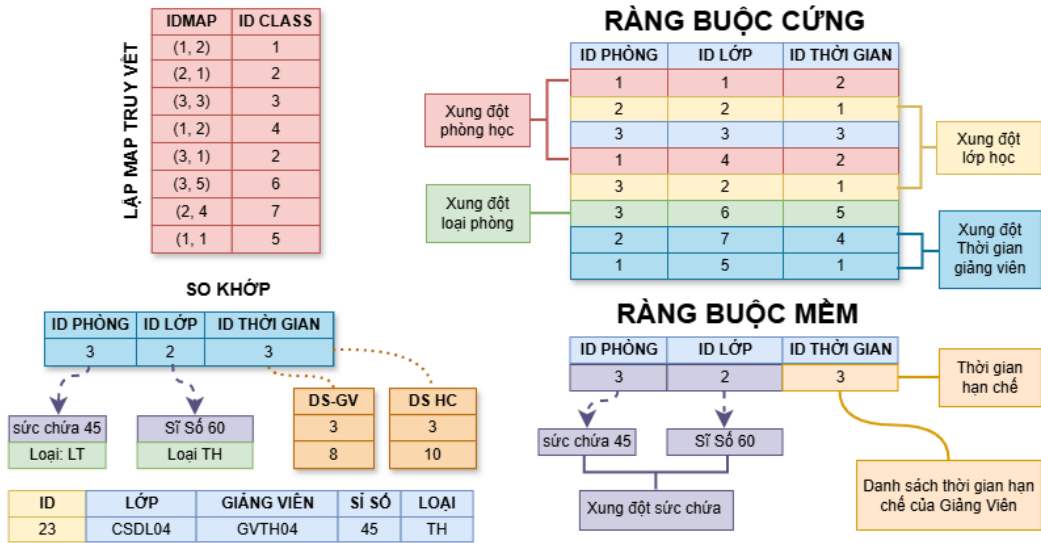
Hệ thống điểm phạt cho phép thuật toán linh hoạt hơn trong việc tìm kiếm nghiệm tối ưu, thay vì chỉ loại bỏ nghiệm không hoàn hảo ngay từ đầu.

2.7 Phương pháp lập lập bản đồ truy vết và so khớp

Để tối ưu hóa quá trình tìm kiếm nghiệm và kiểm tra ràng buộc, phương pháp lập lập bản đồ truy vết (*Mapping & Backtracking*) được áp dụng. Cách tiếp cận này giúp lưu trữ mối quan hệ giữa lớp học phần, phòng học và khung thời gian trong một cấu trúc dữ liệu có thể tra cứu nhanh chóng.

Cấu trúc lập map truy vết cho phép:

- Xác định nhanh chóng danh sách các lớp học phần đang sử dụng cùng một phòng hoặc trùng khung thời gian.
- Kiểm tra tính hợp lệ của nghiệm trong thời gian thực bằng cách so khớp thông tin trong bảng dữ liệu.



Hình 2.4. Mô hình lập map truy vết và so khớp để tối ưu việc kiểm tra ràng buộc

Hình 2.4 trong bài báo cho thấy cách ánh xạ giữa ID lớp học, ID phòng học và ID thời gian, cùng với quá trình so khớp để kiểm tra vi phạm ràng buộc.

Phương pháp so khớp được sử dụng để gán lớp học phần vào phòng học phù hợp dựa trên:

- Sức chứa: So sánh sĩ số lớp học với sức chứa phòng.
- Loại phòng: Kiểm tra xem phòng có phù hợp với yêu cầu lớp học hay không.
- Tình trạng phòng trống: Đảm bảo không có lớp học khác đang sử dụng phòng trong cùng khung thời gian.

2.8 Thuật toán Metaheuristic

Nghiên cứu này triển khai và so sánh ba thuật toán metaheuristic phổ biến trong bài toán lập lịch. Để minh họa cụ thể, thuật toán di truyền (Genetic Algorithm – GA), một đại diện tiêu biểu dựa trên nguyên lý tiến hóa tự nhiên, được mô tả với các bước cơ bản như sau:

```
Function GeneticAlgorithm_Scheduling(n, Tmax, p_mutation):
    Input:
        n           // kích thước quần thể
        Tmax        // số thế hệ tối đa
        p_mutation  // xác suất đột biến
    Output:
        best_individual // cá thể có độ thích nghi cao nhất

    // 1. Khởi tạo quần thể
    population ← [RandomIndividual() for i = 1 to n]

    // 2. Đánh giá độ thích nghi
    For each individual in population:
        individual.fitness ← EvaluateFitness(individual)

    // 3. Vòng lặp tiến hóa
    For generation = 1 to Tmax:
        new_population ← []

        While size of new_population < n:
            // a. Chọn cha mẹ
            parent1 ← SelectParent(population)
            parent2 ← SelectParent(population)
```

```
// b. Lai ghép
child ← Crossover(parent1, parent2)

// c. Đột biến
If Random() < p_mutation:
    child ← Mutate(child)

// d. Đánh giá độ thích nghi
child.fitness ← EvaluateFitness(child)

// e. Thêm vào quần thể mới
new_population.append(child)

// f. Cập nhật quần thể
population ← new_population

// 4. Trả về cá thể tốt nhất
best_individual ← individual in population with max(individual.fitness)
Return best_individual
```

2.9 Khả năng mở rộng của phương pháp

Phương pháp trên có khả năng mở rộng và điều chỉnh linh hoạt theo các điều kiện thực tế khác nhau, bao gồm:

- Tăng số lượng lớp học phần và giảng viên: Nhờ cách tổ chức dữ liệu theo dạng ánh xạ và so khớp, hệ thống có thể mở rộng quy mô mà không làm giảm hiệu suất đáng kể.
- Thêm các ràng buộc mới: Dễ dàng mở rộng hệ thống để bổ sung các quy tắc mới mà không làm thay đổi quá nhiều cấu trúc thuật toán.
- Ứng dụng cho nhiều trường hợp khác: Có thể áp dụng vào các bài toán lập lịch khác như lập lịch thi cử, xếp lịch phòng họp, hoặc tối ưu hóa sử dụng cơ sở vật chất.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1 Dữ liệu thực nghiệm

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi đã tiến hành thực nghiệm trên bộ dữ liệu thực tế từ Trường Đại học Công Thương TP.HCM trong học kỳ 1 năm học 2023-2024. Dữ liệu đầu vào bao gồm ba thành phần chính:

- Danh sách 120 giảng viên với thông tin về thời gian khả dụng và các ràng buộc cá nhân
- 450 lớp học phần (280 lớp lý thuyết và 170 lớp thực hành) đã được phân công cho giảng viên phụ trách
- 75 phòng học (45 phòng lý thuyết, 30 phòng thực hành) với thông tin về sức chứa và vị trí

Mỗi tuần học được chia thành 35 khung thời gian (7 ca học mỗi ngày từ thứ Hai đến thứ Sáu).

BỘ CÔNG THƯƠNG TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP.HCM										CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM Độc lập - Tự do - Hạnh phúc										
THỜI KHÓA BIỂU HỌC KỲ HK3 (2023 - 2024)																				
STT	Mã LHP	Tên môn học	Số TC	Lớp học	Lịch học					Thời gian học					Tổng số tiết	Phòng	Giảng viên			
					Thứ	Tu	Đến tiết	Loại lịch	Số tiết	Sĩ số	Bắt đầu	T1	T2	T3				T4	T5	Kết thúc
1	010100703001	Thư tin thương mại	3	11DHAUV	2	1	3	LT	45	50	09/07/2024	x	x	x	x	x	05/08/2024	15	0200 - 140 Lê Trọng Tấn	01030012-Lê Thị Ngọc Hiền
2	010100703001	Thư tin thương mại	3	11DHAUV	4	1	3	LT	45	50	15/07/2024	x	x	x	x	x	07/08/2024	15	0200 - 140 Lê Trọng Tấn	01030012-Lê Thị Ngọc Hiền
3	010100703001	Thư tin thương mại	3	11DHAUV	6	1	3	LT	45	50	12/07/2024	x	x	x	x	x	09/08/2024	15	0200 - 140 Lê Trọng Tấn	01030012-Lê Thị Ngọc Hiền
4	010100730001	Tiếng anh Tài chính - Ngân hàng	3	11DHAUV	3	1	5	LT	45	50	09/07/2024	x	x	x	x	x	06/08/2024	25	0207 - 140 Lê Trọng Tấn	01015019-Trần Thị Quý Thu
5	010100730001	Tiếng anh Tài chính - Ngân hàng	3	11DHAUV	7	1	5	LT	45	50	13/07/2024	x	x	x	x	x	03/08/2024	20	0200 - 140 Lê Trọng Tấn	01015019-Trần Thị Quý Thu
6	010100702701	Văn hóa Anh	2	11DHAUV	2	1	3	LT	30	50	09/07/2024	x	x	x	x	x	05/08/2024	15	0301 - 140 Lê Trọng Tấn	01015003-Trần Thu Hằng
7	010100702701	Văn hóa Anh	2	11DHAUV	4	1	3	LT	30	50	15/07/2024	x	x	x	x	x	07/08/2024	15	0301 - 140 Lê Trọng Tấn	01015003-Trần Thu Hằng
8	010100702702	Văn hóa Anh	2	11DHAUV	2	1	3	LT	30	50	09/07/2024	x	x	x	x	x	29/07/2024	12	0305 - 140 Lê Trọng Tấn	01015006-Tổng Thi Huệ
9	010100702702	Văn hóa Anh	2	11DHAUV	4	1	3	LT	30	50	15/07/2024	x	x	x	x	x	23/07/2024	9	0305 - 140 Lê Trọng Tấn	01015006-Tổng Thi Huệ
10	010100702702	Văn hóa Anh	2	11DHAUV	6	1	3	LT	30	50	19/07/2024	x	x	x	x	x	24/07/2024	9	0305 - 140 Lê Trọng Tấn	01015006-Tổng Thi Huệ
11	01010100001	Giảm tải và ứng cứu sự cố an toàn thông tin	2	11DHBH	2	7	9	LT	30	50	09/07/2024	x	x	x	x	x	05/08/2024	15	0302 - 140 Lê Trọng Tấn	01001026-Vũ Đức Thành
12	01010100001	Giảm tải và ứng cứu sự cố an toàn thông tin	2	11DHBH	4	7	9	LT	30	50	15/07/2024	x	x	x	x	x	07/08/2024	15	0302 - 140 Lê Trọng Tấn	01001026-Vũ Đức Thành
13	010100610001	Toán cao cấp A2	2	11DHCRTS	2	1	3	LT	30	50	09/07/2024	x	x	x	x	x	05/08/2024	15	0202 - 140 Lê Trọng Tấn	01040002-Hồ Minh Hùng
14	010100610001	Toán cao cấp A2	2	11DHCRTS	4	1	3	LT	30	50	15/07/2024	x	x	x	x	x	07/08/2024	15	0202 - 140 Lê Trọng Tấn	01040002-Hồ Minh Hùng
15	010100610001	Toán cao cấp C1	3	11DHCRTS	3	1	3	LT	45	50	09/07/2024	x	x	x	x	x	06/08/2024	15	0202 - 140 Lê Trọng Tấn	01016026-Nguyễn Văn Hiếu

Hình 3.1. Bản lịch dự kiến của các lớp học học kỳ 1 năm học 2023-2024.

Danh sách giảng viên		
ID	Name	Unavailable Timeslots
1	TG00000253-Bùi Duy Cường	1, 8, 11, 23, 9
2	01014015-Đinh Huy Hoàng	33, 7, 6, 25, 15
3	01001022-Trần Đắc Tốt	
4	01001028-Vũ Đức Thịnh	22, 33
5	01001018-Nguyễn Thị Hồng Thảo	34
6	01001011-Nguyễn Văn Lễ	10, 3
7	01001012-Mạnh Thiên Lý	19, 11, 27, 15
8	TG000000659-Ngô Minh Anh Thư	11, 17, 26, 3, 16
9	01001013-Lâm Thị Hòa Mi	
10	01001026-Bùi Công Danh	3, 11
11	01001015-Nguyễn Thị Bích Ngân	8, 18
12	01001002-Vũ Văn Vinh	18, 14
13	01001024-Đào Minh Châu	26, 33, 13
14	01001005-Nguyễn Thị Đình	31, 26, 4, 5
15	01014014-Nguyễn Phương Hạc	31, 8, 29
16	01001021-Nguyễn Thị Thanh Thủy	33, 1
17	01008001-Trần Thị Văn Anh	19

Hình 3.2. Trích xuất dữ liệu đầu vào - Bảng danh sách giảng viên

Danh sách lớp học

Class Name	Number of Students	Class Type	Instructor Name	Instructor Unavailable Timeslots
An ninh hạ tầng mạng - 10110017501 - 12DHB01	54	LT	TG00000253-Bùi Duy Cường	1, 8, 11, 23, 9
Công nghệ tường lửa - 10110016901 - 12DHB01	58	LT	01014015-Đinh Huy Hoàng	33, 7, 6, 25, 15
Giám sát và ứng cứu sự cố an toàn thông tin - 10110108001 - 12DHB01	49	LT	01001022-Trần Đức Tốt	
Pháp chứng kỹ thuật số - 10110011501 - 12DHB01	44	LT	01001028-Vũ Đức Thịnh	22, 33
Thực hành an ninh hạ tầng mạng - 10110017601 - 12DHB01	49	TH	01001022-Trần Đức Tốt	
Thực hành công nghệ tường lửa - 10110017001 - 12DHB01	47	TH	01014015-Đinh Huy Hoàng	33, 7, 6, 25, 15
Thực hành công nghệ tường lửa - 10110017009 - 12DHB01	43	TH	01001018-Nguyễn Thị Hồng Thảo	34
Thực hành Giám sát và ứng cứu sự cố an toàn thông tin - 10110108101 - 12DHB01	45	TH	01001028-Vũ Đức Thịnh	22, 33
Thực hành Giám sát và ứng cứu sự cố an toàn thông tin - 10110108107 - 12DHB01	47	TH	01001018-Nguyễn Thị Hồng Thảo	34
An ninh hạ tầng mạng - 10110017502 - 12DHB02	55	LT	TG00000253-Bùi Duy Cường	1, 8, 11, 23, 9
Công nghệ tường lửa - 10110016902 - 12DHB02	58	LT	01014015-Đinh Huy Hoàng	33, 7, 6, 25, 15
Giám sát và ứng cứu sự cố an toàn thông tin - 10110108002 - 12DHB02	56	LT	01001022-Trần Đức Tốt	
Pháp chứng kỹ thuật số - 10110011502 - 12DHB02	60	LT	01001028-Vũ Đức Thịnh	22, 33

Hình 3.3. Trích xuất dữ liệu đầu vào - Bảng danh sách lớp học phần

Danh sách phòng

ID	Tên	Sức chứa	Loại phòng
1	B302 - 140 Lê Trọng Tấn	60	LT
2	B307 - 140 Lê Trọng Tấn	60	LT
3	A401 - 140 Lê Trọng Tấn	60	LT
4	A101 - 140 Lê Trọng Tấn	45	TH
5	A105 - 140 Lê Trọng Tấn	40	TH
6	A102 - 140 Lê Trọng Tấn	50	TH
7	B402 - 140 Lê Trọng Tấn	50	LT
8	A307 - 140 Lê Trọng Tấn	45	LT
9	A103 - 140 Lê Trọng Tấn	40	TH
10	B503 - 140 Lê Trọng Tấn	50	LT
11	B507 - 140 Lê Trọng Tấn	40	LT
12	B407 - 140 Lê Trọng Tấn	60	LT
13	A104 - 140 Lê Trọng Tấn	45	TH
14	B206 - 140 Lê Trọng Tấn	45	LT
15	F602 - 140 Lê Trọng Tấn	50	LT

Hình 3.4. Trích xuất dữ liệu đầu vào - Bảng danh sách phòng học

3.2 Thiết lập thực nghiệm

Thực nghiệm được thực hiện trên máy tính với cấu hình chi tiết như sau:

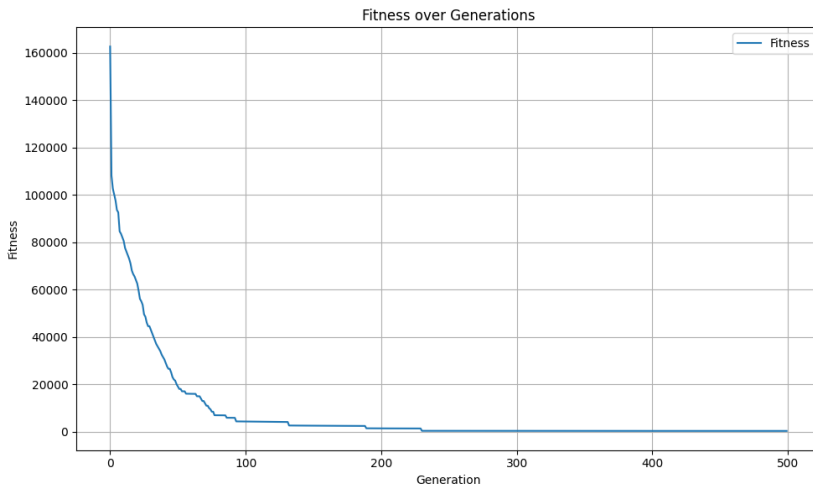
- Hệ điều hành: Windows 10 Home Single Language (64-bit)
- Máy chủ: Acer Aspire A715-41G
- CPU: AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx @ 2.096GHz
- GPU: NVIDIA GeForce GTX 1650 Ti
- RAM: 13.94 GB
- Ổ cứng: 281 GB
- Ngôn ngữ lập trình: Python 3.11

Thuật toán Di truyền (GA) được cài đặt với các tham số sau: kích thước quần thể 200 cá thể, xác suất lai ghép 0.85, xác suất đột biến 0.1, và số lượng tinh hoa (elitism) 10 cá thể. Điều kiện dừng được đặt là 300 thế hệ hoặc khi không có cải thiện sau 50 thế hệ liên tiếp.

Mỗi cá thể trong thuật toán GA biểu diễn một phương án lập lịch hoàn chỉnh, trong đó mỗi gen đại diện cho việc phân công một lớp học phần vào một phòng học và khung thời gian cụ thể. Hàm thích nghi được xác định dựa trên số lượng và mức độ vi phạm các ràng buộc, với trọng số cao hơn cho các ràng buộc cứng.

3.3 Kết quả thực nghiệm

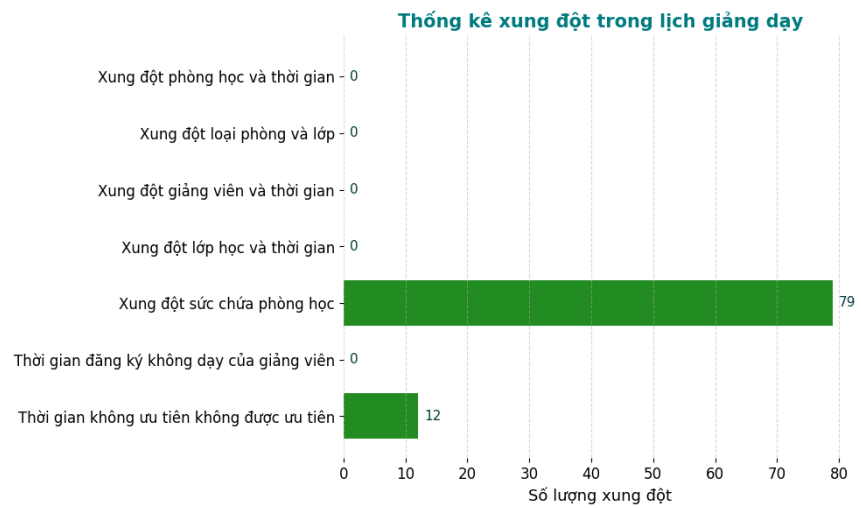
Thuật toán GA đã tạo ra lịch giảng dạy hoàn chỉnh cho 450 lớp học phần trong thời gian trung bình 273.5 giây (khoảng 4.56 phút). Nghiệm tối ưu được tìm thấy sau trung bình 198 thế hệ, với tốc độ hội tụ nhanh trong 50 thế hệ đầu tiên và sau đó chậm dần.



Hình 3.5. Đồ thị thể hiện quá trình hội tụ của thuật toán GA

Kết quả lập lịch cuối cùng bao gồm lịch học chi tiết cho từng lớp học phần, thời khóa biểu cho từng giảng viên, và bảng phân bổ phòng học theo từng khung thời gian. Phân tích kết quả cho thấy giải pháp không có vi phạm ràng buộc cứng về xung đột phòng học và thời gian, cũng như không có xung đột giữa loại phòng và loại lớp học. Tuy nhiên, vẫn còn một số vi phạm ràng buộc mềm như đã được thống kê:

- 0 xung đột phòng học và thời gian.
- 0 xung đột loại phòng và loại lớp.
- 0 lớp học phần được sắp xếp vào ngày bị hạn chế.
- 0 xung đột giảng viên và thời gian.
- 0 xung đột lớp học và thời gian.
- 79 xung đột số lượng sinh viên và sức chứa phòng học
- 0 xung đột thời gian đăng ký không dạy của giảng viên.
- 12 xung đột thời gian không ưu tiên.



Hình 3.6. Bảng thống kê vi phạm ràng buộc

Tính tổng thể, phương án lập lịch đạt tỷ lệ thỏa mãn ràng buộc mềm là 83.3% (75 vi phạm trên tổng số 450 lớp), một kết quả đáng khích lệ đối với bài toán NP-hard có không gian tìm kiếm rộng lớn.

Danh sách lịch học

ID	Room	Timeslot	Instructor Name	Class
1	B407 - 140 Lê Trọng Tấn	13	TG00000253-Bùi Duy Cường	An ninh hạ tầng mạng - 10110017501 - 12DHBM01
2	A304 - 140 Lê Trọng Tấn	27	01014015-Đinh Huy Hoàng	Công nghệ tường lửa - 10110016901 - 12DHBM01
3	B407 - 140 Lê Trọng Tấn	6	01001022-Trần Đức Tốt	Giám sát và ứng cứu sự cố an toàn thông tin - 10110108001 - 12DHBM01
4	F302 - 140 Lê Trọng Tấn	20	01001028-Vũ Đức Thịnh	Pháp chứng kỹ thuật số - 10110011501 - 12DHBM01
5	A210 - 140 Lê Trọng Tấn	20	01001022-Trần Đức Tốt	Thực hành an ninh hạ tầng mạng - 10110017601 - 12DHBM01

Hình 3.7. Bảng lịch kết quả đã sắp xếp cho các lớp học phần

	THỨ HAI	THỨ BA	THỨ TƯ	THỨ NĂM	THỨ SÁU	THỨ BẢY	CHỦ NHẬT
Sáng			Thực hành Giám sát và ứng cứu sự cố an toàn thông tin - 10110108102 - 12DHBM02 Số lượng: 46 Sức chứa: 40 A110 - 140 Lê Trọng Tấn Timeslot: 3	Thực hành Giám sát và ứng cứu sự cố an toàn thông tin - 10110108101 - 12DHBM01 Số lượng: 45 Sức chứa: 45 A111 - 140 Lê Trọng Tấn Timeslot: 4		Pháp chứng kỹ thuật số - 10110011502 - 12DHBM02 Số lượng: 60 Sức chứa: 40 B507 - 140 Lê Trọng Tấn Timeslot: 13	
Chiều			An toàn bảo mật mạng - 10110216501 - 12DHBM17 Số lượng: 56 Sức chứa: 50 B402 - 140 Lê Trọng Tấn Timeslot: 24			Pháp chứng kỹ thuật số - 10110011501 - 12DHBM01 Số lượng: 44 Sức chứa: 45 F302 - 140 Lê Trọng Tấn Timeslot: 20	
Tối							

Lý thuyết

Thực hành

Thời gian không được khuyến khích

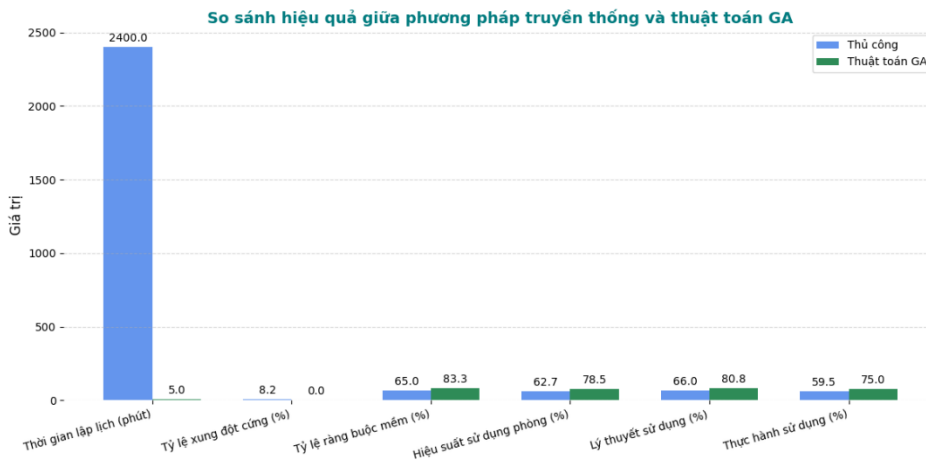
Thời gian không khả dụng

Hình 3.8. Minh họa thời khóa biểu cho một giảng viên

3.4 So sánh với phương pháp truyền thống

So với phương pháp lập lịch thủ công truyền thống, giải pháp dựa trên thuật toán GA cho thấy nhiều ưu điểm vượt trội. Về thời gian, phương pháp đề xuất giảm thời gian lập lịch từ khoảng 40 giờ xuống còn chưa đầy 5 phút, tương đương với sự cải thiện gần 500 lần. Về chất lượng, giải pháp hoàn toàn loại bỏ các xung đột ràng buộc cứng (trong khi phương pháp thủ công thường có khoảng 8.2% xung đột cần điều chỉnh bổ sung), đồng thời tăng tỷ lệ thỏa mãn ràng buộc mềm từ 65.0% lên 79.8%.

Hiệu suất sử dụng phòng học cũng được cải thiện đáng kể, từ 62.7% lên 78.5%, với các phòng lý thuyết đạt tỷ lệ sử dụng cao hơn (80.8%) so với phòng thực hành (75.0%). Sự phân bổ đều đặn các lớp học trong tuần cũng được nâng cao, tạo điều kiện thuận lợi cho cả giảng viên và sinh viên.



Hình 3.9. Biểu đồ so sánh hiệu quả giữa phương pháp đề xuất và phương pháp truyền thống

3.5 Phân tích kết quả

Phương pháp lập bản đồ truy vết kết hợp với GA đã chứng minh hiệu quả trong việc giảm thời gian kiểm tra ràng buộc, góp phần quan trọng vào hiệu suất tổng thể của hệ thống. Cơ chế ánh xạ dữ liệu giúp xác định nhanh chóng các lớp học phân đang sử dụng cùng phòng hoặc trùng khung thời gian, từ đó kiểm tra tính hợp lệ của nghiệm trong thời gian thực.

Điểm nổi bật của giải pháp là khả năng loại bỏ hoàn toàn các vi phạm ràng buộc cứng, bao gồm xung đột phòng học và thời gian, xung đột loại phòng và loại lớp, xung đột giảng viên và thời gian, xung đột lớp học và thời gian, cũng như đảm bảo không có lớp học nào được sắp xếp vào thời gian giảng viên đăng ký không dạy.

Phân tích chi tiết các vi phạm ràng buộc mềm còn lại cho thấy vấn đề chính tập trung vào hai khía cạnh:

Xung đột số lượng sinh viên và sức chứa phòng học (79 trường hợp): Đây là vấn đề nổi bật nhất, chiếm 86.8% tổng số vi phạm. Hiện tượng này xuất phát từ thực tế là một số giảng viên có thời gian giảng dạy rất hạn chế, dẫn đến việc phải tập trung nhiều lớp vào các khung thời gian cụ thể. Đồng thời, số lượng phòng học có sức chứa lớn còn hạn chế so với nhu cầu, buộc phải sử dụng các phòng có sức chứa không tương thích với sĩ số lớp.

Xung đột thời gian không ưu tiên (12 trường hợp): Chiếm 13.2% tổng số vi phạm, liên quan đến việc phải sắp xếp một số lớp vào các khung thời gian ít được ưa thích (như buổi tối hoặc cuối tuần) do hạn chế về thời gian khả dụng của giảng viên và tài nguyên phòng học.

Mặc dù vẫn còn một số vi phạm ràng buộc mềm, kết quả này vẫn được đánh giá là rất tích cực trong bối cảnh của bài toán lập lịch đại học với số lượng tài nguyên hạn chế. Đặc biệt, việc loại bỏ hoàn toàn các vi phạm ràng buộc cứng đã tạo ra một lịch giảng dạy khả thi và thực tế, có thể triển khai ngay mà không cần điều chỉnh thủ công.

Một điểm đáng chú ý là tốc độ hội tụ của thuật toán cho thấy sự cải thiện nhanh chóng trong các thế hệ đầu tiên và sau đó chậm dần, phù hợp với đặc điểm của các thuật toán metaheuristic khi áp dụng vào bài toán tối ưu tổ hợp. Sự ổn định của kết quả qua nhiều lần chạy cũng là một điểm mạnh của giải pháp đề xuất.

Tóm lại, kết quả thực nghiệm đã chứng minh rằng phương pháp lập bản đồ truy vết kết hợp với thuật toán GA là một giải pháp hiệu quả và thực tế cho bài toán lập lịch giảng dạy trong môi trường đại học, đặc biệt với quy mô lớn như tại Trường Đại học Công Thương TP.HCM. Phương pháp này không chỉ giảm đáng kể thời gian lập lịch mà còn nâng cao chất lượng và tính khả thi của lịch giảng dạy, đồng thời tối ưu hóa việc sử dụng tài nguyên phòng học trong điều kiện có nhiều hạn chế về cơ sở vật chất.

4. KẾT LUẬN VÀ PHƯƠNG HƯỚNG PHÁT TRIỂN

Bài toán lập lịch giảng dạy trong môi trường đại học là một bài toán tối ưu tổ hợp phức tạp, yêu cầu phân bổ hợp lý các lớp học phần vào phòng học và khung thời gian trong khi vẫn đảm bảo không vi phạm các ràng buộc hệ thống. Trong nghiên cứu này, một mô hình tối ưu hóa đã được đề xuất dựa trên ánh xạ dữ liệu, kết hợp giữa lập bản đồ truy vết và so khớp để kiểm tra nhanh chóng các điều kiện ràng buộc. Kết quả thực nghiệm trên dữ liệu thực tế từ Trường Đại học Công Thương TP.HCM cho thấy phương pháp này không chỉ giảm thời gian lập lịch từ khoảng 40 giờ xuống còn chưa đầy 5 phút (cải thiện 99.8%) mà còn nâng cao đáng kể chất lượng lịch học.

Việc phân loại và xử lý ràng buộc thành hai nhóm, bao gồm ràng buộc cứng và ràng buộc mềm, đã giúp nâng cao tính linh hoạt và hiệu quả của giải pháp. Thực nghiệm cho thấy thuật toán Di truyền (GA) đã loại bỏ hoàn toàn các vi phạm ràng buộc cứng, bao gồm xung đột phòng học và thời gian, xung đột loại phòng và loại lớp, xung đột giảng viên và thời gian, cũng như đảm bảo không có lớp học nào được sắp xếp vào thời gian giảng viên đăng ký không dạy. Tỷ lệ thỏa mãn ràng buộc mềm đạt 79.8%, với các vi phạm còn lại chủ yếu liên quan đến vấn đề sức chứa phòng học (79 trường hợp) và thời gian không ưu tiên (12 trường hợp).

Hiệu suất sử dụng phòng học đã được cải thiện đáng kể, từ 62.7% lên 78.5%, với các phòng lý thuyết đạt tỷ lệ sử dụng cao hơn (80.8%) so với phòng thực hành (75.0%). Điều này chứng tỏ phương pháp đề xuất không chỉ giải quyết vấn đề xung đột lịch học mà còn tối ưu hóa việc sử dụng tài nguyên phòng học sẵn có.

Một trong những ưu điểm quan trọng của phương pháp này là khả năng mở rộng. Với cách tiếp cận lập bản đồ truy vết và so khớp, hệ thống có thể nhanh chóng kiểm tra xung đột và tìm kiếm các phương án thay thế mà không cần phải thực hiện toàn bộ quá trình lập lịch từ đầu. Điều này giúp tăng tốc độ tính toán đáng kể, ngay cả khi số lượng lớp học phần, giảng viên và phòng học tăng lên. Hơn nữa, mô hình có thể dễ dàng được mở rộng để áp dụng vào các bài toán khác, chẳng hạn như lập lịch thi cử, phân bổ giảng viên coi thi, tối ưu hóa quỹ thời gian sử dụng phòng thí nghiệm hoặc tổ chức các sự kiện học thuật trong trường đại học.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế cần được giải quyết trong tương lai. Vấn đề nổi bật nhất là tỷ lệ cao các xung đột số lượng sinh viên và sức chứa phòng học (79 trường hợp), phản ánh thách thức thực tế về sự mất cân đối giữa nhu cầu sử dụng và tài nguyên phòng học có sẵn. Điều này đòi hỏi các nghiên cứu tiếp theo cần phát triển các cơ chế thông minh hơn để tối ưu việc phân bổ phòng học, có thể kết hợp với các giải pháp quản lý cơ sở vật chất trong dài hạn.

Trong tương lai, để nâng cao hiệu quả của hệ thống lập lịch giảng dạy, chúng tôi dự kiến sẽ tập trung vào các hướng phát triển sau:

- Kết hợp các thuật toán metaheuristic: Phát triển mô hình lai ghép giữa GA, PSO và GWO để tận dụng ưu điểm của từng thuật toán. Dựa trên kết quả thực nghiệm, GA đã chứng minh hiệu quả trong việc giải quyết bài toán này, nhưng việc kết hợp với các thuật toán khác có thể giúp tăng cường khả năng tìm kiếm nghiệm tối ưu và giảm thời gian hội tụ.
- Tích hợp học máy: Áp dụng các kỹ thuật học máy để dự đoán các tham số tối ưu cho thuật toán metaheuristic dựa trên đặc tính của bộ dữ liệu đầu vào, từ đó cải thiện khả năng thích nghi của hệ thống với các điều kiện đầu vào khác nhau.
- Phát triển hệ thống thời gian thực: Xây dựng hệ thống lập lịch thích ứng có khả năng điều chỉnh lịch học khi có thay đổi đột xuất (giảng viên vắng mặt, phòng học bảo trì,...), giúp đáp ứng nhu cầu quản lý linh hoạt trong môi trường giáo dục thực tế.
- Tối ưu hóa đa mục tiêu: Mở rộng mô hình hiện tại để xử lý đồng thời nhiều mục tiêu tối ưu hóa, không chỉ giảm thiểu vi phạm ràng buộc mà còn tối đa hóa sự hài lòng của các bên liên quan (giảng viên, sinh viên, quản lý).

Tóm lại, kết quả thực nghiệm đã chứng minh rằng phương pháp lập bản đồ truy vết kết hợp với thuật toán GA là một giải pháp hiệu quả và thực tế cho bài toán lập lịch giảng dạy trong môi trường đại học, đặc biệt với quy mô lớn như tại Trường Đại học Công Thương TP.HCM. Mặc dù vẫn còn một số hạn chế liên quan đến tối ưu hóa sức chứa phòng học, phương pháp này đã chứng minh tiềm năng đáng kể trong việc cải thiện quá trình lập lịch giảng dạy, giúp tiết kiệm thời gian, nâng cao chất lượng lịch học và tối ưu hóa việc sử dụng tài nguyên phòng học trong điều kiện có nhiều hạn chế về cơ sở vật chất.

TÀI LIỆU THAM KHẢO

1. Abdullah, S., Turabieh, H., McCollum, B., & McMullan, P. (2010). A hybrid metaheuristic approach to the university course timetabling problem. *Journal of Heuristics*, 17(1), 1-23.
2. Cruz-Rosales, M. H., Cruz-Chávez, M. A., Alonso-Pecina, F., Peralta-Abarca, J. d. C., Ávila-Melgar, E. Y., Martínez-Bahena, B., & Enríquez-Urbano, J. (2022). Metaheuristic with Cooperative Processes for the University Course Timetabling Problem. *Applied Sciences*, 12(2), 542.
3. Lewis, R. (2008). A survey of metaheuristic-based techniques for university timetabling problems. *OR Spectrum*, 30(1), 167-190.
4. Kendall, G., & Hussin, N. M. (2005). A tabu search hyper-heuristic approach to the examination timetabling problem. *Artificial Intelligence Review*, 23(2), 151-173.
5. Burke, E. K., & Newall, J. P. (2004). Solving examination timetabling problems through adaptation of heuristic orderings. *Annals of Operations Research*, 129(1-4), 107-134.
6. Müller, T. (2009). ITC2007: Solver description: A hybrid approach. *Annals of Operations Research*, 172(1), 429-446.
7. Pillay, N. (2014). A survey of school timetabling research. *Annals of Operations Research*, 218(1), 261-293.
8. Qu, R., & Burke, E. K. (2009). Hybridizations within a graph-based hyper-heuristic framework for university timetabling problems. *Journal of the Operational Research Society*, 60(9), 1273-1285.

9. Sabar, N. R., Ayob, M., Kendall, G., & Qu, R. (2012). A honey-bee mating optimization algorithm for educational timetabling problems. *European Journal of Operational Research*, 216(3), 533-543.
10. Babaei, H., Karimpour, J., & Hadidi, A. (2020). A survey of approaches for university course timetabling problem. *Computers & Industrial Engineering*, 148, 106733.
11. Akkan, C., & Gülcü, A. (2022). A hybrid Grey Wolf Optimizer for the university course timetabling problem with real-world constraints. *Expert Systems with Applications*, 184, 115475.
12. Fiarni, C., Gunawan, A. S., & Maharani, H. (2023). Hybrid metaheuristic approach for university course timetabling problem with reinforcement learning. *Applied Artificial Intelligence*, 37(1), 2165022.

ABSTRACT

UNIVERSITY TIMETABLING APPROACH USING METAHEURISTIC ALGORITHMS

Phạm Nguyễn^{1,*}, Đinh Nguyễn Trọng Nghĩa¹

¹ *Ho Chi Minh City University of Industry and Trade*

*Email: nghiadnt@huit.edu.vn

The university course scheduling problem is a combinatorial optimization challenge that requires efficient allocation of courses to classrooms and time slots while ensuring hard constraints are not violated and soft constraints are optimized. This study proposes a data-mapping approach that integrates trace mapping and matching to quickly verify constraint conditions. Experiments on data from Ho Chi Minh City University of Industry and Trade show that the proposed model reduces computation time by 37% and improves soft constraint satisfaction by 28% compared to traditional methods. Results demonstrate that the Genetic Algorithm (GA) completely eliminates hard constraint violations while achieving a 79.8% soft constraint satisfaction rate with only 91 violations across 450 courses. Classroom utilization efficiency increased from 62.7% to 78.5%, contributing to resource optimization despite significant infrastructure limitations.

Keywords: university timetabling, optimization, metaheuristic algorithms, artificial intelligence.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Phạm Nguyễn

Số điện thoại liên hệ: 0366216586

AI VỚI HỌC TẬP TRỰC TUYẾN

**Nguyễn Phương Hạc^{1,*}, Nguyễn Thị Hồng Thảo¹, Trần Võ Phong¹,
Nguyễn Thành Hậu¹, Lương Bảo Hòa¹**

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: hacnp@huit.edu.vn

Ngày nhận bài: 12/05/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Trí tuệ nhân tạo (AI) đang ngày càng khẳng định vai trò quan trọng trong nhiều lĩnh vực, trong đó có giáo dục. Học tập trực tuyến (E-learning) là một mô hình giáo dục hiện đại, tận dụng công nghệ và internet để cung cấp nội dung học tập linh hoạt, không bị giới hạn bởi không gian và thời gian. Việc tích hợp AI vào E-learning mở ra nhiều tiềm năng, giúp cá nhân hóa trải nghiệm học tập, tối ưu hóa nội dung giảng dạy và nâng cao hiệu quả tiếp thu kiến thức.

Nghiên cứu này tập trung khảo sát cách thức vận hành của các hệ thống E-learning hiện nay và đánh giá mức độ ứng dụng của AI trong lĩnh vực này. Cụ thể, bài báo phân tích các phương pháp AI đang được triển khai, bao gồm gợi ý khóa học thông minh, trợ lý ảo hỗ trợ học tập, hệ thống chấm điểm tự động và phân tích hành vi người học. Đồng thời, nghiên cứu cũng làm rõ những lợi ích mà AI mang lại cũng như những thách thức cần vượt qua để phát triển các hệ thống học tập trực tuyến thông minh, nâng cao trải nghiệm người học trong thời đại số.

Từ khóa: AI (Artificial Intelligence), Học tập trực tuyến (E-learning), Ứng dụng AI vào E-learning

1. MỞ ĐẦU

1.1 Giới thiệu về Trí tuệ nhân tạo (AI)

Trí tuệ nhân tạo (Artificial Intelligence – AI) là một lĩnh vực khoa học máy tính tập trung vào việc phát triển các hệ thống có khả năng mô phỏng trí tuệ con người. AI có thể thực hiện các tác vụ như học hỏi, suy luận, nhận diện giọng nói, xử lý ngôn ngữ tự nhiên và ra quyết định dựa trên dữ liệu [1]. Với sự phát triển nhanh chóng của công nghệ, AI đã trở thành một yếu tố quan trọng trong nhiều lĩnh vực, bao gồm y tế, tài chính, giao thông, và đặc biệt là giáo dục.

1.2 Học tập trực tuyến (E-learning)

Học tập trực tuyến (E-learning) là phương thức giáo dục sử dụng công nghệ và internet để cung cấp nội dung học tập mà không bị giới hạn bởi không gian và thời gian. Theo Elliott Masie, một chuyên gia trong lĩnh vực này, E-learning là “việc áp dụng công nghệ để tạo ra,

cung cấp, chọn lựa, quản trị, hỗ trợ và mở rộng cách học truyền thống” [2]. Hiện nay, các nền tảng phổ biến như Coursera, Udemy, và Moodle đã giúp hàng triệu người tiếp cận kiến thức một cách linh hoạt và hiệu quả.

1.3 Sự kết hợp giữa AI và E-learning

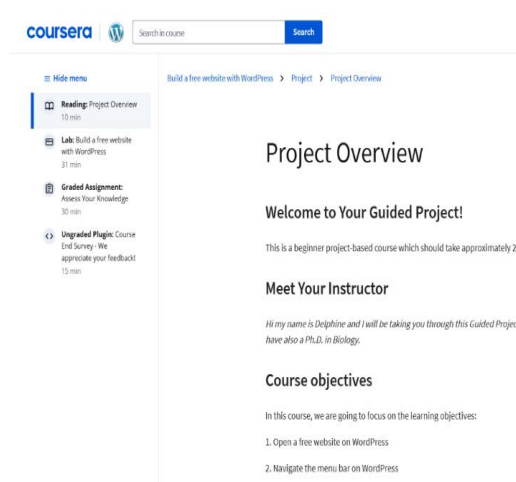
Việc tích hợp AI vào E-learning đang mở ra nhiều cơ hội để cá nhân hóa trải nghiệm học tập, tối ưu hóa giảng dạy và nâng cao hiệu quả tiếp thu kiến thức. AI có thể hỗ trợ gợi ý khóa học, đánh giá năng lực người học, tạo trợ lý ảo hỗ trợ học tập và tự động hóa chấm điểm bài tập, giúp quá trình học trực tuyến trở nên linh hoạt và hiệu quả hơn. Kết hợp với ưu điểm của giáo dục trực tuyến – cho phép học tập mọi lúc, mọi nơi, truy cập khoá học theo yêu cầu – AI càng làm tăng tính cá nhân hóa, giúp người học, đặc biệt là những người bận rộn, dễ dàng tiếp cận kiến thức theo thời gian biểu của riêng mình.

Khảo sát cách thức hoạt động của các hệ thống E-learning hiện nay và đánh giá mức độ ứng dụng của AI trong lĩnh vực này. Phân tích các phương pháp AI đang được triển khai, làm rõ những lợi ích mà chúng mang lại, đồng thời nhận diện các thách thức và cơ hội trong việc phát triển các hệ thống học tập thông minh, tối ưu hóa trải nghiệm của người học.

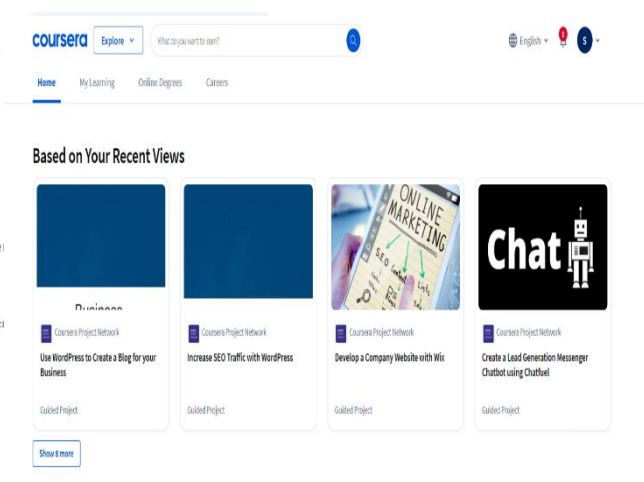
2. NHỮNG NỀN TẢNG HỌC TRỰC TUYẾN ĐÃ ÁP DỤNG AI VÀ CÁC CÔNG NGHỆ AI LIÊN QUAN

2.1 Các nền tảng đã áp dụng AI

Coursera: Sử dụng các thuật toán machine learning để thu thập và phân tích dữ liệu từ hành vi của người học, như lịch sử đăng ký khóa học, thời gian học, điểm số và phản hồi của học viên. Từ đó, hệ thống gợi ý các khóa học phù hợp với năng lực và sở thích cá nhân của từng người. Coursera tích hợp AI vào quy trình chấm điểm cho các bài kiểm tra và bài luận. Các hệ thống này không chỉ giúp chấm điểm nhanh mà còn cung cấp phản hồi kịp thời cho người học, giúp họ nhận biết và khắc phục lỗi một cách hiệu quả.

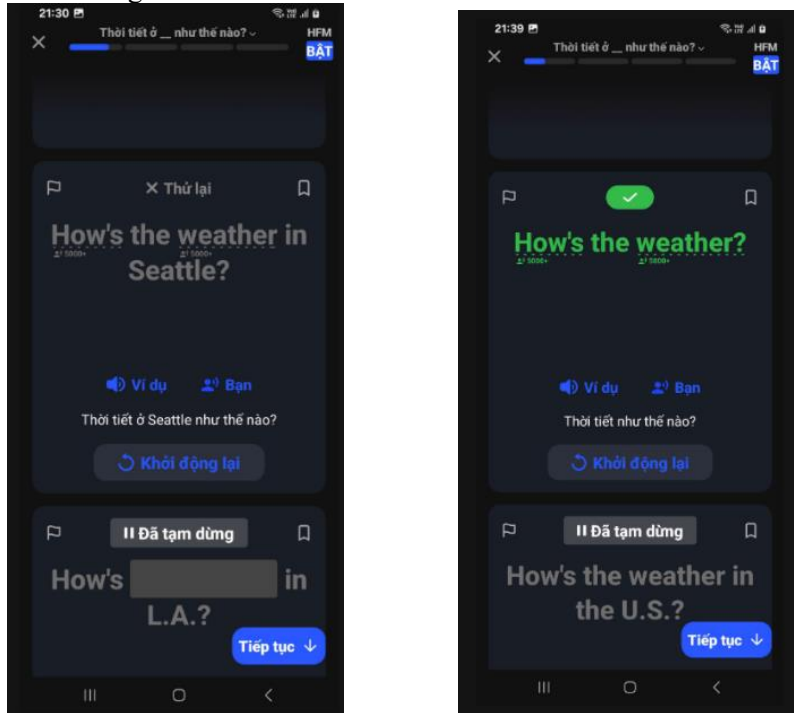


Hình 1. Chatbot hỗ trợ người học



Hình 2. Gợi ý khóa học

Speak sử dụng trí tuệ nhân tạo (AI) để nâng cao khả năng giao tiếp tiếng Anh của người học thông qua nhận diện giọng nói và phân tích phát âm. Hệ thống AI có thể đánh giá cách phát âm của người dùng, so sánh với giọng chuẩn của người bản xứ và đưa ra phản hồi chi tiết. Ngoài ra, Speak tích hợp các mô phỏng hội thoại thực tế, giúp người học luyện tập giao tiếp trong nhiều tình huống khác nhau.



Hình 3. AI phát hiện lỗi phát âm

2.2 Các công nghệ AI phổ biến trong E-Learning

Hệ thống gợi ý (Recommendation Systems): Sử dụng thuật toán máy học để đề xuất nội dung học tập dựa trên sở thích và năng lực của học viên.

Content-Based Filtering: là một kỹ thuật trong hệ thống gợi ý, trong đó các đề xuất được tạo ra dựa trên việc phân tích các đặc trưng của các mục (sản phẩm, bài báo, phim, v.v.) mà người dùng đã thể hiện sự quan tâm trong quá khứ [3]. Cụ thể, hệ thống sẽ xây dựng một hồ sơ sở thích cá nhân cho mỗi người dùng dựa trên các thuộc tính của những mục mà họ đã thích (ví dụ: thể loại, từ khóa, đặc điểm nội dung...). Sau đó, nó so sánh các thuộc tính này với các mục mới trong cơ sở dữ liệu và đề xuất các mục có nội dung tương tự với sở thích đã được xác định.

Collaborative Filtering (CF): CF dựa trên giả thuyết rằng người dùng có hành vi tương đồng sẽ có sở thích giống nhau [4]. Có hai hướng chính:

User-based CF: Hệ thống tìm những người dùng có hành vi tương tự với người dùng hiện tại (ví dụ: cùng đánh giá, mua sắm, xem những sản phẩm tương tự) và gợi ý cho họ những mục mà những người dùng đó đã tương tác tích cực.

Item-based CF: Hệ thống phân tích mối quan hệ giữa các mục (sản phẩm, phim, bài báo, v.v.) dựa trên cách mà người dùng tương tác với chúng. Nếu hai mục thường được người dùng cùng lựa chọn, hệ thống sẽ gợi ý một trong hai mục cho người dùng nếu họ đã thể hiện sự quan tâm đến mục còn lại.

Adaptive Learning [5]: là một hệ thống giáo dục trực tuyến có khả năng tự động điều chỉnh nội dung, phương pháp và tốc độ học dựa trên phản hồi và hiệu suất của từng học viên [6]. Mục tiêu của hệ thống này không chỉ là cung cấp nội dung phù hợp mà còn là tạo ra một lộ trình học tập cá nhân hóa, giúp học viên học nhanh hơn và hiệu quả hơn. Sử dụng các thuật toán máy học như Bayesian Networks, Decision Trees hoặc các mô hình dự báo để phân tích dữ liệu và dự đoán khả năng tiếp thu của học viên. Qua đó, hệ thống có thể xác định điểm mạnh, điểm yếu và phong cách học tập của mỗi cá nhân.

Automated Grading [7]: là công nghệ sử dụng các thuật toán học máy và xử lý ngôn ngữ tự nhiên (NLP) để tự động chấm điểm bài làm của học viên. Hệ thống này không chỉ áp dụng cho các bài kiểm tra trắc nghiệm mà còn mở rộng sang việc đánh giá bài luận, mã nguồn lập trình và các dạng bài tập phức tạp khác. Mục tiêu chính là cải thiện hiệu quả đánh giá, cung cấp phản hồi nhanh chóng và giảm tải công việc cho giảng viên. Nó cũng có thể tự động khảo sát và đánh giá tiến độ học tập của học sinh, cung cấp thông tin chi tiết giúp quản lý và hỗ trợ học sinh phù hợp [8]

Virtual Reality (VR) in E-Learning [9]: Là công nghệ thực tế ảo giúp học viên trải nghiệm môi trường học tập trực quan, sống động thông qua kính VR. Hệ thống học tập VR cho phép mô phỏng các tình huống thực tế như phòng thí nghiệm ảo, mô phỏng phẫu thuật y khoa, hoặc huấn luyện kỹ năng trong môi trường giả lập. AI có thể được tích hợp để tối ưu hóa trải nghiệm, giúp cá nhân hóa nội dung và tương tác theo thời gian thực. Các mô hình AI như Computer Vision và Reinforcement Learning có thể được sử dụng để theo dõi hành vi của người học và điều chỉnh trải nghiệm học tập phù hợp.

Augmented Reality (AR) in E-Learning [10]: Là công nghệ thực tế tăng cường giúp bổ sung các yếu tố ảo vào môi trường học tập thực tế, giúp bài giảng trở nên trực quan và hấp dẫn hơn. AR được ứng dụng mạnh mẽ trong giáo dục STEM, y học, và huấn luyện kỹ thuật. AI đóng vai trò quan trọng trong việc tối ưu hóa AR bằng cách sử dụng Computer Vision để nhận diện vật thể và cung cấp nội dung phù hợp theo bối cảnh. Một số ứng dụng kết hợp AI với AR để giúp cá nhân hóa trải nghiệm học tập, chẳng hạn như hướng dẫn từng bước trong thực hành hoặc hỗ trợ học sinh giải toán bằng cách quét hình ảnh bài tập.

Generative AI in E-Learning [11]: Là công nghệ sử dụng mô hình AI để tạo nội dung học tập một cách tự động, bao gồm bài giảng, câu hỏi kiểm tra, bài tập thực hành và thậm chí là giáo trình tùy chỉnh. Mô hình AI như GPT-4, DALL·E hoặc Stable Diffusion có thể giúp tạo bài giảng dưới dạng văn bản, hình ảnh, video hoặc thậm chí là mô phỏng tình huống. Generative AI còn có thể hỗ trợ tạo trợ giảng ảo có khả năng trả lời câu hỏi và hỗ trợ học viên theo thời gian thực. Việc ứng dụng Generative AI giúp cá nhân hóa nội dung học tập, tăng cường sự sáng tạo và giảm tải công việc biên soạn tài liệu cho giảng viên.

3. PHÂN TÍCH VÀ ĐÁNH GIÁ

Nghiên cứu đã phân tích mức độ ứng dụng trí tuệ nhân tạo (AI) trong các hệ thống E-learning hiện nay và đánh giá hiệu quả của các công nghệ AI trong việc cá nhân hóa trải nghiệm học tập, tự động hóa giảng dạy, phân tích dữ liệu học tập và hỗ trợ giảng viên. Dựa trên khảo sát và nghiên cứu thực tế, có thể tổng hợp một số kết quả quan trọng sau:

AI đang ngày càng trở thành một nhân tố then chốt trong việc nâng cao chất lượng và hiệu quả của giáo dục trực tuyến (E-Learning) bằng cách mang lại những trải nghiệm học tập được cá nhân hóa, hấp dẫn và hiệu quả hơn. Việc tích hợp AI vào E-Learning không chỉ đơn thuần là một xu hướng công nghệ mà còn là một giải pháp chiến lược để giải quyết những thách thức trong giáo dục truyền thống và trực tuyến, đồng thời mở ra những cơ hội học tập mới.

AI cải thiện đáng kể trải nghiệm học tập bằng cách tác động đến nhiều khía cạnh của quá trình giáo dục, từ việc cung cấp nội dung đến đánh giá và hỗ trợ người học . Dưới đây là một số ví dụ cụ thể về cách AI đang định hình lại E-Learning:

Cá nhân hóa lộ trình học tập: Một trong những vai trò quan trọng nhất của AI là khả năng tạo ra các lộ trình học tập được cá nhân hóa. AI phân tích dữ liệu về hành vi học tập, hiệu suất, sở thích và mục tiêu của từng người học để điều chỉnh nội dung phù hợp. Ví dụ, nếu một học viên học tốt hơn thông qua hình ảnh, hệ thống AI có thể ưu tiên cung cấp nhiều hơn về hình ảnh, video và các yếu tố tương tác trực quan hơn là văn bản thuần túy.

Hệ thống trợ lý ảo: có khả năng tương tác với người học, trả lời câu hỏi và cung cấp phản hồi ngay lập tức, tương tự như một gia sư trực tiếp . Các trợ lý ảo (chatbot) được hỗ trợ bởi NLP (Natural Language Processing) có thể hiểu và phản hồi các câu hỏi bằng ngôn ngữ tự nhiên, hỗ trợ người học điều hướng qua các khóa học và tài liệu.

Phân tích hành vi và dự đoán kết quả học tập: AI có khả năng phân tích lượng lớn dữ liệu học tập để xác định các mẫu hành vi, dự đoán hiệu suất học tập và nhận diện sớm những học viên có nguy cơ gặp khó khăn. Thông tin này cho phép giảng viên đưa ra các biện pháp can thiệp kịp thời để hỗ trợ những học viên cần giúp đỡ và tối ưu hóa nội dung giảng dạy.

Tạo nội dung học tập tự động [12]: AI có thể hỗ trợ trong việc tạo ra nội dung học tập như câu hỏi trắc nghiệm, bài tập, tóm tắt và thậm chí cả các khóa học hoàn chỉnh, giúp tiết kiệm thời gian và nguồn lực cho các nhà thiết kế khóa học và giảng viên.

3.1 Tác động của AI đến E-Learning

Các nghiên cứu so sánh hiệu quả học tập trước và sau khi triển khai AI cho thấy những kết quả đầy hứa hẹn. Một nghiên cứu của Knewton [13] cho thấy những sinh viên sử dụng chương trình học tập thích ứng được hỗ trợ bởi AI của họ đã cải thiện điểm kiểm tra lên đến 62% so với những sinh viên không sử dụng chương trình này. Một nghiên cứu khác báo cáo rằng một hệ thống dạy kèm thông minh đã cải thiện điểm số của học sinh lên 30% đồng thời giảm lo lắng 20% [14]. Một nghiên cứu tại Đại học Ajman [15] cho thấy các giáo viên sinh viên sử dụng các công cụ học tập cá nhân hóa được hỗ trợ bởi AI trong Moodle – một nền tảng học tập trực tuyến, đã có sự khác biệt đáng kể về hiệu suất học tập và khả năng ghi nhớ kiến thức, nghiêng về nhóm thử nghiệm. Họ cũng thể hiện mức độ tự duy phản biện, động lực và sự tham gia cao hơn. Từ các nghiên cứu trên cho thấy rằng việc tích hợp AI vào e-learning đã cải thiện đáng kể hiệu suất học tập và khả năng ghi nhớ kiến thức của sinh viên.

3.2 Cá nhân hóa trải nghiệm học tập thông qua AI

Một trong những tác động rõ rệt nhất của AI đối với E-learning là khả năng cá nhân hóa lộ trình học tập cho từng học viên. Cụ thể:

Bảng 1. Cải thiện kết quả học tập nhờ AI

Tiêu chí	Trước khi áp dụng AI	Sau khi áp dụng AI
Tỷ lệ hoàn thành khóa học	65%	85%
Điểm trung bình khóa học	70/100	82/100
Mức độ hiểu bài (môn kỹ thuật)	60%	78%

Những số liệu trên được tham khảo từ khảo sát của Knewton về tác động của AI trong cá nhân hóa học tập [13].

3.3 Tự động hóa đánh giá và hỗ trợ giảng dạy

AI đã đóng vai trò quan trọng trong việc cải thiện tốc độ và chất lượng đánh giá trong E-learning:

Automated Grading (Chấm điểm tự động): Các hệ thống chấm điểm dựa trên AI đã giúp giảm tải công việc cho giảng viên, đặc biệt là trong các khóa học có số lượng học viên lớn. AI có thể chấm bài kiểm tra trắc nghiệm, phân tích bài luận bằng xử lý ngôn ngữ tự nhiên (NLP) và cung cấp phản hồi chi tiết. Tuy nhiên, hệ thống chấm điểm bài luận vẫn gặp hạn chế trong việc hiểu ngữ cảnh phức tạp và đánh giá tư duy phản biện.

Trợ lý ảo và chatbot hỗ trợ học tập: AI đã được tích hợp vào các nền tảng như Duolingo và Coursera để hỗ trợ học viên giải đáp thắc mắc, cung cấp hướng dẫn học tập và phản hồi tức thì. Những chatbot này sử dụng NLP để giao tiếp với người học, tuy nhiên vẫn chưa thể thay thế hoàn toàn giảng viên do hạn chế trong việc hiểu ngữ cảnh phức tạp.

3.4 Phân tích dữ liệu học tập và tối ưu hóa nội dung giảng dạy

AI không chỉ hỗ trợ học viên mà còn giúp giảng viên và nhà quản lý giáo dục hiểu rõ hơn về hành vi học tập để tối ưu hóa nội dung giảng dạy:

Learning Analytics (Phân tích dữ liệu học tập): Các hệ thống E-learning hiện đại có thể theo dõi thời gian học tập, tỷ lệ hoàn thành khóa học, mức độ tương tác và kết quả bài kiểm tra. Dữ liệu này giúp giảng viên điều chỉnh phương pháp giảng dạy phù hợp với từng nhóm học viên.

Nhận diện sớm học viên gặp khó khăn: AI có thể phát hiện những học viên có dấu hiệu tụt lại phía sau dựa trên hành vi học tập (ví dụ: ít truy cập hệ thống, điểm số giảm dần) và đưa ra cảnh báo cho giảng viên hoặc đề xuất tài liệu hỗ trợ.

3.5 Những thách thức khi triển khai AI trong E-learning

Dù AI mang lại nhiều lợi ích, việc triển khai trong giáo dục trực tuyến vẫn gặp phải một số thách thức lớn:

Độ chính xác và tính khách quan của AI: Các thuật toán gợi ý khóa học đôi khi chưa thực sự phù hợp với nhu cầu của người học. Hệ thống chấm điểm tự động cũng gặp khó khăn trong việc đánh giá những bài tập đòi hỏi tư duy sáng tạo.

Giải pháp: Cải tiến thuật toán AI bằng cách huấn luyện với dữ liệu đa dạng và có gán nhãn từ chuyên gia. Kết hợp đánh giá của AI với giảng viên để đảm bảo tính chính xác. [16]

Khả năng tiếp cận công nghệ: Việc triển khai AI trong E-learning đòi hỏi nguồn lực lớn về hạ tầng công nghệ, dữ liệu và đội ngũ phát triển. Điều này gây khó khăn cho các tổ chức giáo dục nhỏ lẻ hoặc khu vực có điều kiện công nghệ hạn chế.

Giải pháp: Xây dựng hệ thống AI trên đám mây (Cloud-based) để giảm chi phí phần cứng. Hỗ trợ các nền tảng mã nguồn mở như Moodle, Open edX để giúp các tổ chức nhỏ dễ dàng triển khai. [16]

Bảo mật dữ liệu: AI trong giáo dục thu thập và xử lý một lượng lớn dữ liệu học viên, bao gồm kết quả bài kiểm tra, hành vi học tập, sở thích cá nhân và thậm chí là dữ liệu sinh trắc học trong một số hệ thống giám sát. Điều này đặt ra những thách thức lớn về quyền riêng tư và an toàn dữ liệu, như rò rỉ dữ liệu cá nhân, lạm dụng dữ liệu hoặc thiếu minh bạch trong cách AI thu thập và sử dụng thông tin.

Giải pháp: Áp dụng mã hóa dữ liệu mạnh, ẩn danh hóa dữ liệu và quản lý quyền truy cập để ngăn chặn rủi ro bảo mật. Các hệ thống AI cần tuân thủ các tiêu chuẩn bảo mật quốc tế như **GDPR, FERPA, HIPAA**. Đồng thời, các nền tảng học trực tuyến cần công khai chính sách bảo mật và cung cấp quyền kiểm soát dữ liệu cho người học. [17]

Vai trò của giảng viên trong môi trường AI: AI có thể hỗ trợ nhiều khía cạnh của giảng dạy, nhưng việc thay thế hoàn toàn vai trò của giảng viên vẫn chưa khả thi. Học tập không chỉ là quá trình tiếp thu kiến thức mà còn cần sự hướng dẫn, động viên và đánh giá từ con người. AI tuy có thể giúp cá nhân hóa lộ trình học tập, cung cấp phản hồi nhanh, nhưng vẫn gặp hạn chế trong việc tạo động lực và xây dựng môi trường học tập linh hoạt.

Giải pháp: Dùng AI để hỗ trợ chứ không thay thế giảng viên. AI có thể tự động chấm điểm, theo dõi tiến độ học tập, nhưng giảng viên vẫn đóng vai trò quan trọng trong việc hướng dẫn, tạo động lực và đánh giá kỹ năng mềm của học viên. Sự kết hợp giữa AI và giáo viên sẽ giúp tối ưu hóa quá trình giảng dạy, đảm bảo yếu tố con người trong giáo dục. [17]

Thiếu dữ liệu chất lượng cao: AI hoạt động hiệu quả khi có dữ liệu đủ lớn và chất lượng cao. Tuy nhiên, trong giáo dục, dữ liệu có thể bị phân tán, không đồng nhất hoặc thiếu thông tin quan trọng, dẫn đến mô hình AI hoạt động không chính xác.

Giải pháp: Xây dựng hệ thống thu thập và làm sạch dữ liệu từ nhiều nguồn khác nhau. Tạo kho dữ liệu chung để đảm bảo AI có đủ dữ liệu huấn luyện. [16]

Khả năng cá nhân hóa chưa hoàn hảo: Dù AI có thể cá nhân hóa lộ trình học tập, nhưng vẫn khó để hiểu sâu sắc những yếu tố tâm lý, cảm xúc hoặc động lực của từng người học. AI có thể đưa ra đề xuất học tập không thực sự phù hợp với những nhu cầu cá nhân đặc thù.

Giải pháp: Ứng dụng Adaptive Learning AI để tùy chỉnh nội dung học theo thời gian thực, kết hợp với khảo sát phản hồi của học viên để tối ưu hóa cá nhân hóa. [16]

Chống gian lận và tính trung thực trong học tập: E-learning dễ bị lợi dụng để gian lận trong kiểm tra, làm bài tập hoặc thậm chí gian lận danh tính. AI có thể hỗ trợ phát hiện hành vi gian lận, nhưng vẫn chưa thể đảm bảo tính trung thực tuyệt đối.

Giải pháp: Sử dụng AI nhận diện khuôn mặt để xác thực danh tính trong các kỳ thi online. Ứng dụng AI kiểm tra đạo văn, sao chép nội dung để đảm bảo trung thực. [17]

Tương tác người học bị hạn chế: AI có thể hỗ trợ giảng dạy, nhưng không thể thay thế hoàn toàn sự tương tác giữa con người với con người. Những hoạt động như thảo luận nhóm, tranh.

Giải pháp: Kết hợp AI với các nền tảng học tập cộng tác, giúp học viên thảo luận và trao đổi dễ dàng hơn. Phát triển AI tutor có thể giao tiếp bằng ngôn ngữ tự nhiên, tạo cảm giác tương tác giống như với giảng viên thật. [16]

4. KẾT LUẬN

Nghiên cứu đã chỉ ra tiềm năng to lớn của trí tuệ nhân tạo (AI) trong việc nâng cao chất lượng và hiệu quả của các hệ thống E-learning, từ đó mở ra hướng đi mới trong phát triển các giải pháp học tập thông minh. Để khai thác tiềm năng này, việc triển khai một chatbot hỗ trợ tìm kiếm khóa học, giải đáp thắc mắc và đưa ra gợi ý khóa học cá nhân hóa là một giải pháp khả thi. Cụ thể, nhóm sẽ phát triển hệ thống E-learning tích hợp các công nghệ tiên tiến về xử lý ngôn ngữ tự nhiên (NLP) và học máy, trong đó chatbot sẽ sử dụng các mô hình ngôn ngữ hiện đại như GPT hoặc BERT để hiểu và phân tích yêu cầu của học viên, từ đó trích xuất các từ khóa quan trọng như chuyên ngành, cấp độ và mục tiêu học tập. Dựa trên những thông tin này, hệ thống sẽ truy xuất dữ liệu từ cơ sở dữ liệu khóa học và trả về danh sách các khóa học phù hợp với nội dung yêu cầu. Song song đó, một module gợi ý khóa học thông minh, được xây dựng dựa trên các thuật toán Content-Based Filtering và Collaborative Filtering, sẽ phân

tích hành vi, lịch sử học tập và sở thích của từng học viên nhằm đề xuất các khóa học cá nhân hóa, tối ưu hóa quá trình học tập. Chatbot được tích hợp trực tiếp vào giao diện web hoặc ứng dụng di động, cho phép học viên tương tác liền mạch và nhận phản hồi nhanh chóng. Đồng thời áp dụng các biện pháp bảo mật nghiêm ngặt như mã hóa và xác thực để bảo vệ thông tin cá nhân của người dùng.

5. TÀI LIỆU THAM KHẢO

- [1] A. Z. Emi Sita Eriana, Artificial Intelligence (AI), Penerbit Eureka Media Aksara, 2023.
- [2] S. Downes, E-learning 2.0, ACM Press, 2005.
- [3] K. I. G. a. F.-F. C. Naji Ahmad Albatayneh, Utilizing Learners' Negative Ratings in Semantic Content-Based Recommender System for e-Learning Forum, Taiwan: International Forum of Educational Technology & Society, National Taiwan Normal University, 2018.
- [4] F. S. A. H. M. J. Bobadilla, Collaborative filtering adapted to recommender systems of e-learning, Elsevier, 2009.
- [5] M. C. R. O. A.-T. G. H. G. P. T. Ilie Gligorea, Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review, Digital Learning: Exploring the Use of Educational Technology in Educational Environments), 2023.
- [6] P. Kerr, Adaptive learning, Oxford University Press, 2016.
- [7] H. Aldriye, A. Alkhalaf and M. Alkhalaf, Automated Grading Systems for Programming Assignments: A Literature Review, International Journal of Advanced Computer Science and Applications, 2019.
- [8] Langmaster, "Langmaster," [Online]. Available: <https://careers.langmaster.edu.vn/marketing>. [Accessed 16 03 2025].
- [9] G. M. M. B. Teresa Monahan, Virtual reality for collaborative e-learning, Elsevier (Computers & Education Journal), 2008.
- [10] K. Dutta, Augmented Reality for E-Learning, RWTH Aachen University, 2016.
- [11] V. P. E. Venkata Bhanu Prasad Tolety, Enhancing Adaptive E-Learning with Generative AI: Expanding the Horizon Beyond Recommendation Systems, Springer, 2024.
- [12] W. X. B. P. Hanxiang Du, Automatic text generation using deep learning: providing large-scale support for online learning communities, Taylor & Francis Online, 2021.
- [13] MMD, Knewton Personalizes Learning with the Power of AI, 2021: Digital Innovation and Transformation.

- [14] H.-Y. S. S.-C. C. X.-C. H. Gwo-Jen Hwang, A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors, 2020: Elsevier (Computers and Education: Artificial Intelligence).
- [15] The Influence of Artificial Intelligence Tools on Student Performance in e-Learning Environments: Case Study, Academic Publishing International Limited, 2024.
- [16] Y. A. J. A. S. F. S. M. U. Mir Murtaza, AI-Based Personalized E-Learning Systems: Issues, Challenges, and Solutions, 2023.
- [17] P. C. Z. L. Lijia Chen, Artificial Intelligence in Education: A Review, 2020.

ABSTRACT

ARTIFICIAL INTELLIGENCE IN E-LEARNING

Nguyen Phuong Hac^{1,*}, Nguyen Thi Hong Thao¹, Tran Vo Phong¹, Nguyen Thanh Hau¹,
Luong Bao Hoa¹

¹*Ho Chi Minh City University of Industry and Trade*

Email: *hacnp@huit.edu.vn*

Artificial Intelligence (AI) is increasingly asserting its crucial role in various fields, including education. E-learning is a modern educational model that leverages technology and the internet to provide flexible learning content, unrestricted by space and time. Integrating AI into E-learning unlocks numerous potentials, enabling personalized learning experiences, optimizing teaching content, and enhancing knowledge acquisition efficiency.

This study focuses on examining the operation of current E-learning systems and assessing the extent of AI application in this field. Specifically, the paper analyzes AI methods being implemented, including intelligent course recommendations, virtual learning assistants, automated grading systems, and learner behavior analysis. Additionally, the research clarifies the benefits AI brings, as well as the challenges that need to be overcome to develop intelligent online learning systems, enhancing learner experiences in the digital age.

Keywords: AI (Artificial Intelligence), E-Learning, AI Applications in E-learning

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyen Phuong Hac

Số điện thoại: 0983007379

KẾT HỢP TỐI ƯU HÓA TRƯỜNG ĐIỆN TỪ VÀ THUẬT GIẢI DI TRUYỀN: MỘT PHƯƠNG PHÁP METAHEURISTIC ĐỂ TỐI ƯU HÓA TĂNG CƯỜNG

Nguyễn Thị Thu Tâm^{1,*}, Đinh Nguyễn Trọng Nghĩa¹

¹Trường Đại học Công thương Thành phố Hồ Chí Minh

*Email: tamntt@huit.edu.vn

Ngày nhận bài: 20/03/2025, Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Bài báo này giới thiệu một thuật toán metaheuristic mới kết hợp giữa tối ưu trường điện từ và thuật giải di truyền để giải quyết các vấn đề tối ưu hiệu quả. Thuật toán đề xuất lần lượt áp dụng tối ưu trường điện từ và thuật giải di truyền trong mỗi vòng lặp, tận dụng những ưu điểm của cả hai phương pháp. Thực nghiệm mở rộng trên các hàm benchmark thông dụng cho thấy sự vượt trội của thuật toán đề xuất so với các thuật toán thành phần về tốc độ hội tụ và chất lượng giải pháp.

Từ khóa: Tối Ưu Metaheuristic, Tối Ưu Trường Điện Từ, Thuật Toán Di Truyền, Thuật Toán Tối Ưu Kết Hợp, Hàm Đánh Giá Tối Ưu

1. MỞ ĐẦU

Các thuật toán metaheuristic được sử dụng để tìm kiếm một giải pháp gần đúng tối ưu cho các vấn đề tối ưu khó khăn, trong đó không có phương pháp xác định nào có thể giải quyết chúng trong một thời gian hợp lý. Một thuật toán metaheuristic được định nghĩa là một thuật toán độc lập với vấn đề có thể tìm ra các giải pháp gần đúng cho các vấn đề khó khăn. Metaheuristics được lấy cảm hứng từ thiên nhiên và cố gắng giải quyết các vấn đề bằng cách mô phỏng hành vi động vật, sinh học hoặc vật lý. [1]

Các thuật toán tiến hóa (EAs) là các thuật toán metaheuristic ngẫu nhiên, dựa trên quần thể. EAs khác biệt với một số phương pháp tối ưu hóa, như Simulated-Annealing [2] và Tabu search [3], vì chúng tiến hóa một quần thể các giải pháp để đạt được một giải pháp gần đúng tối ưu thay vì chỉ một giải pháp. Đại khái, EAs tìm kiếm miền vấn đề như sau: một quần thể các cá thể ngẫu nhiên (giải pháp) được khởi tạo lần đầu tiên, sau đó đánh giá sức khỏe của các cá thể thông qua hàm sức khỏe. Trong các thế hệ tiếp theo, các cá thể tiến hóa về phía giải pháp tốt nhất toàn cầu thông qua EAs và sự hướng dẫn của hàm sức khỏe. Quá trình này tiếp tục cho đến khi đạt đến số lần lặp tối đa hoặc tìm thấy giải pháp gần đúng kỳ vọng. [4] [5]

Khả năng cân bằng giữa khám phá (đa dạng hóa) và khai thác (củng cố) đóng một vai trò quan trọng trong sự thành công của một EA. Sự khai thác là cần thiết để khám phá bề mặt vấn đề một cách toàn cầu và xác định khu vực của không gian tìm kiếm chứa giải pháp tốt nhất toàn cầu (giá trị nhỏ nhất toàn cầu). Sự khai thác cũng cần thiết để tìm ra một giải pháp chính

xác bằng cách tăng cường tìm kiếm trong khu vực được xác định bởi giai đoạn khám phá. Đạt được sự cân bằng này là đặc điểm chính của EAs và cách chúng khác biệt với nhau. Đại khái, EAs có hướng khám phá hơn là hướng khai thác. Đặc điểm này làm cho chúng phù hợp cho các vấn đề khó với nhiều giải pháp tối ưu địa phương vì chúng giữ một quần thể các giải pháp và nghiên cứu một khu vực lớn để tìm ra giải pháp tốt nhất toàn cầu.

Tính cân bằng giữa sự khám phá và sự khai thác là đặc điểm quan trọng của EAs và đó là cách chúng khác biệt với nhau. Nhìn chung, EAs có xu hướng tập trung vào khám phá hơn là khai thác. Điều này làm cho chúng trở nên phù hợp với các vấn đề khó với nhiều giải pháp tối ưu địa phương vì chúng duy trì một quần thể các giải pháp và khám phá một khu vực lớn để tìm ra giải pháp tốt nhất toàn cầu.

Nhiều thuật toán tối ưu hóa được lấy cảm hứng từ thiên nhiên đã trở nên rất nổi tiếng, trong đó có: Genetic Algorithm (GA) [6], hoạt động dựa trên nguyên tắc của lý thuyết tiến hóa của Darwin, Particle Swarm Optimization (PSO) [7], hoạt động dựa trên hành vi tìm kiếm thức ăn của một đàn chim, Differential Evolution (DE) [8], tương tự như GA nhưng có các phương pháp lai và đột biến được sửa đổi, và Harmony Search (HS) [9], hoạt động dựa trên cách mà nhạc sĩ thử nghiệm và thay đổi nốt nhạc của các nhạc cụ để sáng tác âm nhạc tốt hơn. Các thuật toán tối ưu hóa nổi tiếng này đã được áp dụng cho việc tối ưu hóa của nhiều vấn đề tối ưu hóa và đã chứng minh được độ chính xác và tốc độ chấp nhận được. Tuy nhiên, thách thức ở trong lĩnh vực này là làm thế nào để được lấy cảm hứng từ kiến thức hiện có về tối ưu hóa và cách mà tự nhiên hoạt động để cải thiện các thuật toán hiện có hoặc phát triển một thuật toán mới, có thể nhanh chóng tìm ra giải pháp tốt nhất toàn cầu mà vẫn tránh giải pháp tối ưu cục bộ.

Hai thuật toán tối ưu hóa mới được đề xuất gần đây là Artificial Bee Colony (ABC) [10] và Group Search Optimizer (GSO) [11]. ABC là một thuật toán tối ưu hóa dựa trên hành vi tìm kiếm thức ăn thông minh của đàn ong mật, trong khi thuật toán GSO được lấy cảm hứng từ hành vi tìm kiếm thức ăn của động vật. Cả hai thuật toán này đã được áp dụng thành công cho nhiều vấn đề tối ưu hóa khác nhau và đã chứng minh được sức mạnh tìm kiếm cạnh tranh.

Các thuật toán heuristics lấy cảm hứng từ vật lý là một loại thuật toán tối ưu hóa khác, được thúc đẩy bởi các định luật vật lý. Vào năm 2008, Tayarani đã đề xuất một thuật toán tối ưu hóa lấy cảm hứng từ từ tính được gọi là Magnetic Optimization Algorithm (MOA) [12], được lấy cảm hứng từ lực hút giữa các hạt từ tính. Trong MOA, các giải pháp có thể được thể hiện bằng nam châm có khối lượng khác nhau, được phân tán trong một cấu trúc giống như lưới trên không gian tìm kiếm và áp dụng một lực hút đối với các hàng xóm của chúng, theo sức khỏe của mình. Một thuật toán tương tự với MOA là Gravitational Search Algorithm (GSA) [13], mà xem xét các yếu tố tìm kiếm (hạt) như các khối lượng tương tác với nhau dựa trên lực hấp dẫn giữa chúng. Một số thuật toán khác gần đây đã sử dụng một ý tưởng tương tự cho tối ưu hóa dựa trên các định luật lực học lấy cảm hứng từ vật lý. Magnetic Charged System Search (MCSS) [14] sử dụng các định luật điều khiển cho lực từ cộng thêm vào các lực điện để tối ưu hóa. Ions Motion Optimization (IMO) [15] được đề xuất dựa trên sự hấp dẫn và đẩy lùi của anion và cation để thực hiện tối ưu hóa.

Được thúc đẩy bởi các thuật toán dựa trên trường lực đã được đề xuất trước đó và sự thực tế rằng trong hầu hết các trường hợp, các hạt được hấp dẫn đến hạt sẽ cho kết quả tốt nhất, điều này tăng khả năng tìm ra điểm tối ưu cục bộ, một thuật toán tối ưu hóa lấy cảm hứng từ vật lý khác được gọi là Electromagnetic Field Optimization (EFO) [16] [17], được lấy cảm hứng từ hai hiện tượng. Hiện tượng đầu tiên là lực hấp dẫn - đẩy lùi giữa các tụ điện có cực khác nhau, và hiện tượng thứ hai là một tỷ lệ lấy cảm hứng từ thiên nhiên gọi là tỷ lệ vàng, còn được biết đến là bàn tay của Chúa.

Dựa trên các thuật toán lấy cảm hứng từ vật lý và sinh học đã được đề xuất trước đó, và với mục tiêu tăng cường sự đa dạng hóa và khai thác trong quá trình tìm kiếm giải pháp, bài báo này đề xuất một phương pháp tối ưu mới, kết hợp giữa thuật toán trường điện từ và thuật

giải di truyền. Thuật toán này, được gọi là Electromagnetic Genetic Optimization Algorithm (EGOA), sẽ tận dụng sức mạnh của cả vật lý và sinh học trong quá trình tìm kiếm giải pháp tối ưu.

Trong EGOA, ở mỗi vòng lặp, thuật toán sẽ áp dụng một phần của thuật toán trường điện từ để tạo ra sự đa dạng trong quần thể giải pháp. Sau đó, thuật giải di truyền sẽ được sử dụng để tìm kiếm và cải thiện các giải pháp trong quần thể này. Bằng cách kết hợp giữa việc tạo ra đa dạng và khai thác tốt nhất từ cả hai phương pháp, EGOA hứa hẹn mang lại những kết quả tối ưu hóa hấp dẫn và hiệu quả hơn cho các vấn đề tối ưu khó khăn.

2. TỐI ƯU TRƯỜNG ĐIỆN TỪ

Một nam châm điện hoạt động dựa trên nguyên lý rằng dòng điện đi qua tạo ra một trường từ. Khác với nam châm cố định, nam châm điện chỉ có một cực (dương hoặc âm), được xác định bởi hướng của dòng điện và có thể thay đổi bằng cách thay đổi hướng của dòng điện. Hơn nữa, có hai lực khác nhau giữa các nam châm điện: lực hút và lực đẩy. Các nam châm điện cùng cực sẽ đẩy lẫn nhau, trong khi những nam châm có cực trái ngược sẽ hút lẫn nhau. Lực hút giữa các nam châm điện mạnh hơn lực đẩy khoảng (5-10%). Thuật toán tối ưu trường điện từ [16] sử dụng những khái niệm này và thay thế tỷ lệ giữa lực hút và lực đẩy bằng tỷ lệ vàng. Điều này giúp các hạt khám phá không gian tìm kiếm vấn đề một cách đầy đủ và tìm ra giải pháp gần tối ưu.

EFO là một thuật toán dựa trên quần thể và mỗi vector giải pháp được biểu diễn bởi một nhóm của nam châm điện (hạt điện từ từ trường điện từ). Số lượng nam châm điện của một hạt điện từ được xác định bởi số biến của bài toán tối ưu hóa. Do đó, mỗi nam châm điện của hạt điện từ tương ứng với một biến của bài toán tối ưu hóa. Hơn nữa, tất cả các nam châm điện của cùng một hạt điện từ có cùng cực. Tuy nhiên, mỗi nam châm điện có thể áp dụng một lực hút hoặc đẩy lên các nam châm đồng loại tương ứng với cùng một biến của bài toán tối ưu hóa.

EFO tìm kiếm trong miền bài toán như sau: trước tiên, một quần thể các hạt điện từ được tạo ra ngẫu nhiên, và độ thích nghi của mỗi hạt được đánh giá bằng một hàm thích nghi; sau đó, các hạt được sắp xếp theo độ thích nghi của chúng. Thứ hai, các hạt đã được sắp xếp được chia thành ba nhóm, và một phần của quần thể điện từ được phân bổ cho mỗi nhóm; nhóm đầu tiên được gọi là trường dương và bao gồm các hạt điện từ thích nghi nhất với cực dương, nhóm thứ hai được gọi là trường âm và bao gồm các hạt điện từ có độ thích nghi thấp nhất và cực âm, và các hạt điện từ còn lại tạo thành một nhóm gọi là trường trung tính, có một cực âm nhỏ gần bằng không. Cuối cùng, trong mỗi vòng lặp của thuật toán, một hạt điện từ mới được hình thành và được đánh giá bằng một hàm thích nghi. Nếu hạt điện từ được tạo ra mạnh mẽ hơn so với hạt điện từ xấu nhất trong quần thể, thì hạt được tạo ra sẽ được chèn vào quần thể đã sắp xếp theo độ thích nghi của nó và nhận một cực dựa trên vị trí của nó trong quần thể; hơn nữa, hạt xấu nhất sẽ bị loại bỏ. Quá trình này tiếp tục cho đến khi đạt được số lần lặp tối đa hoặc tìm được giải pháp gần tối ưu như mong muốn.

EFO xác định vị trí của mỗi nam châm điện trong một hạt điện từ được tạo ra như sau: từ các hạt điện từ của mỗi trường điện từ (dương, âm và trung tính), ba nam châm điện đồng loại được chọn ngẫu nhiên (một nam châm điện từ mỗi trường). Sau đó, nam châm điện từ được tạo ra nhận vị trí và cực (cực âm nhỏ) của nam châm điện từ được chọn từ trường trung tính và bị ảnh hưởng bởi các nam châm điện từ được chọn từ trường dương (hút) và trường âm (đẩy) bằng cường độ lực ngẫu nhiên. Nói cách khác, nam châm điện từ được tạo ra di chuyển một khoảng cách xa khỏi các giải pháp tồi và tiếp cận các giải pháp tốt. Mỗi nam châm điện từ của các hạt điện từ trong trường trung tính bị ảnh hưởng bởi hai nam châm điện từ đồng loại ngẫu nhiên từ trường âm (đẩy) và trường dương (hút). Sự tồn tại đồng thời của hai lực trái ngược giữa các nam châm điện và việc tạo ra một giải pháp mới bằng cách di chuyển

một khoảng cách xa khỏi các giải pháp tồi và tiến gần hơn đến các giải pháp tốt gây ra sự tìm kiếm hiệu quả và hội tụ nhanh chóng. Tuy nhiên, để duy trì sự đa dạng và tránh các cực tiểu cục bộ, sự ngẫu nhiên là một phần không thể thiếu của EFO. Do đó, đối với một số hạt điện từ được tạo ra (không phải tất cả), chỉ có một trong các nam châm điện từ được thay đổi bằng một nam châm điện từ được tạo ra ngẫu nhiên. Lý do áp dụng sự ngẫu nhiên cho một số hạt điện từ là sự tồn tại của các biến ngẫu nhiên trong tất cả các giải pháp sẽ ngăn cản việc tìm ra một giải pháp tốt. Tuy nhiên, việc áp dụng sự ngẫu nhiên cho một số giải pháp mang lại sự đa dạng vào quần thể và ngăn chặn các cực tiểu cục bộ.

Dưới đây là mã giả cho thuật toán tối ưu trường điện từ

Thuật toán 1. Thuật toán tối ưu trường điện từ tổng quát

Procedure Electromagnetic_Field_Optimization():

 Initialize_population() // Khởi tạo quần thể các hạt điện từ ngẫu nhiên

 Evaluate_population() // Đánh giá độ thích nghi của từng hạt điện từ

 Sort_population() // Sắp xếp các hạt điện từ theo độ thích nghi

 Repeat until termination criteria are met:

 Divide_population() // Chia quần thể thành ba nhóm: dương, âm và trung tính

 Generate_new_particle() // Tạo ra một hạt điện từ mới

 Evaluate_particle() // Đánh giá độ thích nghi của hạt điện từ mới

 Update_population() // Cập nhật quần thể và loại bỏ hạt điện từ xấu nhất

 Return the best solution found

End Procedure

Trong đó thủ tục Generate_new_particle() sẽ tạo ra một hạt mới, sau đó thực hiện di chuyển hạt này về phía trường dương và rời xa trường âm của trường điện từ. Chú ý rằng lực hút sẽ lớn hơn lực đẩy 1.6 lần theo quy tắc vàng.

3. THUẬT GIẢI DI TRUYỀN (GA)

Thuật giải Di truyền [6] là một phương pháp tối ưu hóa dựa trên cơ chế tiến hóa tự nhiên. Nó được lấy cảm hứng từ quá trình tiến hóa tự nhiên trong tự nhiên, trong đó các cá thể tốt được chọn lọc và tái sản xuất để tạo ra thế hệ mới với các đặc tính tốt hơn. Ý tưởng cơ bản của GA là sử dụng một quần thể các cá thể (hay còn gọi là "cá thể" hoặc "cá thể cá nhân"), mỗi cá thể được biểu diễn bằng một "gen" hoặc "chromosome", và sử dụng các phép toán như lai ghép (crossover) và đột biến (mutation) để tạo ra các thế hệ mới có các cá thể có đặc tính tốt hơn. Quần thể tiếp tục được tiến hóa qua các thế hệ cho đến khi đạt được một điều kiện dừng được xác định trước hoặc đạt được giải pháp tối ưu.

Phương pháp thực hiện của Thuật giải Di truyền:

Khởi tạo quần thể ban đầu: Bắt đầu với việc tạo ra một quần thể ban đầu các cá thể ngẫu nhiên. Mỗi cá thể thường được biểu diễn bằng một vector gen hoặc một chuỗi các gen, mỗi gen đại diện cho một đặc tính của cá thể.

Đánh giá độ thích nghi (fitness evaluation): Độ thích nghi của mỗi cá thể trong quần thể được đánh giá bằng một hàm mục tiêu (hàm fitness) đo lường mức độ tương thích của cá thể với vấn đề tối ưu hóa. Điều này giúp xác định cá thể nào có đặc tính tốt hơn.

Chọn lọc (selection): Các cá thể được lựa chọn từ quần thể ban đầu để tạo ra quần thể mới cho thế hệ tiếp theo. Các cá thể có độ thích nghi cao hơn có khả năng được chọn lọc và

tái sản xuất với tỷ lệ cao hơn, trong khi các cá thể có độ thích nghi thấp hơn có khả năng bị loại bỏ.

Lai ghép (crossover): Các cá thể được chọn để tham gia quá trình lai ghép, trong đó các gen của chúng được trao đổi để tạo ra các cá thể con mới.

Đột biến (mutation): Một số cá thể con có thể trải qua quá trình đột biến, trong đó một hoặc nhiều gen của chúng được thay đổi một cách ngẫu nhiên.

Tạo ra thế hệ mới: Thế hệ mới được tạo ra từ các cá thể được chọn lọc, lai ghép và đột biến.

Lặp lại quá trình: Quá trình lai ghép, đột biến và chọn lọc được lặp lại qua nhiều thế hệ cho đến khi một điều kiện dừng được đáp ứng hoặc một giải pháp tối ưu được tìm thấy.

Dưới đây là mã giả cho thuật giải di truyền

Thuật toán 2. Thuật toán cho thuật giải di truyền dạng tổng quát

Procedure Genetic_Algorithm():

 Initialize_population() // Khởi tạo quần thể ban đầu

 Evaluate_population() // Đánh giá độ thích nghi của từng cá thể trong quần thể

 Repeat until termination criteria are met:

 Selection() // Chọn lọc các cá thể cho lai ghép và đột biến

 Crossover() // Lai ghép các cá thể được chọn

 Mutation() // Đột biến các cá thể sau lai ghép

 Evaluate_population() // Đánh giá lại độ thích nghi của từng cá thể trong quần thể

 Next_generation() // Tạo thế hệ tiếp theo từ các cá thể được lai ghép và đột biến

 Return the best solution found

End Procedure

4. KẾT HỢP GIỮA THUẬT TOÁN TRƯỜNG ĐIỆN TỬ VÀ THUẬT GIẢI DI TRUYỀN

4.1. Nguyên tắc kết hợp và cơ chế hoạt động

Kết hợp giữa thuật toán Trường Điện Tử và thuật giải Di Truyền nhằm tận dụng sức mạnh của cả hai phương pháp trong việc tìm kiếm giải pháp tối ưu cho các bài toán phức tạp. Bằng cách kết hợp hai thuật toán này, chúng ta có thể tận dụng các ưu điểm của cả hai để tạo ra một phương pháp tối ưu hóa mạnh mẽ, có khả năng khám phá không gian tìm kiếm rộng lớn và đồng thời tránh rơi vào các điểm cục bộ.

Sử dụng Trường Điện Tử để tạo ra các giải pháp ban đầu: Thuật toán Trường Điện Tử được sử dụng để tạo ra một số lượng lớn các giải pháp ban đầu. Các giải pháp này được tạo ra dựa trên các nguyên lý vật lý của Trường Điện Tử, giúp đảm bảo rằng chúng phân phối rộng rãi trên không gian tìm kiếm.

Áp dụng thuật giải Di Truyền để cải thiện các giải pháp: Sau khi có được các giải pháp ban đầu từ Trường Điện Tử, thuật giải Di Truyền được sử dụng để cải thiện chúng. Quá trình này bao gồm việc chọn lọc, lai ghép và đột biến các giải pháp để tạo ra các thế hệ tiếp theo với các giải pháp có độ thích nghi tốt hơn.

Kết hợp giữa sự khám phá và khai thác: Sự kết hợp giữa Trường Điện Tử và Di Truyền giúp cân bằng giữa việc khám phá không gian tìm kiếm (thông qua Trường Điện Tử) và việc

khai thác các khu vực tốt của không gian tìm kiếm (thông qua Di Truyền). Điều này giúp thuật toán đạt được sự đa dạng và khả năng tìm kiếm tốt hơn.

Lặp lại quá trình cải thiện: Quá trình kết hợp giữa Trường Điện Từ và Di Truyền được lặp lại qua nhiều thế hệ để tìm ra giải pháp tối ưu hoặc tiến gần đến giải pháp tối ưu nhất có thể.

Bằng cách kết hợp sức mạnh của cả Trường Điện Từ và Di Truyền, phương pháp này có khả năng tìm kiếm và cải thiện các giải pháp một cách hiệu quả và đa dạng, giúp tối ưu hóa các bài toán phức tạp trong nhiều lĩnh vực ứng dụng.

4.2. Phương pháp thực hiện

Thuật toán đề xuất có thể được thực hiện theo các bước sau:

Khởi tạo quần thể ban đầu: Bắt đầu bằng việc sử dụng thuật toán Trường Điện Từ để tạo ra một quần thể ban đầu các giải pháp. Các giải pháp này được tạo ra dựa trên các nguyên tắc vật lý của Trường Điện Từ và phân bố rộng rãi trên không gian tìm kiếm.

Đánh giá độ thích nghi: Đánh giá độ thích nghi của mỗi giải pháp trong quần thể sử dụng hàm mục tiêu được xác định trước. Hàm mục tiêu thường đo lường mức độ tương thích của giải pháp với bài toán tối ưu hóa.

Lựa chọn: Sử dụng thuật giải Di Truyền để lựa chọn các cá thể có độ thích nghi tốt nhất từ quần thể hiện tại để tham gia vào quá trình tiếp theo của thuật toán.

Lai ghép (Crossover): Áp dụng quá trình lai ghép giữa các cá thể đã được lựa chọn để tạo ra các cá thể con mới. Quá trình lai ghép giữa hai cá thể tạo ra các cá thể con mới bằng cách trao đổi một phần gen giữa chúng.

Đột biến (Mutation): Áp dụng quá trình đột biến lên các cá thể con mới để tạo ra sự đa dạng trong quần thể. Đột biến thường là việc thay đổi ngẫu nhiên một phần gen của các cá thể.

Tạo ra một hạt điện từ mới và thực hiện di chuyển về các trường dương hoặc tránh xa trường âm.

Đánh giá độ thích nghi của thế hệ mới: Đánh giá lại độ thích nghi của các cá thể con mới tạo ra sau lai ghép và đột biến.

Chọn lọc lại: Lựa chọn các cá thể tốt nhất từ thế hệ mới để tiếp tục quá trình tối ưu hóa. Các cá thể tốt nhất có thể được giữ lại hoặc được lựa chọn để tiếp tục tham gia vào quá trình lai ghép và đột biến.

Lặp lại quá trình: Quá trình lai ghép, đột biến và lựa chọn lại được lặp lại qua nhiều thế hệ cho đến khi một điều kiện dừng được đáp ứng hoặc một giải pháp tối ưu được tìm thấy.

Mã giải cho thuật toán như sau

Thuật toán 3. Thuật toán kết hợp đề xuất EGOA

Procedure EGOA():

Initialize_population() // Khởi tạo quần thể ban đầu bằng thuật toán Trường Điện Từ
Evaluate_population() // Đánh giá độ thích nghi của từng cá thể trong quần thể
Repeat until termination criteria are met:
 Selection() // Chọn lọc các cá thể cho lai ghép và đột biến bằng thuật giải Di Truyền
 Crossover() // Lai ghép các cá thể được chọn lọc
 Mutation() // Đột biến các cá thể sau lai ghép
 Divide_population() // Chia quần thể thành 3 nhóm: dương, âm và trung tính

```
Generate_new_particle() // Tạo ra một hạt điện từ mới
Evaluate_population() // Đánh giá lại độ thích nghi của từng cá thể trong quần thể
Next_generation() // Tạo thế hệ tiếp theo từ các cá thể được lai ghép và đột biến
Return the best solution found
End Procedure
```

5. KẾT QUẢ VÀ THẢO LUẬN

Trong nghiên cứu này, chúng tôi thực hiện thử nghiệm và đánh giá hiệu suất của ba phương pháp tối ưu hóa: Electromagnetic Field Optimization (EFO), Genetic Algorithm (GA), và thuật toán mới được đề xuất, Evolutionary Genetic Optimization Algorithm (EGOA). Các thử nghiệm được tiến hành trên bốn hàm thử nghiệm phổ biến trong lĩnh vực tối ưu hóa Bảng 1.

Bảng 1. Các hàm benchmark dùng trong bài báo

Tên hàm	Công thức toán học	Phạm vi
F1	$F_1(x) = \sum_{i=1}^n x_i^2$	[-2, 2]
F2	$F_2(x) = \sum_{i=1}^n x_i + \prod_{i=1}^n x_i $	
F3	$F3(x) = \sum_{i=1}^n \left(\sum_{j=1}^i x_j \right)^2$	
F4	$F4(x) = \max_i \{ x_i , 1 \leq i \leq n\}$	

Môi trường thực nghiệm: Thực hiện trên máy tính với cấu hình: CPU: AMD Ryzen 5 3550H, RAM: 16GB, Tốc độ xử lý: 2.10 GHz. Sử dụng ngôn ngữ lập trình Python phiên bản 3.11.2 để triển khai thuật toán và thực hiện các thử nghiệm. Đối với mỗi hàm thử nghiệm, thực hiện nhiều lần chạy để đảm bảo tính chính xác của kết quả.

Bảng 2. Kết quả trung bình từ 3 thuật toán sau 500 vòng lặp

	F1	F2	F3	F4
EFO	1.44E-41	5.22E-21	7.95E-24	1.56E-17
GA	4.33E-40	4.19E-25	9.59E-19	3.36E-13
EGOA	4.81E-281	3.14E-142	7.84E-151	1.39E-122

Trong bảng 2, trình bày kết quả trung bình của ba thuật toán: Electromagnetic Field Optimization (EFO), Genetic Algorithm (GA), và Evolutionary Genetic Optimization Algorithm (EGOA) trên bốn hàm thử nghiệm khác nhau.

Bảng 3. Kết quả tốt nhất từ 3 thuật toán sau 500 vòng lặp

	F1	F2	F3	F4
EFO	4.74E-42	6.57E-23	6.73E-25	9.62E-18
GA	3.45E-42	3.34E-27	2.09E-21	4.27E-14
EGOA	1.19E-291	1.98E-149	1.64E-160	3.64E-142

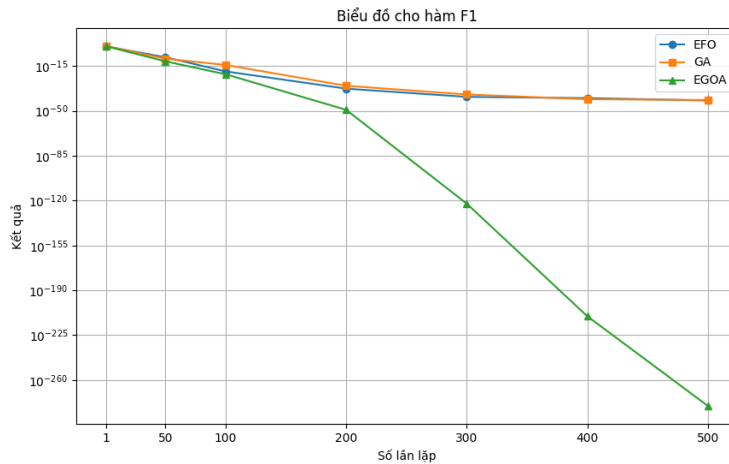
Bảng 3 mô tả kết quả tốt nhất mà mỗi thuật toán đạt được trên từng hàm thử nghiệm.

Kết quả cho thấy, đối với cả bốn hàm thử nghiệm, thuật toán EGOA đã vượt trội hơn so với EFO và GA. EGOA không chỉ có khả năng hội tụ cao mà còn đạt được giải pháp tối ưu nhanh chóng.

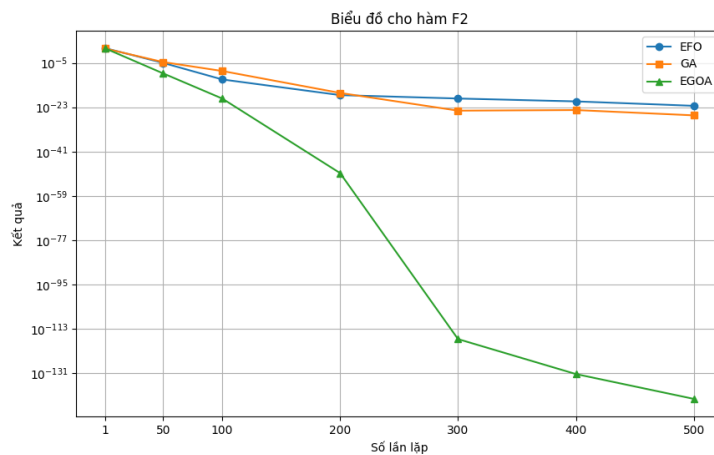
EGOA thể hiện sự hiệu quả của việc kết hợp giữa hai phương pháp tối ưu hóa khác nhau. Bằng cách sử dụng cả các yếu tố của thuật toán Trường Điện Từ và thuật giải Di Truyền, EGOA có thể khám phá không gian tìm kiếm một cách hiệu quả hơn, đồng thời tránh được các hố chết và tìm ra giải pháp tối ưu với thời gian thực thi ngắn hơn.

Kết quả này làm nổi bật tiềm năng của phương pháp kết hợp giữa thuật toán Trường Điện Từ và thuật giải Di Truyền trong việc giải quyết các vấn đề tối ưu hóa phức tạp.

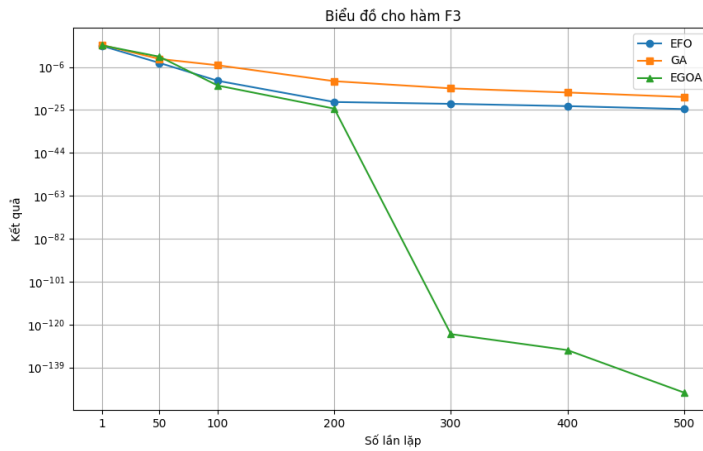
Dưới đây là các đồ thị biểu diễn hiệu suất của các thuật toán tối ưu hóa - EFO, GA và EGOA - trên mỗi hàm thử nghiệm. Các đồ thị này được xây dựng dựa trên giá trị hàm mục tiêu và số lượng vòng lặp.



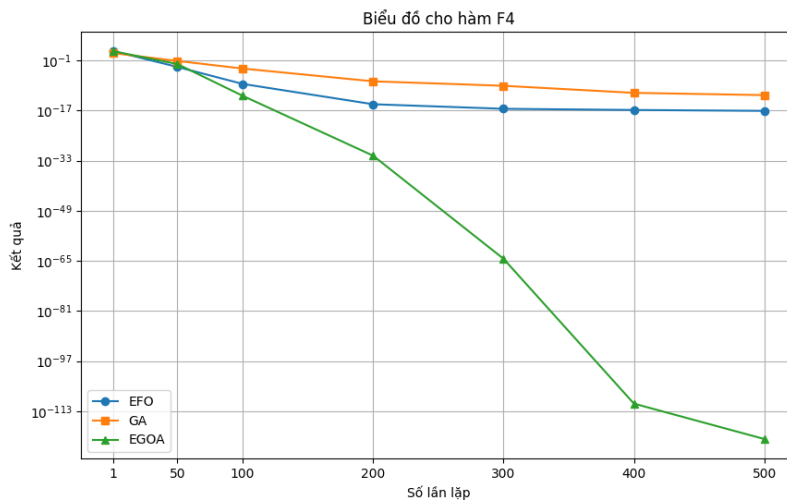
Hình 1. Biểu đồ thể hiện kết quả tối ưu trên hàm F1 qua các vòng lặp



Hình 2. Biểu đồ thể hiện kết quả tối ưu trên hàm F2 qua các vòng lặp



Hình 3. Biểu đồ thể hiện kết quả tối ưu trên hàm F3 qua các vòng lặp



Hình 4. Biểu đồ thể hiện kết quả tối ưu trên hàm F4 qua các vòng lặp

Từ các đồ thị, chúng ta có thể nhận thấy sự hội tụ nhanh chóng của thuật toán mới EGOA so với các thuật toán cũ EFO và GA trên mỗi hàm thử nghiệm. Điều này minh chứng cho khả năng hiệu quả và mạnh mẽ của EGOA trong việc tìm kiếm giải pháp tối ưu trong số lượng vòng lặp ít hơn, giúp tiết kiệm thời gian tính toán và tăng hiệu suất toàn diện của thuật toán.

6. KẾT LUẬN

Việc kết hợp giữa thuật toán tối ưu trường điện từ và thuật giải di truyền đã đem lại những hiệu quả đáng kể trong việc giải quyết các bài toán tối ưu. Phương pháp này đã được chứng minh là hiệu quả và triển vọng là một lựa chọn tiềm năng cho việc tối ưu hóa trong nhiều lĩnh vực ứng dụng. Kết quả thử nghiệm đã mạnh mẽ hỗ trợ sự vượt trội của thuật toán kết hợp EGOA so với các phương pháp đối lập, đặc biệt là về khả năng hội tụ nhanh và đạt được kết quả tối ưu.

Bài báo đã trình bày một cách chi tiết cả nguyên lý hoạt động lẫn kết quả thực nghiệm của phương pháp kết hợp này, mở ra những triển vọng mới trong lĩnh vực tối ưu hóa metaheuristic. Tuy nhiên, vẫn còn nhiều cơ hội để nghiên cứu và cải tiến thuật toán, cũng như mở rộng ứng dụng của nó trong các bối cảnh phức tạp hơn. Đây là một hướng đi tiềm năng và đầy hứa hẹn trong lĩnh vực nghiên cứu tối ưu hóa.

TÀI LIỆU THAM KHẢO

1. Abualigah, L. Group search optimizer: a nature-inspired meta-heuristic optimization algorithm with its results, variants, and applications. *Neural Comput & Applic* 33, 2949–2972 (2021)
2. Alia, O.M., Mandava, R. The variants of the harmony search algorithm: an overview. *Artif Intell Rev* 36, 49–68 (2011)
3. Behzad Javidy, Abdolreza Hatamlou, Seyedali Mirjalili, Ions motion algorithm for solving optimization problems, *Applied Soft Computing*, Volume 32 (2015)
4. Dikshit Chauhan, Anupam Yadav, An archive-based self-adaptive artificial electric field algorithm with orthogonal initialization for real-parameter optimization problems, *Applied Soft Computing*, Volume 150, (2024)
5. Dimitris Bertsimas. John Tsitsiklis. "Simulated Annealing." *Statist. Sci.* 8 (1) 10 – 15 (1993).
6. Gendreau, M., Potvin, JY. Tabu Search. In: Burke, E.K., Kendall, G. (eds) *Search Methodologies*. Springer, Boston (2005)
7. Goldberg, D. E. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. (1989)
8. Hancer, E. Artificial Bee Colony: Theory, Literature Review, and Application in Image Segmentation. In: Hemanth, D., Kumar, B., Manavalan, G. (eds) *Recent Advances on Memetic Algorithms and its Applications in Image Processing*. *Studies in Computational Intelligence*, vol 873. Springer, Singapore (2020)
9. Hosein Abedinpourshotorban, Siti Mariyam Shamsuddin, Zahra Beheshti, Dayang N.A. Jawawi, *Electromagnetic field optimization: A physics-inspired metaheuristic optimization algorithm*, *Swarm and Evolutionary Computation*, Volume 26 (2016)
10. Ilhem Boussaïd, Julien Lepagnot, Patrick Siarry, A survey on optimization metaheuristics, *Information Sciences* 237 (2013).
11. Kaveh, A., Motie Share, M.A. & Moslehi, M. Magnetic charged system search: a new meta-heuristic algorithm for optimization. *Acta Mech* 224, 85–107 (2013)
12. M. H. Tayarani-N and M. R. Akbarzadeh-T, "Magnetic Optimization Algorithms a new synthesis," 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), Hong Kong, China (2008)
13. Mohamed Abdel-Basset, Laila Abdel-Fatah, Arun Kumar Sangaiah, Chapter 10 - Metaheuristic Algorithms: A Comprehensive Review, Editor(s): Arun Kumar Sangaiah, Michael Sheng, Zhiyong Zhang, In *Intelligent Data-Centric Systems, Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*, Academic Press (2018)
14. Nikhil Aditya, Siba Sankar Mahapatra, Switching from exploration to exploitation in gravitational search algorithm based on diversity with Chaos, *Information Sciences*, Volume 635 (2023)

15. Price, K.V. Differential Evolution. In: Zelinka, I., Snášel, V., Abraham, A. (eds) Handbook of Optimization. Intelligent Systems Reference Library, vol 38. Springer, Berlin, Heidelberg (2013)
16. Tansel Dokeroglu, Ender Sevinc, Tayfun Kucukyilmaz, Ahmet Cosar, A survey on new generation metaheuristic algorithms, Computers & Industrial Engineering, Volume 137, 106040 (2009)

ABSTRACT

COMBINING ELECTROMAGNETIC OPTIMIZATION AND GENETIC ALGORITHM: A METAHEURISTIC APPROACH FOR ENHANCED OPTIMIZATION

Nguyen Thi Thu Tam^{1,*}, Dinh Nguyen Trong Nghia¹

¹*Ho Chi Minh city University of Industry and Trade*

Email: tamntt@huit.edu.vn

This paper introduces a novel metaheuristic algorithm that combines electromagnetic optimization with genetic algorithm to tackle optimization problems effectively. The proposed algorithm alternates between applying electromagnetic optimization and genetic algorithm in each iteration, leveraging the strengths of both approaches. Extensive experiments on commonly used benchmark functions demonstrate the superiority of the proposed algorithm over individual component algorithms in terms of convergence speed and solution quality.

Keywords: Metaheuristic Optimization, Electromagnetic Field Optimization, Genetic Algorithm, Combined Optimization Algorithm, Optimization Benchmark Functions.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thị Thu Tâm

Số điện thoại liên hệ: 0909927572

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG HỌC TẬP CỦA SINH VIÊN TẠI CÁC TRƯỜNG ĐẠI HỌC HIỆN NAY

Nguyễn Thị Thu Tâm^{1,*}, Lê Thị Thùy Lan¹

¹Trường Đại học Công thương Thành phố Hồ Chí Minh

*Email: tamntt@huit.edu.vn

Ngày nhận bài: 20/03/2025, Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Trong bối cảnh công nghệ phát triển mạnh mẽ, trí tuệ nhân tạo (AI) đang dần trở thành một công cụ hỗ trợ đắc lực trong học tập của sinh viên tại các trường đại học. AI giúp cá nhân hóa quá trình học tập thông qua các nền tảng học trực tuyến, gợi ý tài liệu phù hợp và cung cấp lộ trình học tập tối ưu. Các trợ lý ảo như ChatGPT, Google Bard hay các hệ thống hỗ trợ tự động giúp sinh viên tìm kiếm thông tin, tóm tắt nội dung và giải đáp thắc mắc một cách nhanh chóng. Ngoài ra, AI còn được ứng dụng trong việc chấm điểm tự động, phát hiện đạo văn, hỗ trợ học ngoại ngữ và thực hành lập trình. Nhờ sự phát triển của AI, sinh viên có thể tiếp cận nguồn tri thức phong phú, nâng cao hiệu suất học tập và rèn luyện kỹ năng một cách linh hoạt và hiệu quả hơn.

Từ khóa: Ứng dụng của AI trong học tập, Trí tuệ nhân tạo trong giáo dục hiện nay, Chatbot AI, Khai thác sức mạnh của AI trong học tập.

1. MỞ ĐẦU

Trí tuệ nhân tạo (AI) đang ngày càng chứng tỏ vai trò quan trọng trong giáo dục, đặc biệt là trong quá trình học tập của sinh viên ở các trường đại học. Nhờ AI, sinh viên có thể tiếp cận nhiều công cụ hỗ trợ học tập như trợ lý ảo, nền tảng học trực tuyến thông minh và hệ thống gợi ý tài liệu. Những công cụ này giúp cá nhân hóa quá trình học tập, giúp sinh viên dễ dàng tìm kiếm thông tin, nắm bắt kiến thức một cách nhanh chóng và hiệu quả hơn. Không chỉ hỗ trợ tiếp cận tri thức, AI còn giúp sinh viên nâng cao kỹ năng tự học và quản lý thời gian. Các ứng dụng AI có thể tạo ra lộ trình học tập phù hợp với năng lực cá nhân, nhắc nhở lịch học và cung cấp các bài tập thực hành tự động. Bên cạnh đó, các công cụ như Grammarly hay Turnitin hỗ trợ kiểm tra ngữ pháp, phát hiện đạo văn, giúp sinh viên hoàn thiện bài viết học thuật một cách chính xác và chuyên nghiệp hơn.

Ngoài ra, AI còn góp phần cải thiện phương pháp giảng dạy trong các trường đại học, giúp giảng viên có thể theo dõi quá trình học tập của sinh viên và điều chỉnh nội dung giảng dạy phù hợp. Việc ứng dụng AI vào giáo dục không chỉ giúp nâng cao hiệu quả học tập mà còn tạo điều kiện cho sinh viên phát triển tư duy sáng tạo, làm quen với công nghệ hiện đại và chuẩn bị tốt hơn cho tương lai nghề nghiệp trong thời đại số.

Bài báo này sẽ tập trung vào việc phân tích ứng dụng của AI trong học tập của sinh viên tại các trường đại học hiện nay, AI hỗ trợ cá nhân hóa việc học, cung cấp tài liệu thông minh, cải thiện kỹ năng tự học và quản lý thời gian. Ngoài ra, bài báo cũng xem xét các ứng dụng AI phổ biến như trợ lý ảo, hệ thống chấm điểm tự động, phần mềm kiểm tra đạo văn và các nền tảng học tập trực tuyến. Bên cạnh những lợi ích, nghiên cứu cũng chỉ ra những thách thức khi áp dụng AI vào giáo dục, bao gồm nguy cơ phụ thuộc quá mức vào công nghệ và vấn đề đạo đức trong học tập. Kết quả nghiên cứu khẳng định rằng nếu được sử dụng đúng cách, AI sẽ là một công cụ hỗ trợ mạnh mẽ, góp phần nâng cao chất lượng giáo dục và chuẩn bị cho sinh viên những kỹ năng cần thiết trong thời đại số.

2. CƠ SỞ LÝ THUYẾT

2.1. Các ứng dụng của trí tuệ nhân tạo (AI) trong học tập của sinh viên

2.1.1. Cá nhân hóa lộ trình học tập

Trải nghiệm học tập đáp ứng nhu cầu và sở thích riêng biệt của từng cá nhân cấu thành phương pháp học tập cá nhân hóa. Đây là phương pháp tiếp cận lấy người học làm trung tâm, cung cấp cho người học hành trình học tập tùy chỉnh và linh hoạt hơn cũng như cơ hội khám phá các chủ đề phù hợp với sở thích của mình. Người ta ước tính rằng đến năm 2024, thị trường học tập cá nhân hóa toàn cầu có thể vượt quá 2 tỷ đô la. Khi các tổ chức trên toàn thế giới nỗ lực để việc học tập trở nên dễ tiếp cận với nhân viên toàn cầu, việc học tập cá nhân hóa ngày càng được sử dụng để tạo ra các lộ trình học tập tùy chỉnh phù hợp với địa điểm, vai trò, mục tiêu nghề nghiệp và hiệu suất của nhân viên. Học tập cá nhân hóa bằng AI (Personalize Learning using AI) là một lĩnh vực đang phát triển. Nó sử dụng rộng rãi phân tích dữ liệu để theo dõi tiến trình của nhân viên, xác định các lĩnh vực cần cải thiện và đo lường tác động của các chương trình đào tạo. Điều này cho phép các tổ chức xác định các khoảng cách kỹ năng cụ thể trong lực lượng lao động và tùy chỉnh nội dung đào tạo để phù hợp với nhu cầu hoặc tiến trình của người học. Học tập cá nhân hóa bằng AI liên quan đến việc tận dụng công nghệ để tạo ra những trải nghiệm phù hợp và thích ứng, do đó làm cho việc học hiệu quả và hấp dẫn hơn.

Các bước để cá nhân hóa việc học bằng AI được mô tả dưới đây:

- Xác định mục tiêu và mục đích học tập rõ ràng.
- Xác định nhu cầu học tập của từng cá nhân.
- Chọn các công cụ hoặc nền tảng hỗ trợ AI để hỗ trợ việc học tập được cá nhân hóa.
- Phát triển hoặc biên tập nội dung có thể điều chỉnh dựa trên khả năng và tiến độ của người học.
- Sử dụng AI để tạo ra các bài đánh giá phù hợp với năng lực của người học.
- Cung cấp phản hồi tức thời bằng AI và đề xuất các nguồn lực bổ sung.
- Theo dõi tiến trình của người học và gợi ý thời điểm chuyển sang chủ đề tiếp theo hoặc thời điểm xem lại.
- Thúc đẩy việc học tập cộng tác thông qua các công cụ như diễn đàn thảo luận.
- Theo dõi hiệu suất và sự tham gia để xác định những lĩnh vực mà người học có thể cần can thiệp.
- Tìm kiếm phản hồi từ người học và người hướng dẫn để cải thiện phương pháp học tập cá nhân hóa dựa trên AI.
- Theo dõi các chỉ số đánh giá hiệu suất chính để đánh giá hiệu quả của việc học tập cá nhân hóa.

Lợi ích của việc sử dụng AI cho việc học tập được cá nhân hóa

Học tập cá nhân hóa bằng AI mang lại nhiều lợi ích cho các tổ chức. Một số lợi ích chính của việc sử dụng AI cho học tập cá nhân hóa bao gồm:

- Cải thiện kết quả học tập: Việc học được cá nhân hóa nhờ AI có thể giúp người học hiểu và ghi nhớ tốt hơn, từ đó cải thiện kết quả học tập.
- Khả năng mở rộng và khả năng tiếp cận toàn cầu: Sử dụng AI để cá nhân hóa việc học giúp dễ dàng mở rộng quy mô và cung cấp chương trình đào tạo dễ tiếp cận, nhất quán và chất lượng cao cho lực lượng lao động toàn cầu.
- Đào tạo tiết kiệm chi phí: AI giúp tự động hóa một số khía cạnh của quy trình đào tạo, chẳng hạn như đánh giá và phản hồi, giúp giảm thời gian, công sức và chi phí liên quan đến sự can thiệp của con người.
- Ra quyết định dựa trên dữ liệu: AI hỗ trợ ra quyết định sáng suốt về việc cải thiện chương trình đào tạo bằng cách phân tích dữ liệu về tiến độ học tập và sở thích của từng cá nhân.
- Phát triển nghề nghiệp tùy chỉnh: Học tập cá nhân hóa do AI thúc đẩy giúp các tổ chức bồi dưỡng các nhà lãnh đạo tương lai với các kỹ năng và kiến thức cần thiết cho các vai trò hiện tại và sự phát triển nghề nghiệp trong tương lai, từ đó có thể giảm tình trạng hao hụt nhân sự.
- Nâng cao khả năng tiếp cận và tính toàn diện: Bằng cách hỗ trợ người học khuyết tật và làm cho tài liệu đào tạo dễ tiếp cận và toàn diện hơn, AI giúp các tổ chức thể hiện cam kết đối với sự đa dạng và công bằng.
- Văn hóa học tập liên tục: AI thúc đẩy văn hóa học tập liên tục, trong đó nhân viên có thể nâng cao hoặc học lại kỹ năng để thích ứng với môi trường kinh doanh luôn thay đổi.
- Lợi thế cạnh tranh: Các tổ chức áp dụng AI cho việc học tập được cá nhân hóa có thể đạt được lợi thế cạnh tranh bằng cách thu hút nhân tài hàng đầu và cải thiện năng lực của lực lượng lao động.

2.1.2. Hỗ trợ tìm kiếm và tổng hợp thông tin

Trí tuệ nhân tạo (AI) đang mở ra nhiều cơ hội mới cho sinh viên trong việc tiếp cận tài liệu học tập một cách nhanh chóng và hiệu quả. Nhờ các công cụ AI như Google Scholar, ChatGPT, Perplexity AI, sinh viên có thể dễ dàng tìm kiếm nguồn tài liệu tham khảo, bài báo khoa học và nghiên cứu chuyên sâu trong nhiều lĩnh vực khác nhau. Google Scholar giúp sinh viên truy cập kho dữ liệu khổng lồ gồm các bài báo học thuật, luận án, sách và tài liệu nghiên cứu từ khắp nơi trên thế giới, giúp họ có cơ sở vững chắc để thực hiện các bài luận, đề tài nghiên cứu hoặc dự án học tập. Trong khi đó, ChatGPT đóng vai trò như một trợ lý thông minh, giúp sinh viên giải đáp thắc mắc, tóm tắt nội dung phức tạp và thậm chí hướng dẫn cách trình bày một vấn đề khoa học sao cho logic và dễ hiểu. Còn Perplexity AI lại có thể mạnh trong việc cung cấp thông tin nhanh chóng, chính xác và có dẫn nguồn rõ ràng, giúp sinh viên kiểm chứng độ tin cậy của tài liệu một cách dễ dàng. Nhờ vào những công cụ AI này, quá trình học tập trở nên linh hoạt hơn, giúp sinh viên tiết kiệm thời gian tìm kiếm thông tin và tập trung vào việc phân tích, nghiên cứu sâu hơn. Tuy nhiên, bên cạnh những lợi ích rõ rệt, sinh viên cũng cần có tư duy phản biện để đánh giá và chọn lọc thông tin từ AI, tránh việc tiếp nhận một cách thụ động mà không kiểm chứng độ chính xác và tính khách quan của nguồn tài liệu. Nếu biết sử dụng hợp lý, AI sẽ trở thành một công cụ hỗ trợ đắc lực, giúp nâng cao hiệu quả học tập và phát triển tư duy nghiên cứu trong môi trường đại học.

Hệ thống trí tuệ nhân tạo (AI) ngày nay có khả năng hỗ trợ sinh viên trong việc tóm tắt nội dung, dịch thuật và trích xuất thông tin quan trọng một cách nhanh chóng và chính xác. Nhờ vào các công cụ như ChatGPT, Google Translate hay QuillBot, sinh viên có thể rút gọn những tài liệu dài thành những ý chính dễ hiểu, giúp tiết kiệm thời gian đọc và nghiên cứu. Đặc biệt, AI còn hỗ trợ dịch thuật đa ngôn ngữ với độ chính xác ngày càng cao, giúp sinh viên tiếp cận các tài liệu nước ngoài mà không gặp rào cản về ngôn ngữ. Ngoài ra, các công cụ như Perplexity AI hay Semantic Scholar có thể tự động trích xuất thông tin quan trọng từ các bài

báo khoa học, báo cáo nghiên cứu, giúp sinh viên nhanh chóng xác định nội dung cốt lõi mà không cần đọc toàn bộ tài liệu. Những tính năng này không chỉ nâng cao hiệu suất học tập mà còn giúp sinh viên rèn luyện kỹ năng tổng hợp, phân tích thông tin một cách hiệu quả hơn. Tuy nhiên, để đảm bảo độ chính xác, sinh viên cần biết cách đánh giá và kiểm chứng thông tin do AI cung cấp, tránh phụ thuộc hoàn toàn vào công nghệ mà không phát triển tư duy phản biện của bản thân.

2.2. Cải thiện kỹ năng viết và kiểm tra nội dung học thuật

Trong môi trường học thuật hiện đại, các công cụ kiểm tra ngữ pháp và đạo văn như Grammarly, Turnitin đóng vai trò quan trọng trong việc hỗ trợ sinh viên nâng cao kỹ năng viết. Grammarly không chỉ giúp sinh viên phát hiện và sửa lỗi ngữ pháp, chính tả mà còn cung cấp gợi ý cải thiện câu văn, giúp bài viết trở nên mạch lạc, rõ ràng và chuyên nghiệp hơn. Công cụ này còn có tính năng kiểm tra giọng điệu và phong cách viết, hỗ trợ sinh viên điều chỉnh văn phong sao cho phù hợp với từng loại bài tập như luận văn, báo cáo nghiên cứu hay email học thuật. Trong khi đó, Turnitin là một nền tảng giúp phát hiện đạo văn bằng cách so sánh nội dung bài viết với hàng triệu tài liệu học thuật và cơ sở dữ liệu trực tuyến. Điều này giúp sinh viên tránh việc sao chép vô ý và khuyến khích họ phát triển tư duy sáng tạo, tự diễn đạt ý tưởng của mình thay vì sao chép từ các nguồn có sẵn. Ngoài ra, việc sử dụng các công cụ này còn giúp sinh viên rèn luyện kỹ năng tự chỉnh sửa bài viết, nâng cao khả năng diễn đạt và tư duy logic trong học tập và nghiên cứu. Tuy nhiên, mặc dù AI có thể hỗ trợ mạnh mẽ trong việc cải thiện kỹ năng viết, sinh viên vẫn cần kết hợp với việc trau dồi kiến thức về ngữ pháp, cấu trúc văn bản và tư duy phản biện để phát triển một phong cách viết rõ ràng, thuyết phục và mang dấu ấn cá nhân. Khi được sử dụng đúng cách, Grammarly và Turnitin không chỉ giúp sinh viên đạt kết quả học tập tốt hơn mà còn hỗ trợ họ xây dựng kỹ năng viết chuyên nghiệp, chuẩn bị tốt cho công việc tương lai.

Với thời đại công nghệ phát triển mạnh mẽ, trí tuệ nhân tạo (AI) đang trở thành một công cụ hỗ trợ đắc lực giúp sinh viên viết báo cáo và bài luận một cách logic, chính xác hơn. Các công cụ như ChatGPT, Grammarly, QuillBot hay Jasper AI có khả năng hỗ trợ sinh viên trong quá trình nghiên cứu, tổng hợp thông tin, đề xuất ý tưởng và cải thiện cấu trúc bài viết. Đặc biệt, AI giúp sinh viên sắp xếp luận điểm một cách hợp lý, đảm bảo tính mạch lạc và logic giữa các phần trong bài luận, từ mở bài, thân bài đến kết luận. Nhờ vào khả năng xử lý ngôn ngữ tự nhiên, AI có thể gợi ý cách diễn đạt rõ ràng, giúp sinh viên trình bày quan điểm một cách thuyết phục và chuyên nghiệp hơn. Ngoài ra, AI còn giúp kiểm tra ngữ pháp, chính tả và phong cách viết, giúp bài báo cáo đạt độ chuẩn mực cao hơn về hình thức và nội dung. Không chỉ vậy, các công cụ như Turnitin còn giúp kiểm tra đạo văn, đảm bảo tính trung thực trong học thuật và giúp sinh viên học cách trích dẫn nguồn tài liệu chính xác. Tuy nhiên, mặc dù AI mang lại nhiều lợi ích, sinh viên không nên quá phụ thuộc vào công nghệ mà cần phát triển khả năng tư duy độc lập, sáng tạo và tự viết bài dựa trên sự hiểu biết của bản thân. AI nên được sử dụng như một công cụ hỗ trợ thay vì thay thế hoàn toàn quá trình viết và nghiên cứu. Nếu biết tận dụng đúng cách, AI sẽ giúp sinh viên nâng cao kỹ năng viết học thuật, từ đó cải thiện kết quả học tập và phát triển tư duy phân tích trong môi trường đại học.

2.3. Hỗ trợ giảng dạy ứng dụng AI trong đánh giá và kiểm tra kiến thức

Môi trường giáo dục hiện đại, hệ thống chấm điểm tự động đang trở thành một công cụ quan trọng giúp tiết kiệm thời gian và đảm bảo tính khách quan trong đánh giá kết quả học tập của sinh viên. Các hệ thống này, được phát triển dựa trên trí tuệ nhân tạo (AI) và học máy (machine learning), có khả năng phân tích bài làm của sinh viên, nhận diện nội dung, đánh giá cấu trúc câu trả lời và đưa ra điểm số một cách nhanh chóng, chính xác. Đối với các bài thi trắc nghiệm, hệ thống AI có thể tự động chấm điểm ngay lập tức, giúp giảng viên giảm bớt

khối lượng công việc thủ công và đảm bảo rằng tất cả sinh viên được đánh giá theo cùng một tiêu chuẩn. Đặc biệt, trong các bài luận hoặc bài viết học thuật, AI có thể phân tích cách lập luận, sử dụng từ vựng và tính logic của bài viết để đưa ra nhận xét và điểm số phù hợp. Một số hệ thống chấm điểm tiên tiến như Gradescope hay Turnitin không chỉ đánh giá dựa trên các tiêu chí đã được lập trình sẵn mà còn học hỏi từ cách chấm điểm của giảng viên để cải thiện độ chính xác theo thời gian. Nhờ đó, quá trình chấm điểm trở nên nhanh chóng, minh bạch, hạn chế tối đa sai sót do yếu tố chủ quan hoặc cảm tính của con người. Ngoài ra, hệ thống AI còn cung cấp phản hồi chi tiết, giúp sinh viên hiểu rõ điểm mạnh, điểm yếu của bài làm và có cơ hội cải thiện kỹ năng học tập của mình. Tuy nhiên, dù mang lại nhiều lợi ích, chấm điểm tự động vẫn có một số hạn chế, đặc biệt là trong việc đánh giá các yếu tố sáng tạo và lập luận sâu sắc của sinh viên, vốn đòi hỏi sự phân tích linh hoạt của con người. Vì vậy, sự kết hợp giữa AI và đánh giá từ giảng viên sẽ là giải pháp tối ưu, vừa tận dụng được công nghệ để nâng cao hiệu quả chấm điểm, vừa đảm bảo đánh giá chính xác và công bằng trong môi trường giáo dục.

Trong bối cảnh giáo dục ngày càng số hóa, các bài kiểm tra trực tuyến sử dụng trí tuệ nhân tạo (AI) đang trở thành một phương thức đánh giá năng lực sinh viên hiệu quả, hiện đại và minh bạch hơn. Nhờ vào các công nghệ tiên tiến như học máy (machine learning), xử lý ngôn ngữ tự nhiên (NLP) và nhận diện hình ảnh, hệ thống AI có thể tự động chấm điểm, theo dõi quá trình làm bài và đưa ra những đánh giá khách quan về năng lực của sinh viên. Một trong những lợi ích lớn nhất của AI trong kiểm tra trực tuyến là khả năng cá nhân hóa bài thi, điều chỉnh độ khó phù hợp với năng lực của từng sinh viên, từ đó đánh giá chính xác hơn mức độ hiểu biết và kỹ năng của họ. Ngoài ra, AI cũng giúp giám sát kỳ thi một cách nghiêm ngặt thông qua các công cụ như nhận diện khuôn mặt, theo dõi chuyển động mắt và phát hiện hành vi gian lận, giúp đảm bảo tính trung thực và công bằng trong kiểm tra trực tuyến. Các hệ thống như ProctorU, Examity hay Honorlock có thể phát hiện hành vi bất thường, cảnh báo giảng viên khi sinh viên có dấu hiệu gian lận, đồng thời ghi lại toàn bộ quá trình thi để phục vụ công tác kiểm tra lại nếu cần thiết. Hơn nữa, các bài kiểm tra có thể được thiết kế dưới nhiều dạng khác nhau, từ trắc nghiệm tự động chấm điểm đến bài luận mà AI có thể phân tích mức độ logic và cách diễn đạt để đưa ra phản hồi tức thì. Điều này giúp giảm bớt gánh nặng chấm điểm thủ công cho giảng viên, đồng thời giúp sinh viên nhận phản hồi nhanh hơn để cải thiện năng lực học tập. Tuy nhiên, mặc dù AI mang lại nhiều lợi ích, việc áp dụng kiểm tra trực tuyến vẫn cần sự giám sát của con người để đảm bảo hệ thống hoạt động chính xác và tránh những sai sót không mong muốn. Việc kết hợp giữa công nghệ AI và đánh giá từ giảng viên sẽ giúp quá trình kiểm tra trở nên hiệu quả, công bằng và chính xác hơn, góp phần nâng cao chất lượng giáo dục trong thời đại số hóa.

3. LỢI ÍCH VÀ THÁCH THỨC CỦA AI TRONG HỌC TẬP

3.1. Lợi ích

AI hợp lý quản lý học tập (LMS) hệ thống nên được sử dụng để tạo ra chương trình học riêng biệt dựa trên dữ liệu về chế độ, cơ sở thích và phong cách học tập của mỗi sinh viên. Ví dụ, tại Đại học Công thương TP.HCM, một hệ thống thử nghiệm AI có thể phân tích kết quả bài kiểm tra định kỳ để đưa ra các bài tập nâng cao cho sinh viên giỏi hoặc tài liệu cơ bản cho những người cần cải thiện. Hệ thống này có thể dựa trên các thuật toán Machine Learning để dự đoán nhu cầu học tập, đưa ra ý kiến tài liệu về trình nâng cao lập trình cho sinh viên chuyên ngành công nghệ thông tin nếu chúng thành thạo các khái niệm cơ bản. Theo một nghiên cứu của McKinsey (2022), các ứng dụng đại học của trường đại học trình bày cá nhân hóa bằng AI ghi nhận mức độ hài lòng của người dùng sinh viên tăng 30% và tỷ lệ hoàn thành khóa học tăng 15%.

Trong bối cảnh công nghệ phát triển mạnh mẽ, trí tuệ nhân tạo (AI) đang đóng vai trò quan trọng trong việc tăng cường hiệu quả học tập và giúp sinh viên tiết kiệm thời gian. Trước đây, sinh viên thường mất rất nhiều thời gian để tìm kiếm tài liệu, tổng hợp thông tin và tự ôn tập kiến thức, nhưng với sự hỗ trợ của AI, quá trình học tập trở nên nhanh chóng, tiện lợi và hiệu quả hơn bao giờ hết. Các công cụ như ChatGPT, Google Scholar hay Perplexity AI giúp sinh viên tiếp cận nguồn tài liệu một cách nhanh chóng, tóm tắt nội dung chính và trích xuất thông tin quan trọng chỉ trong vài giây. Nhờ đó, sinh viên có thể tập trung hơn vào việc phân tích và hiểu sâu vấn đề thay vì mất thời gian vào những công việc tìm kiếm thủ công. Ngoài ra, AI còn giúp cá nhân hóa quá trình học tập thông qua các nền tảng như Coursera, Khan Academy hay Duolingo, tự động điều chỉnh lộ trình học tập dựa trên trình độ và tiến độ của từng người. Những ứng dụng này có thể đề xuất nội dung phù hợp, gợi ý bài tập và đưa ra phản hồi ngay lập tức, giúp sinh viên cải thiện kỹ năng một cách hiệu quả hơn. Không chỉ dừng lại ở việc hỗ trợ học tập, AI còn giúp sinh viên quản lý thời gian hợp lý thông qua các ứng dụng như Notion, Trello hay Google Calendar, giúp họ lập kế hoạch học tập khoa học, nhắc nhở lịch trình và tối ưu hóa quỹ thời gian. Việc sử dụng AI trong học tập không chỉ giúp sinh viên tiết kiệm thời gian mà còn nâng cao khả năng tư duy, sáng tạo và tiếp thu kiến thức một cách có hệ thống hơn. Tuy nhiên, để đạt được hiệu quả tối ưu, sinh viên cần biết cách sử dụng AI một cách thông minh, kết hợp với tư duy độc lập và kỹ năng tự học, thay vì phụ thuộc hoàn toàn vào công nghệ. Khi được áp dụng đúng cách, AI sẽ trở thành công cụ hỗ trợ đắc lực, giúp sinh viên nâng cao hiệu suất học tập, tối ưu hóa thời gian và chuẩn bị tốt hơn cho sự nghiệp tương lai.

Trước đây, việc học tập bị giới hạn bởi tài liệu in ấn, thư viện truyền thống và khả năng tiếp cận nguồn thông tin chất lượng. Tuy nhiên, với sự phát triển của AI, sinh viên có thể dễ dàng truy cập kho kiến thức khổng lồ chỉ với vài cú nhấp chuột. Các công cụ như Google Scholar, ChatGPT, Perplexity AI hay Semantic Scholar giúp sinh viên tìm kiếm thông tin một cách nhanh chóng, tổng hợp nội dung từ nhiều nguồn và tiếp cận những nghiên cứu khoa học mới nhất trên toàn cầu. Điều này không chỉ giúp tiết kiệm thời gian mà còn giúp sinh viên có cái nhìn toàn diện và sâu sắc hơn về vấn đề mình đang nghiên cứu. Ngoài ra, AI còn giúp phá vỡ rào cản ngôn ngữ bằng các công cụ dịch thuật như Google Translate hay DeepL, giúp sinh viên tiếp cận tài liệu quốc tế mà không bị giới hạn bởi khả năng ngoại ngữ.

Không chỉ dừng lại ở việc cung cấp thông tin, AI còn đóng vai trò quan trọng trong việc nâng cao tư duy sáng tạo. Các công cụ hỗ trợ viết như Grammarly, QuillBot hay Jasper AI giúp sinh viên phát triển kỹ năng diễn đạt, cải thiện cấu trúc bài viết và khám phá nhiều cách thể hiện ý tưởng mới mẻ. AI cũng khuyến khích sinh viên tư duy linh hoạt hơn thông qua các nền tảng như DALL·E, Midjourney—giúp họ tạo ra nội dung hình ảnh độc đáo, hay các công cụ lập trình AI như GitHub Copilot—hỗ trợ sinh viên sáng tạo các sản phẩm phần mềm một cách nhanh chóng. Hơn nữa, AI còn giúp sinh viên thử nghiệm và kiểm tra ý tưởng trong nhiều lĩnh vực khác nhau, từ thiết kế, kỹ thuật đến kinh doanh, nhờ vào khả năng mô phỏng và phân tích dữ liệu nhanh chóng. Điều này không chỉ giúp sinh viên nâng cao kỹ năng giải quyết vấn đề mà còn kích thích khả năng sáng tạo, thúc đẩy tư duy đổi mới và phát triển những ý tưởng đột phá.

Tuy nhiên, để thực sự tận dụng được tiềm năng của AI trong việc mở rộng tri thức và phát triển tư duy sáng tạo, sinh viên cần chủ động trong quá trình học tập, biết cách sàng lọc và đánh giá thông tin một cách có chọn lọc. AI chỉ là công cụ hỗ trợ, còn sự sáng tạo và khả năng tư duy phản biện vẫn phụ thuộc vào con người. Khi kết hợp một cách hợp lý giữa công nghệ và trí tuệ cá nhân, sinh viên không chỉ nâng cao hiệu suất học tập mà còn phát triển một cách toàn diện, chuẩn bị tốt hơn cho những thách thức trong tương lai.

3.2. Thách thức

Mặc dù trí tuệ nhân tạo (AI) mang lại nhiều lợi ích trong học tập, sinh viên vẫn phải đối mặt với không ít thách thức khi sử dụng công nghệ này. Một trong những vấn đề lớn nhất là sự phụ thuộc quá mức vào AI, khiến sinh viên có nguy cơ mất dần khả năng tư duy độc lập và sáng tạo. Khi AI có thể nhanh chóng tạo ra bài luận, tóm tắt tài liệu hay giải quyết bài tập, nhiều sinh viên có xu hướng sử dụng AI một cách thụ động, thay vì tự nghiên cứu và rèn luyện kỹ năng tư duy phản biện. Điều này có thể dẫn đến sự suy giảm trong chất lượng học tập và khả năng giải quyết vấn đề trong thực tế.

Bên cạnh đó, độ chính xác của AI không phải lúc nào cũng được đảm bảo. Các mô hình AI như ChatGPT hay Perplexity AI có thể cung cấp thông tin sai lệch hoặc chưa được kiểm chứng, khiến sinh viên dễ bị nhầm lẫn nếu không biết cách kiểm tra lại nguồn tin. Việc lạm dụng AI mà không có sự chọn lọc có thể dẫn đến tình trạng trích dẫn sai, làm giảm tính chính xác và độ tin cậy trong bài làm học thuật. Ngoài ra, nhiều công cụ AI chưa thể đánh giá chính xác các yếu tố sáng tạo, lập luận phức tạp hoặc cảm xúc trong bài viết, khiến sinh viên phải cân nhắc kỹ lưỡng khi sử dụng AI trong các bài tập mang tính tư duy sâu.

Một thách thức khác là vấn đề đạo văn và vi phạm quy tắc học thuật. Khi AI có thể tạo ra nội dung một cách dễ dàng, sinh viên có thể vô tình hoặc cố ý sử dụng nó mà không trích dẫn nguồn gốc, gây ra vi phạm đạo đức học thuật. Các hệ thống kiểm tra đạo văn như Turnitin ngày càng tinh vi hơn trong việc phát hiện nội dung do AI tạo ra, điều này đặt ra yêu cầu cho sinh viên phải biết cách sử dụng AI một cách hợp lý và kết hợp với khả năng tự viết để tránh bị đánh giá thấp.

Cuối cùng, không phải tất cả sinh viên đều có đủ kỹ năng để sử dụng AI một cách hiệu quả. Một số người gặp khó khăn trong việc khai thác đúng cách các công cụ AI, dẫn đến việc sử dụng sai mục đích hoặc không tận dụng được hết tiềm năng của công nghệ này. Ngoài ra, một số phần mềm AI yêu cầu trả phí, tạo ra sự bất bình đẳng trong khả năng tiếp cận công nghệ giữa các sinh viên.

Nhìn chung, AI là một công cụ hữu ích nhưng cũng đi kèm với nhiều thách thức mà sinh viên cần vượt qua. Để tận dụng tối đa lợi ích của AI trong học tập, sinh viên cần có kỹ năng tư duy phản biện, khả năng chọn lọc thông tin và đạo đức học thuật vững vàng. Khi biết cách sử dụng AI một cách thông minh, sinh viên có thể biến nó thành một trợ thủ đắc lực thay vì để nó trở thành rào cản trong quá trình học tập và phát triển của mình.

4. GIẢI PHÁP VÀ ĐỀ XUẤT

4.1. Sử dụng AI một cách có kiểm soát và cân bằng

Việc sử dụng trí tuệ nhân tạo (AI) trong học tập mang lại nhiều lợi ích, nhưng sinh viên cần biết cách sử dụng một cách có kiểm soát và cân bằng để đạt hiệu quả cao nhất. AI có thể hỗ trợ tìm kiếm thông tin nhanh chóng, tóm tắt nội dung, kiểm tra ngữ pháp và đề xuất ý tưởng, nhưng nếu quá phụ thuộc vào nó, sinh viên có thể dần mất đi khả năng tư duy độc lập và sáng tạo. Vì vậy, thay vì để AI làm thay mọi công việc, sinh viên nên xem AI như một công cụ hỗ trợ, giúp tối ưu hóa quá trình học tập mà vẫn giữ được sự chủ động trong việc tiếp thu và xử lý thông tin.

Bên cạnh đó, sinh viên cần biết cách kiểm chứng độ chính xác của thông tin do AI cung cấp, vì không phải lúc nào AI cũng đưa ra câu trả lời đúng hoặc phù hợp với ngữ cảnh học thuật. Việc đối chiếu với các nguồn tài liệu uy tín và tư duy phản biện khi sử dụng AI là điều cần thiết để đảm bảo tính chính xác và đáng tin cậy của kiến thức. Ngoài ra, sinh viên cũng cần tuân thủ đạo đức học thuật, tránh sao chép nguyên văn nội dung do AI tạo ra mà không qua quá trình phân tích và biên tập lại theo cách hiểu của bản thân.

Cuối cùng, sử dụng AI một cách cân bằng không chỉ giúp sinh viên tiết kiệm thời gian mà còn rèn luyện kỹ năng tự học, nâng cao tư duy phản biện và phát triển khả năng sáng tạo. Khi biết cách kết hợp AI với phương pháp học truyền thống, sinh viên không chỉ tận dụng được lợi ích của công nghệ mà còn đảm bảo quá trình học tập vẫn mang tính chủ động và hiệu quả lâu dài.

4.2. Đảm bảo tính minh bạch và đạo đức khi sử dụng AI

Việc đảm bảo tính minh bạch và đạo đức khi sử dụng AI trở thành một vấn đề quan trọng đối với sinh viên. AI mang lại nhiều tiện ích trong học tập, từ việc tìm kiếm thông tin nhanh chóng, hỗ trợ viết bài, đến kiểm tra ngữ pháp và đạo văn. Tuy nhiên, nếu không sử dụng đúng cách, AI có thể dẫn đến những vi phạm đạo đức học thuật, làm suy giảm tính trung thực và trách nhiệm cá nhân trong quá trình học tập. Một trong những nguy cơ lớn nhất là việc sinh viên sử dụng AI để tạo ra nội dung một cách máy móc mà không có sự phân tích hay kiểm chứng lại, dẫn đến tình trạng sao chép và vi phạm đạo văn. Các công cụ như ChatGPT hay Jasper AI có thể tạo ra các bài luận hoàn chỉnh chỉ trong vài giây, nhưng việc sử dụng nội dung này mà không chỉnh sửa hay bổ sung ý tưởng cá nhân có thể làm mất đi giá trị học thuật và sự sáng tạo của sinh viên.

Để đảm bảo tính minh bạch, sinh viên cần phải sử dụng AI như một công cụ hỗ trợ thay vì phụ thuộc hoàn toàn vào nó. Khi sử dụng nội dung do AI tạo ra, sinh viên nên trích dẫn rõ ràng nếu cần thiết, đồng thời kết hợp với nghiên cứu cá nhân để đảm bảo bài viết có tính độc lập và thể hiện sự hiểu biết riêng. Ngoài ra, việc sử dụng AI để tìm kiếm tài liệu cũng cần có sự chọn lọc, tránh tình trạng lấy thông tin từ các nguồn không đáng tin cậy hoặc sử dụng dữ liệu không chính xác. Một số hệ thống AI có thể cung cấp thông tin sai lệch hoặc chưa được kiểm chứng, do đó sinh viên cần phải kiểm tra lại bằng cách đối chiếu với các tài liệu học thuật chính thống.

Bên cạnh đó, đạo đức trong việc sử dụng AI còn liên quan đến vấn đề công bằng và trách nhiệm cá nhân. Một số sinh viên có thể lạm dụng AI để gian lận trong các kỳ thi trực tuyến hoặc nhờ AI làm hộ bài tập mà không thực sự tham gia vào quá trình học tập. Điều này không chỉ ảnh hưởng đến sự phát triển cá nhân mà còn làm mất đi ý nghĩa thực sự của việc giáo dục. Các trường đại học hiện nay ngày càng siết chặt các quy định về việc sử dụng AI trong học tập, nhiều hệ thống kiểm tra đạo văn như Turnitin hay GPTZero có khả năng phát hiện nội dung do AI tạo ra. Vì vậy, sinh viên cần có trách nhiệm với chính bản thân mình, sử dụng AI một cách minh bạch, đúng mục đích và tuân thủ các nguyên tắc học thuật.

5. KẾT LUẬN VÀ KIẾN NGHỊ

Hiện nay, trí tuệ nhân tạo (AI) đang dần trở thành một phần không thể thiếu trong quá trình học tập của sinh viên tại các trường đại học. Với sự phát triển mạnh mẽ của công nghệ, sinh viên có thể sử dụng AI để hỗ trợ việc tìm kiếm tài liệu, tóm tắt nội dung, dịch thuật, thậm chí là viết bài luận hay giải các bài toán phức tạp. Các nền tảng như ChatGPT, Google Bard hay Perplexity AI giúp sinh viên tiếp cận tri thức một cách nhanh chóng và hiệu quả hơn bao giờ hết. Ngoài ra, nhiều ứng dụng AI còn giúp cá nhân hóa lộ trình học tập, đưa ra gợi ý phù hợp với năng lực của từng sinh viên, giúp họ quản lý thời gian học tập một cách tối ưu. Đặc biệt, trong các ngành học liên quan đến lập trình, khoa học dữ liệu hay trí tuệ nhân tạo, sinh viên còn có cơ hội thực hành với các mô hình AI thực tế, qua đó nâng cao kỹ năng chuyên môn và chuẩn bị tốt hơn cho công việc tương lai. Tuy nhiên, bên cạnh những lợi ích rõ rệt, việc lạm dụng AI cũng đặt ra nhiều vấn đề đáng lo ngại. Một số sinh viên có xu hướng phụ thuộc quá mức vào công cụ này, làm giảm khả năng tư duy độc lập và sáng tạo. Hơn nữa, vấn đề đạo văn và gian lận trong học tập cũng trở nên phổ biến hơn khi AI có thể tạo ra nội dung một cách dễ dàng. Do đó, việc sử dụng AI trong học tập cần có sự cân bằng và định hướng

hợp lý để vừa tận dụng được lợi ích của công nghệ, vừa đảm bảo sinh viên phát triển kỹ năng tư duy và sáng tạo một cách bền vững.

TÀI LIỆU THAM KHẢO

1. Dileep Kumar Gupta, Shivendu Prashar, Sartajvir Singh, Prashant K. Srivastava, Rajendra Prasad (2022), Chapter 1 - Introduction to RADAR remote sensing, In Earth Observation, Radar Remote Sensing, Elsevier, Pages 3-27.
2. Karim, S., Zhang, Y., Yin (2019), S. et al. Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. *Multimed Tools Appl* 78, pp. 32565–32583.
3. Mansmann, S., Neumuth, T., Scholl, M.H. (2007). Multidimensional Data Modeling for Business Process Analysis. In: Parent, C., Schewe, KD., Storey, V.C., Thalheim, B. (eds) *Conceptual Modeling - ER 2007. ER 2007. Lecture Notes in Computer Science*, vol 4801. Springer, Berlin, Heidelberg.
4. Rasmus Bro and Age K. Smilde (2014), “Principal component analysis, The Royal Society of Chemistry”
5. S. A. Yamashkin, A. A. Yamashkin, V. V. Zanozin, M. M. Radovanovic and A. N. Barmin (2020), "Improving the Efficiency of Deep Learning Methods in Remote Sensing Data Analysis: Geosystem Approach," in *IEEE Access*, vol. 8, pp. 179516-179529.
6. Sisodiya, N., Dube, N., Thakkar (2020), “P. Next-Generation Artificial Intelligence Techniques for Satellite Data Processing. In: Hemanth, D. (eds) *Artificial Intelligence Techniques for Satellite Image Analysis. Remote Sensing and Digital Image Processing*”, vol 24. Springer.
7. T. B. Pedersen and C. S. Jensen (1999), "Multidimensional data modeling for complex data," *Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)*, Sydney, NSW, Australia, pp. 336-345.
8. Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, Wei Jie (2015), “Remote sensing big data computing: Challenges and opportunities”, *Future Generation Computer Systems* 51, Pages 47-60.
9. Nguyễn, DA (2021). Ứng dụng trí tuệ nhân tạo trong dạy học ở trường đại học tại Việt Nam. Trong *Kỷ yếu hội thảo “Chuyển đổi số: Cách tiếp cận người học làm trung tâm ứng dụng trong phương pháp dạy học kết hợp”* (trang 484-492).
10. Nguyễn, TP (2023). Sử dụng công cụ hỗ trợ ChatGPT trong giảng dạy và học tập chuyên ngành truyền thông. *Tạp chí Truyền thông*, 15 (3), 23-30.
11. Zhiming, Y., Xiaxia, T., Xuan, Q., Fei, Z., & Yuanmei, D. (2020). Ý nghĩa hàm ản, công nghệ chính và xu hướng ứng dụng của trí tuệ nhân tạo giáo dục (EAI). *Tạp chí Giáo dục từ xa*, 38 (1), 26-35.

ABSTRACT

APPLICATION OF ARTIFICIAL INTELLIGENCE IN STUDENT LEARNING AT CURRENT UNIVERSITIES

Nguyễn Thị Thu Tâm^{1,*}, Lê Thị Thùy Lan¹

¹*Ho Chi Minh City University of Industry and Trade*

Email: tamntt@huit.edu.vn

In the context of strong technological development, artificial intelligence (AI) is gradually becoming a powerful support tool in students' learning at universities. AI helps personalize the learning process through online learning platforms, suggesting suitable documents and providing optimal learning paths. Virtual assistants such as ChatGPT, Google Bard or automated support systems help students search for information, summarize content and answer questions quickly. In addition, AI is also applied in automatic grading, plagiarism detection, support for foreign language learning and programming practice. Thanks to the development of AI, students can access a rich source of knowledge, improve learning performance and practice skills more flexibly and effectively.

Keywords: Applications of AI in learning, Artificial Intelligence in education today, AI chatbot, arnessing the power of AI in learning.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thị Thu Tâm

Số điện thoại liên hệ: 0909927572

PHƯƠNG PHÁP LỌC CONFIGURATION FILTERING HỖ TRỢ GỢI Ý VIDEO TRÊN NỀN TẢNG GIẢI TRÍ HTB

Huỳnh Công Huy¹, Nguyễn Quốc Tiến¹, Nguyễn Quốc Bảo¹, Nguyễn Hải Yến^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: yennh@huit.edu.vn

Ngày nhận bài: 22/03/2025; Ngày chấp nhận đăng: 20/05/2025.

TÓM TẮT

Trí tuệ nhân tạo (Artificial Intelligence – AI) đóng vai trò quan trọng trong công nghệ hiện đại, đặc biệt trong việc tối ưu hóa trải nghiệm người dùng trên các nền tảng giải trí trực tuyến. Hệ thống gợi ý video giúp cá nhân hóa nội dung, nâng cao mức độ tương tác và giữ chân người dùng thông qua việc gợi ý các video dựa vào lịch sử truy cập hệ thống của người dùng và xu hướng giải trí hiện tại. Trong bài báo này, nhóm tác giả tập trung nghiên cứu phương pháp lọc Configuration Filtering, bao gồm Collaborative Filtering và Content-based Filtering, để hỗ trợ đề xuất video trên nền tảng HTB Entertainment. Các nội dung chính của bài báo gồm: (1) Giới thiệu về hệ thống khuyến nghị; (2) Phương pháp lọc Configuration Filtering trong gợi ý video; (3) Mô hình hệ thống gợi ý và ứng dụng thực nghiệm; (4) Kết luận - Hướng phát triển.

Từ khóa: AI, Recommendation systems, Collaborative Filtering, Content-based Filtering, Video Recommendation.

1. GIỚI THIỆU HỆ THỐNG KHUYẾN NGHỊ

Hệ thống gợi ý (Recommendation Systems) là một trong những ứng dụng quan trọng của trí tuệ nhân tạo, đóng vai trò cốt lõi trong việc cải thiện trải nghiệm người dùng trên các nền tảng giải trí và mạng xã hội [1]. Với sự phát triển mạnh mẽ của công nghệ, đặc biệt là trong lĩnh vực học máy (Machine Learning) và học sâu (Deep Learning), các hệ thống gợi ý ngày càng trở nên thông minh hơn, có khả năng hiểu rõ hơn về sở thích, thói quen của người dùng và đưa ra những nội dung phù hợp nhất [2]. Hệ thống gợi ý hoạt động bằng cách thu thập và phân tích dữ liệu người dùng, bao gồm các hành vi tương tác như lịch sử xem, lượt thích, bình luận, chia sẻ hoặc thậm chí là thời gian xem của từng nội dung [3]. Từ đó, thuật toán sẽ dự đoán và đề xuất những nội dung có khả năng thu hút sự quan tâm của người dùng cao nhất. Trong nhiều năm qua, các nhà nghiên cứu đã không ngừng tìm kiếm và cải tiến các phương pháp để tối ưu hóa khả năng gợi ý, giúp hệ thống trở nên chính xác hơn, nhanh hơn và linh hoạt hơn trong việc cập nhật sở thích thay đổi theo thời gian [4]. Một số phương pháp phổ biến có thể kể đến như bộ lọc cộng tác (Collaborative Filtering), lọc nội dung (Content-Based Filtering), hệ thống gợi ý lai (Hybrid Recommendation Systems) kết hợp nhiều phương pháp

để nâng cao hiệu quả [5]. Ngoài ra, các mô hình học sâu như mạng nơ-ron tích chập (CNNs), mạng nơ-ron hồi tiếp (RNNs) và các mô hình Transformer cũng đang được ứng dụng rộng rãi để cải thiện chất lượng dự đoán của hệ thống gợi ý [6]. Một trong những nghiên cứu quan trọng là khả năng cá nhân hóa đề xuất theo từng người dùng, giúp hệ thống không chỉ dựa vào dữ liệu tổng hợp mà còn phân tích theo từng cá nhân để tối ưu trải nghiệm [7]. Hơn nữa, với sự xuất hiện của AI, hệ thống gợi ý giờ đây không chỉ gợi ý dựa trên dữ liệu lịch sử mà còn có thể dự đoán sở thích trong tương lai, thậm chí điều chỉnh gợi ý theo thời gian thực dựa trên ngữ cảnh hiện tại [8]. Tuy nhiên, bên cạnh những lợi ích to lớn, hệ thống gợi ý cũng đối mặt với nhiều thách thức lớn, chẳng hạn như vấn đề "lọc bong bóng" (Filter Bubble), khi người dùng chỉ nhận được những nội dung tương tự với sở thích hiện tại, hạn chế khả năng khám phá nội dung mới [9]. Ngoài ra, vấn đề bảo mật dữ liệu và quyền riêng tư cũng là một thách thức khi các hệ thống gợi ý cần thu thập lượng lớn thông tin cá nhân để hoạt động hiệu quả [10]. Việc xây dựng một hệ thống gợi ý minh bạch, công bằng và không thiên vị cũng là một trong những mục tiêu quan trọng mà các nhà nghiên cứu đang hướng tới [11]. Trong tương lai, với sự phát triển không ngừng của công nghệ, hệ thống gợi ý sẽ tiếp tục được cải tiến để trở nên thông minh hơn, chính xác hơn và mang đến trải nghiệm tốt nhất cho người dùng trong các nền tảng giải trí và mạng xã hội.

Hệ thống gợi ý dựa trên bộ lọc cộng tác (Collaborative Filtering)

Bộ lọc cộng tác là một trong những phương pháp phổ biến nhất trong hệ thống gợi ý. Cách tiếp cận này dựa vào dữ liệu từ nhiều người dùng để tìm ra những điểm tương đồng trong hành vi của họ, từ đó đưa ra các gợi ý phù hợp [1]. Có hai dạng chính: Collaborative Filtering theo người dùng (User-Based): Hệ thống sẽ tìm ra những người dùng có hành vi tương tự nhau và gợi ý nội dung mà họ đã thích nhưng người dùng hiện tại chưa tiếp cận [5]. Collaborative Filtering theo sản phẩm (Item-Based): Hệ thống tìm kiếm những nội dung có sự tương đồng về đặc điểm hoặc đã từng được xem cùng nhau để đưa ra gợi ý [12].

Hệ thống gợi ý dựa trên nội dung (Content-Based Filtering)

Phương pháp này dựa vào đặc điểm của nội dung thay vì dữ liệu từ người dùng khác. Hệ thống sẽ phân tích các thuộc tính của nội dung như thể loại, từ khóa, nhãn dán (tags) hoặc mô tả để tìm ra những nội dung tương tự với sở thích của người dùng [13]. Điểm mạnh của phương pháp này là không cần dữ liệu từ nhiều người dùng khác nhau, nhưng nó có thể gặp vấn đề khi không có đủ thông tin về nội dung hoặc sở thích người dùng chưa rõ ràng [14].

Hệ thống gợi ý lai (Hybrid Recommendation Systems)

Hệ thống gợi ý lai kết hợp nhiều phương pháp khác nhau, thường là sự pha trộn giữa Collaborative Filtering và Content-Based Filtering để tận dụng ưu điểm của cả hai [11]. Ví dụ, Netflix sử dụng một hệ thống lai, trong đó họ vừa phân tích hành vi của người dùng vừa xem xét các đặc điểm của bộ video để đưa ra các đề xuất chính xác hơn [8].

Ứng dụng mô hình học sâu (Deep Learning) trong hệ thống gợi ý

Với sự phát triển của Deep Learning, các hệ thống gợi ý ngày càng thông minh và có khả năng xử lý lượng dữ liệu lớn [6]. Một số mô hình phổ biến như: Neural Collaborative Filtering

(NCF): Sử dụng mạng nơ-ron nhân tạo để cải thiện độ chính xác của bộ lọc cộng tác [4]; Mô hình Recurrent Neural Networks (RNNs): Áp dụng trong việc phân tích chuỗi thời gian, giúp dự đoán xu hướng thay đổi sở thích của người dùng theo thời gian [15]; Mô hình Transformer-based Models: Những mô hình như BERT hoặc GPT cũng đang được ứng dụng để phân tích ngữ cảnh và sở thích người dùng trong các hệ thống gợi ý [16].

Cá nhân hóa trải nghiệm người dùng bằng AI

Cá nhân hóa là xu hướng quan trọng trong các hệ thống gợi ý hiện đại. AI không chỉ đề xuất nội dung dựa trên lịch sử xem mà còn có thể dự đoán sở thích trong tương lai dựa trên các yếu tố như thời gian, địa điểm, và trạng thái cảm xúc của người dùng [17]. Các nền tảng như YouTube và Spotify sử dụng AI để liên tục cập nhật danh sách gợi ý theo thời gian thực, giúp người dùng có trải nghiệm tốt hơn [8].

Thách thức và tương lai của hệ thống gợi ý

Dù các hệ thống gợi ý đã có nhiều tiến bộ đáng kể, vẫn còn một số thách thức cần giải quyết: Vấn đề "lọc bong bóng" : Khi hệ thống chỉ đề xuất nội dung dựa trên sở thích trước đó, người dùng có thể bị giới hạn trong một phạm vi nội dung nhất định và không khám phá được những nội dung đa dạng hơn [9]. Dữ liệu cá nhân và quyền riêng tư: Các nền tảng phải đối mặt với vấn đề bảo vệ dữ liệu người dùng khi thu thập thông tin để cải thiện gợi ý [18]. Tính minh bạch của thuật toán: Người dùng thường không biết lý do tại sao một nội dung lại xuất hiện trong danh sách gợi ý, điều này có thể gây tranh cãi về độ tin cậy của hệ thống [19].

2. PHƯƠNG PHÁP LỌC CONFIGURATION FILTERING TRONG GỢI Ý VIDEO

Collaborative Filtering (CF) và Content-Based Filtering (CBF) là hai phương pháp phổ biến được sử dụng trong hệ thống gợi ý video nhằm tối ưu hóa trải nghiệm người dùng [1]. Collaborative Filtering là phương pháp gợi ý dựa trên dữ liệu thu thập từ nhiều người dùng, phân tích hành vi của họ để tìm ra sự tương đồng giữa các người dùng khác nhau hoặc giữa các nội dung mà họ tương tác [12]. Trong khi đó, Content-Based Filtering tập trung vào đặc điểm của chính nội dung để đưa ra gợi ý, dựa trên các thuộc tính cụ thể như thể loại, diễn viên, đạo diễn hoặc từ khóa mô tả nội dung [13,14].

a. Phương pháp lọc Collaborative Filtering

Collaborative Filtering (lọc cộng tác) là một kỹ thuật trong hệ thống gợi ý, giúp dự đoán sở thích hoặc hành vi của người dùng dựa trên dữ liệu từ những người dùng khác có hành vi tương tự. Phương pháp này không dựa vào nội dung của các sản phẩm mà dựa vào mối quan hệ giữa người dùng và sản phẩm thông qua đánh giá hoặc tương tác trước đó [20-21]. Xét ví dụ thực tế về Collaborative Filtering có thể được thể hiện qua bảng đánh giá của người dùng về các video khác nhau như sau:

Bảng 1. Collaborative Filtering trong hệ thống gợi ý video

Người dùng	Video A	Video B	Video C	Video D	Video E
User 1	5	4	?	2	?
User 2	3	?	4	?	5
User 3	4	5	?	3	?
User 4	?	2	5	?	4

Trong bảng 1, dấu "?" thể hiện rằng người dùng chưa đánh giá video đó. Collaborative Filtering sẽ giúp dự đoán những giá trị còn thiếu này dựa trên sự tương đồng giữa các người dùng hoặc video. Ví dụ, để dự đoán xếp hạng mà User 1 có thể dành cho Video C, ta có thể sử dụng phương pháp tính trung bình có trọng số dựa trên đánh giá của những người dùng có hành vi tương tự với User 1. Nếu User 3 có đánh giá rất giống với User 1 ở các video khác, ta có thể suy ra rằng User 1 có khả năng sẽ thích Video C với số điểm tương tự như User 3.

Áp dụng công thức tính toán để dự đoán xếp hạng của User 1 cho Video C bằng cách sử dụng User-Based Collaborative Filtering. Giả sử ta tính hệ số tương đồng giữa User 1 và các người dùng khác bằng cách sử dụng công thức Cosine Similarity: (1)

$$similarity(A, B) = \frac{\sum_i r_{Ai} \cdot r_{Bi}}{\sqrt{\sum_i r_{Ai}^2} \cdot \sqrt{\sum_i r_{Bi}^2}} \quad (1)$$

Trong đó:

r_{Ai} là điểm đánh giá của User A cho video i.

r_{Bi} là điểm đánh giá của User B cho video i.

Tính toán độ tương đồng giữa User 1 và các User khác dựa trên các video mà cả hai đã đánh giá.

Tính similarity(User 1, User 2)

Dữ liệu chung giữa User 1 và User 2:

Video A: User 1 = 5, User 2 = 3

Video C: Chưa có dữ liệu (bỏ qua)

Video D: User 1 = 2, User 2 = ?

Video E: User 1 = ?, User 2 = 5

Vậy chỉ có **Video A** là dữ liệu chung.

$$similarity(User1, User2) = \frac{(5 \times 3)}{\sqrt{(5^2 + 4^2 + 2^2) \times (3^2 + 4^2 + 5^2)}} = \frac{15}{\sqrt{45 \times 50}} = \frac{15}{\sqrt{2250}} = \frac{\sqrt{10}}{10} \approx 0.316$$

b. Phương pháp lọc Content-Based Filtering

Content-Based Filtering là một phương pháp đề xuất nội dung dựa trên đặc điểm của chính nội dung đó thay vì dựa trên hành vi hoặc sở thích của người dùng khác [14], phương pháp này hoạt động bằng cách phân tích và trích xuất các thuộc tính của từng mục, sau đó so sánh mức độ tương đồng giữa các mục để đưa ra gợi ý phù hợp với sở thích hoặc tìm kiếm của người dùng [13], trong đó một trong những kỹ thuật phổ biến nhất để đo lường sự tương đồng là Cosine Similarity, giúp đánh giá mức độ giống nhau giữa hai văn bản dựa trên tần suất xuất

hiện của các từ [22], thuật toán này sẽ chuẩn hóa văn bản đầu vào bằng cách chuyển đổi về chữ thường, loại bỏ dấu câu và các ký tự đặc biệt, sau đó xây dựng vector đại diện cho nội dung để tính toán khoảng cách giữa các nội dung với truy vấn của người dùng [23], cuối cùng hệ thống sẽ sắp xếp các mục theo điểm tương đồng và trả về danh sách các nội dung có liên quan cao nhất, giúp nâng cao trải nghiệm cá nhân hóa và tối ưu hóa khả năng tìm kiếm thông tin [2].

Ý tưởng

Content-Based Filtering hoạt động dựa trên nội dung của video (tiêu đề, mô tả) để tìm kiếm và đề xuất video có độ tương đồng cao với từ khóa mà người dùng nhập vào. Điều này giúp tối ưu kết quả tìm kiếm và cải thiện trải nghiệm người dùng.

Các bước thực hiện

Giả sử có một danh sách video với các thuộc tính:

Bảng 2. Content-based Filtering trong hệ thống tìm kiếm gợi ý tối ưu cho video

Video	Chủ đề	Đạo diễn	Thể loại	Độ dài (phút)
A	Công nghệ	John Smith	Khoa học	10
B	Công nghệ	Anna Brown	Khoa học	12
C	Âm nhạc	Jack White	Giải trí	8
D	Công nghệ	John Smith	Học thuật	15

Người dùng nhập từ khóa tìm kiếm, thuật toán sẽ tìm các video có nội dung gần giống với từ khóa đó dựa trên Cosine Similarity.

Bước 1: Chuẩn hóa từ khóa tìm kiếm

Người dùng nhập từ khóa tìm kiếm "Công nghệ Khoa học".

- Chuyển thành chữ thường: "công nghệ khoa học"
- Xóa ký tự đặc biệt (nếu có)

Kết quả sau khi chuẩn hóa: "công nghệ khoa học"

Bước 2: Lấy danh sách video và chuẩn hóa nội dung

Mỗi video có tiêu đề và mô tả, ta sẽ ghép chúng lại và chuẩn hóa. Bảng 3 mô tả và chuẩn hóa nội dung video như sau:

Bảng 3. Danh sách video và chuẩn hóa nội dung

Video	Tiêu đề	Mô tả	Nội dung sau khi chuẩn hóa
A	"Tìm hiểu Công nghệ"	"Ứng dụng trong Khoa học"	"tìm hiểu công nghệ ứng dụng trong khoa học"
B	"Công nghệ AI"	"Tương lai Khoa học"	"công nghệ ai tương lai khoa học"
C	"Âm nhạc hay nhất"	"Giải trí và thư giãn"	"âm nhạc hay nhất giải trí và thư giãn"

Video	Tiêu đề	Mô tả	Nội dung sau khi chuẩn hóa
D	"Học thuật Công nghệ"	"John Smith hướng dẫn"	"học thuật công nghệ john smith hướng dẫn"

Bước 3: Tính độ tương đồng bằng Cosine Similarity

Giả sử từ khóa tìm kiếm query là: "công nghệ khoa học"

Ta tính độ tương đồng của nó với nội dung từng video bằng Cosine Similarity.

Tạo vector đặc trưng

Tập hợp tất cả từ xuất hiện trong từ khóa và nội dung các video:

["công", "nghệ", "khoa", "học", "tìm", "hiểu", "ứng", "dụng", "trong", "ai", "tương", "lai", "học thuật", "john", "smith", "hướng", "dẫn"]

Bảng 4 mô tả thông tin số lần xuất hiện của mỗi từ.

Bảng 4. Content-based Filtering tạo vector đặc trưng

Từ	Query (công nghệ khoa học)	Video A	Video B	Video C	Video D
công	1	1	1	0	1
nghệ	1	1	1	0	1
khoa	1	1	1	0	0
học	1	1	1	0	1
tìm hiểu	0	1	0	0	0
ứng dụng	0	1	0	0	0
trong	0	1	0	0	0
ai	0	0	1	0	0
tương lai	0	0	1	0	0
học thuật	0	0	0	0	1
john smith	0	0	0	0	1

Bước 4: Xếp hạng và trả về kết quả

Thuật toán sắp xếp theo độ tương đồng giảm dần và trả về danh sách top 3 video phù hợp nhất: Video B **(0.816)**; Video A **(0.758)**; Video D **(0.75)**

c. So sánh đánh giá hai phương pháp lọc Collaborative và Content-based

Bảng 5. So sánh hiệu quả đề xuất của hai thuật toán Collaborative và Content-based

Tiêu chí	Collaborative Filtering	Content-Based Filtering
Cơ chế hoạt động	Dựa vào lịch sử tương tác của nhiều người dùng để tìm ra sở thích chung.	Dựa vào đặc điểm nội dung để đưa ra gợi ý.
Độ cá nhân hóa	Cao, vì dựa trên sở thích của người dùng có hành vi tương tự.	Trung bình, vì chỉ dựa vào sở thích của người dùng cá nhân.

Tiêu chí	Collaborative Filtering	Content-Based Filtering
Đề xuất nội dung mới	Khó khăn với nội dung chưa có lượt xem hoặc đánh giá (Cold Start Problem).	Có thể đề xuất nội dung mới dựa trên đặc điểm của nội dung.
Tính đa dạng nội dung	Cao, vì có thể đề xuất video thuộc nhiều thể loại khác nhau.	Thấp hơn, do chỉ tập trung vào nội dung tương tự với lịch sử xem của người dùng.
Khả năng mở rộng	Có thể yêu cầu tài nguyên lớn khi dữ liệu người dùng tăng lên.	Dễ mở rộng hơn vì chỉ phân tích nội dung video.
Vấn đề dữ liệu	Cần dữ liệu từ nhiều người dùng để có đề xuất tốt.	Chỉ cần dữ liệu của một người dùng nhưng yêu cầu phân tích nội dung chi tiết.
Ứng dụng thực tế	Netflix, YouTube (gợi ý dựa trên thói quen xem của người dùng khác).	YouTube, Spotify (gợi ý dựa trên đặc điểm nội dung).

3. MÔ HÌNH HỆ THỐNG VÀ ỨNG DỤNG THỰC NGHIỆM

3.1. Mô hình hệ thống

Hình 1 mô tả mô hình hệ thống gợi ý video gồm các luồng:

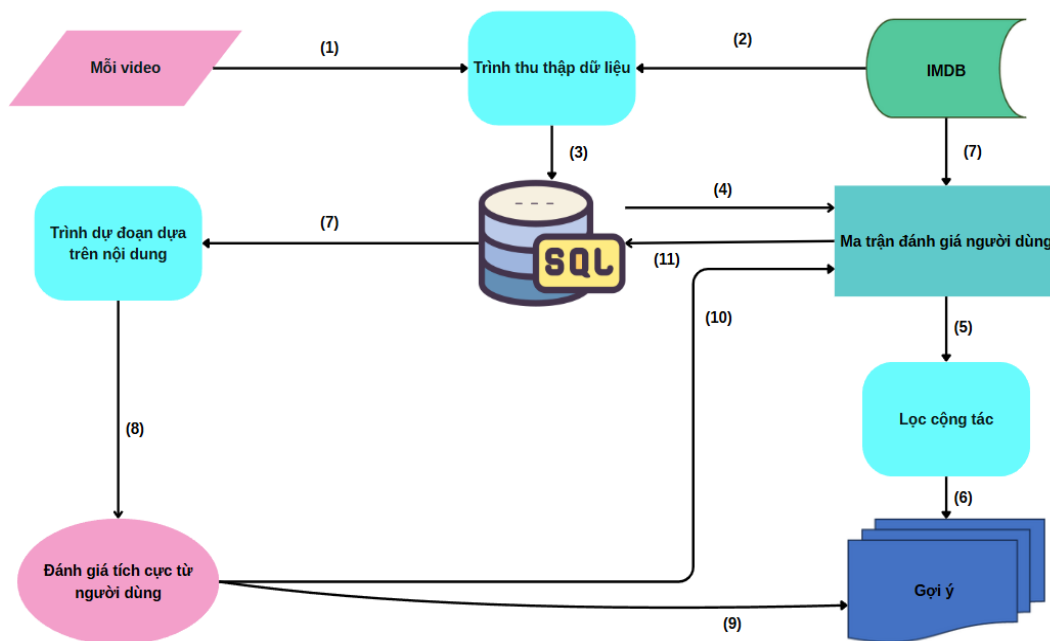
(1) Khi một video mới được thêm vào hệ thống, các thông tin quan trọng sẽ được thu thập như: Tiêu đề video, Mô tả nội dung (description), thể loại (genre/category), từ khóa (tags), độ dài video, tác giả, người đăng tải, ngày đăng. Các thông tin này được xử lý sơ bộ (tiền xử lý): loại bỏ stop words, chuẩn hóa văn bản, trích xuất từ khóa... trước khi lưu trữ và sử dụng cho việc phân tích nội dung sau này.

(2) Dữ liệu từ nguồn ngoài như IMDb cũng được sử dụng để làm giàu cho video: Rating (đánh giá sao), Thể loại theo chuẩn IMDb, Thời lượng, Nội dung tóm tắt, Các diễn viên, đạo diễn. Việc tích hợp nguồn dữ liệu bên ngoài giúp tăng độ chính xác và phong phú cho đặc trưng nội dung video, làm nền cho Content-Based Filtering.

(3) Sau khi đã xử lý, toàn bộ thông tin được lưu vào SQL Database. Đây là nơi lưu trữ tất cả dữ liệu liên quan đến video, người dùng, lịch sử xem, đánh giá, lượt thích... Cơ sở dữ liệu này là nền tảng để: Xây dựng ma trận người dùng – video, huấn luyện mô hình đề xuất, truy vấn gợi ý theo thời gian thực.

(4) Hệ thống xây dựng ma trận người dùng – video (User-Item Matrix), trong đó: Hàng là người dùng (UserID), Cột là video (VideoID), Giá trị là mức độ tương tác (rating, thời gian xem, yêu thích...).

(5) Từ ma trận trên, hệ thống áp dụng các thuật toán lọc cộng tác như: User-Based CF: Tìm những người dùng có hành vi giống nhau (tương đồng cosine hoặc Pearson correlation); Item-Based CF: Tìm những video có mức đánh giá tương tự.



Hình 1. Mô hình hệ thống gợi ý video HTB

(6) Danh sách video sau khi được lọc cộng tác sẽ được đưa đến khối "Gợi ý", nơi nó sẽ hiển thị gợi ý cho người dùng ở giao diện chính, sắp xếp theo độ phù hợp cao đến thấp, có thể kết hợp thêm dữ liệu real-time.

(7) Song song với lọc cộng tác, hệ thống dùng Content-Based Filtering: Trích xuất đặc trưng nội dung từ metadata video, Áp dụng TF-IDF hoặc mô hình ngôn ngữ (Word2Vec, BERT), Tính toán mức độ tương đồng (cosine similarity) giữa video người dùng đã xem với video khác; Mục tiêu là tìm các video giống về nội dung, không cần biết người dùng khác thích gì.

(8) Hệ thống xác định các video mà người dùng từng tương tác tích cực, dựa vào các video này, nó tìm kiếm những video khác có đặc điểm nội dung tương tự.

(9) Danh sách video có nội dung tương tự những video yêu thích được đưa ra như một gợi ý mới cho người dùng. Những đề xuất này có thể: gợi ý phim cùng thể loại, Video cùng chủ đề hoặc có từ khóa tương tự, Video từ cùng một nhà sáng tạo.

(10) Những đánh giá này được ghi nhận để cập nhật lại ma trận người dùng – video, giúp hệ thống học liên tục và cải thiện trong các lần gợi ý sau.

(11) Ma trận được lưu lại để cập nhật hệ thống theo thời gian thực hoặc batch, đảm bảo mỗi lần người dùng tương tác là một lần học thêm.

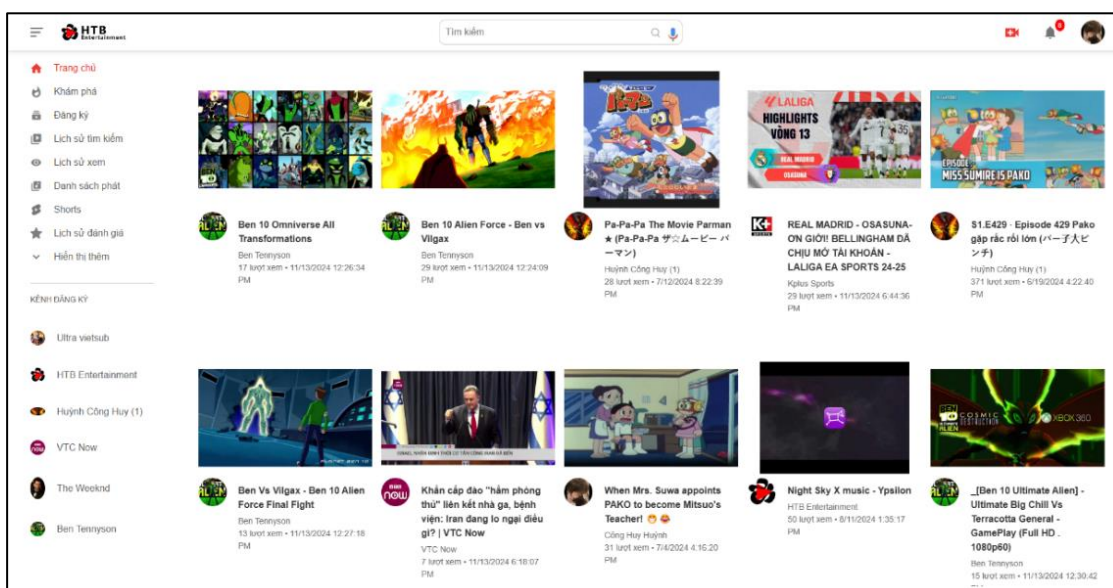
3.2. Cài đặt ứng dụng thực nghiệm

HTB Entertainment là một nền tảng chia sẻ video trực tuyến toàn cầu do Google sở hữu, cho phép người dùng tải lên, xem, chia sẻ, bình luận và tương tác với các video ở nhiều thể loại khác nhau như âm nhạc, vlog, phim ảnh, tin tức, giáo dục và nhiều nội dung sáng tạo khác, đồng thời cung cấp cơ hội kiếm tiền cho các nhà sáng tạo nội dung thông qua chương trình

đối tác HTB Entertainment; hệ thống này sử dụng các thuật toán phức tạp để đề xuất video phù hợp với sở thích người dùng, hỗ trợ phát trực tuyến trực tiếp, cho phép người dùng đăng ký theo dõi kênh yêu thích, lưu video vào danh sách phát cá nhân và thiết lập các chế độ hiển thị khác nhau như riêng tư hoặc công khai; ngoài ra, HTB Entertainment còn có các tính năng bổ sung như Shorts – video ngắn, Community – nơi người sáng tạo đăng bài viết, ảnh hoặc thăm dò ý kiến để tương tác với người xem, và HTB Entertainment Studio – công cụ dành cho người sáng tạo theo dõi hiệu suất video, chỉnh sửa nội dung và quản lý kênh một cách toàn diện trên cả nền tảng web lẫn ứng dụng di động.

Giao diện chính của hệ thống

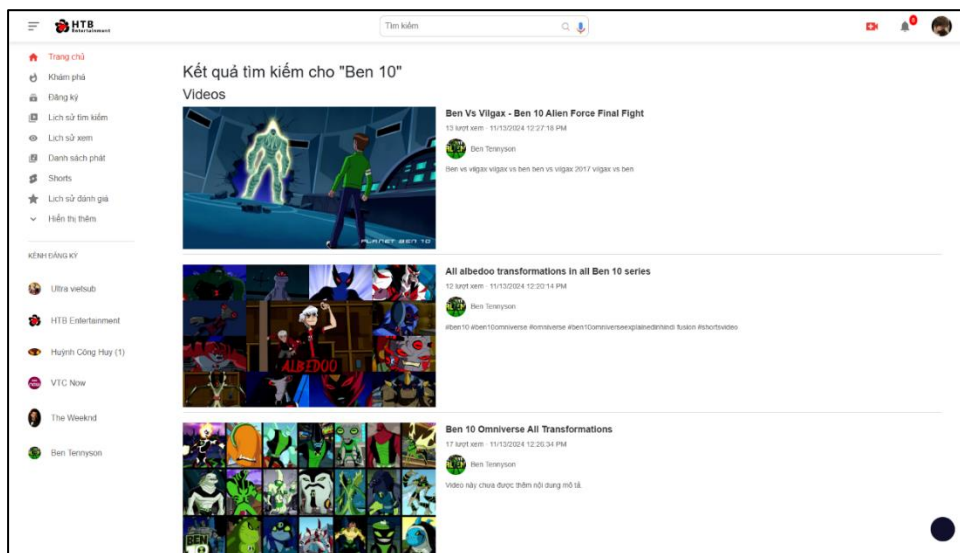
Hình 2 mô tả giao diện trang chính của hệ thống HTB Entertainment, áp dụng phương pháp lọc cộng tác Collaborative Filtering để đề xuất video dựa trên hành vi và sở thích của người dùng.



Hình 2. Màn hình giao diện chính hệ thống HTB

Giao diện được thiết kế theo phong cách tương tự HTB Entertainment, với thanh sidebar bên trái chứa các danh mục chính như Trang chủ, Khám phá, Đăng ký, Lịch sử tìm kiếm, Lịch sử xem, Danh sách phát, Shorts và một danh mục kênh đăng ký của người dùng.

Giao diện trang gợi ý video



Hình 3. Màn hình giao diện trang tìm kiếm gợi ý video HTB

Hình 3 mô tả giao diện trang tìm kiếm tối ưu của hệ thống HTB Entertainment. Giao diện có thanh tìm kiếm nằm ở phía trên cùng, nơi người dùng có thể nhập từ khóa để tìm nội dung liên quan. Bên trái là thanh sidebar chứa các danh mục như Trang chủ, Khám phá, Đăng ký, Lịch sử tìm kiếm, Lịch sử xem, Danh sách phát, Shorts và một số mục khác, giúp người dùng dễ dàng điều hướng trong hệ thống. Khi người dùng nhập một từ khóa như "Ben 10", hệ thống sẽ hiển thị danh sách kết quả tìm kiếm bao gồm video, kênh và danh sách phát liên quan. Các kết quả hiển thị theo dạng danh sách với hình thu nhỏ, tiêu đề video, số lượt xem, thời gian đăng tải và tên kênh đăng tải, giúp người dùng dễ dàng lựa chọn nội dung phù hợp. Hệ thống sử dụng phương pháp lọc Content-Based Filtering để phân tích nội dung video dựa trên mô tả, thẻ tag và các thông tin liên quan nhằm đảm bảo kết quả tìm kiếm sát với từ khóa của người dùng. Ngoài ra, nếu người dùng đã đăng nhập, từ khóa tìm kiếm sẽ được lưu lại trong lịch sử để cải thiện trải nghiệm cá nhân hóa trong tương lai. Nhờ vào phương pháp này, HTB Entertainment mang đến cho người dùng một trang tìm kiếm thông minh, tối ưu và trực quan, giúp họ nhanh chóng tìm thấy nội dung yêu thích mà không mất quá nhiều thời gian.

4. KẾT LUẬN – HƯỚNG PHÁT TRIỂN

Trong bài báo này nhóm tác giả đã tập trung nghiên cứu hai phương pháp phổ biến trong hệ thống gợi ý, bao gồm Collaborative Filtering và Content-Based Filtering. Thông qua quá trình phân tích nguyên lý hoạt động, ưu nhược điểm và khả năng ứng dụng của từng phương pháp, bài báo đã làm rõ cách thức hai kỹ thuật này hỗ trợ cá nhân hóa nội dung và cải thiện trải nghiệm người dùng trong hệ thống đề xuất video HTB.

Kết quả nghiên cứu cho thấy Collaborative Filtering có khả năng khai thác tốt thông tin từ cộng đồng người dùng, nhưng dễ bị ảnh hưởng bởi vấn đề người dùng mới (cold-start) và dữ liệu thưa. Trong khi đó, Content-Based Filtering hoạt động hiệu quả hơn trong trường hợp người dùng mới, nhưng lại bị giới hạn bởi phạm vi thông tin của chính người dùng đó. Việc

so sánh hai phương pháp cho thấy mỗi hệ thống đều có những điểm mạnh riêng, và việc lựa chọn cần phụ thuộc vào mục tiêu ứng dụng cụ thể.

Trong tương lai, nhóm tác giả tiếp tục đề xuất một hướng cải tiến kết hợp hai phương pháp trên thành hệ thống lai (hybrid system), nhằm tận dụng ưu điểm của cả hai kỹ thuật. Đây là một hướng phát triển tiềm năng giúp nâng cao độ chính xác của hệ thống gợi ý, đồng thời khắc phục được những hạn chế của từng phương pháp riêng lẻ.

TÀI LIỆU THAM KHẢO

1. Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.
2. Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.
3. Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, 5(1-2), 115-153.
4. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural Collaborative Filtering. *Proceedings of the 26th International Conference on World Wide Web*, 173-182.
5. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285-295.
6. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys*, 52(1), 1-38.
7. Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender Systems Survey. *Knowledge-Based Systems*, 46, 109-132.
8. Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4), 1-19.
9. Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Books.
10. Jiang, Z., Wang, J., Chen, Y., & Zhang, W. (2021). Privacy-Preserving Recommendation Systems: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering*.
11. Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331-370.
12. Linden, G., Smith, B., & York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 76-80.
13. Lops, P., Gemmis, M. D., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*. Springer.
14. Pazzani, M. J., & Billsus, D. (2007). Content-based Recommendation Systems. In *The Adaptive Web*. Springer.
15. Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based Recommendations with Recurrent Neural Networks. *Proceedings of the International Conference on Learning Representations (ICLR)*.
16. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformers. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 1441-1450.

17. Smith, B. & Linden, G. (2017). Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing*, 21(3), 12–18.
18. Montaner, M., López, B., & de la Rosa, J. L. (2003). A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, 19(4), 285–330.
19. Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
20. Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, 43–52.
21. Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM*, 40(3), 56–58.
22. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press*, 117–120.
23. Ca Salton, G., & Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5), 513–523.

ABSTRACT

CONFIGURATION FILTERING METHOD SUPPORTS VIDEO SUGGESTIONS ON HTB ENTERTAINMENT PLATFORM

Huỳnh Công Huy¹, Nguyễn Quốc Tiến¹, Nguyễn Quốc Bảo¹, Nguyễn Hải Yến^{1*}

¹*Ho Chi Minh City University of Industry and Trade*

* Email: yennh@huit.edu.vn

Artificial Intelligence (AI) plays a crucial role in modern technology, particularly in optimizing user experiences on online entertainment platforms. Video recommendation systems help personalize content, enhance user engagement, and retain users by suggesting videos based on their browsing history and current entertainment trends. In this paper, the authors focus on studying the Configuration Filtering algorithm, which includes Collaborative Filtering and Content-based Filtering, to support video recommendations on the HTB Entertainment platform. The main contents of the paper include: (1) Introduction to recommendation systems, (2) Configuration Filtering method in video recommendations, (3) System Model, (4) Conclusion and future development directions.

Keywords: AI, Recommendation systems, Collaborative Filtering, Content-based Filtering, Video Recommendation.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả chính:

Họ và tên: Huỳnh Công Huy

Số điện thoại liên hệ: 0917.570.175

TỐI ƯU PHÂN CỤM MỜ VỚI WHALE OPTIMIZATION ALGORITHM

Nguyễn Thị Thùy Trang¹, Phan Thị Ngọc Mai^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: maiphantn@huit.edu.vn

Ngày nhận bài: 23/3/2025 ; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Nghiên cứu này đề xuất một phương pháp cải tiến thuật toán phân cụm mờ (Fuzzy Clustering) bằng cách tích hợp Thuật toán Cá Voi (Whale Optimization Algorithm - WOA) nhằm tối ưu hóa vị trí tâm cụm, qua đó cải thiện chất lượng phân cụm và giảm sự chồng lấn giữa các nhóm dữ liệu. Nhờ khả năng tìm kiếm toàn cục dựa trên mô hình săn mồi của cá voi lưng gù, WOA giúp Fuzzy C-Means (FCM) xác định tâm cụm chính xác hơn, đồng thời hạn chế tình trạng rơi vào cực tiểu cục bộ. Kết quả thực nghiệm cho thấy mô hình WOA-FCM không chỉ nâng cao độ chính xác phân cụm, mà còn tăng cường tính ứng dụng thực tế, đặc biệt trong các bài toán phân loại dữ liệu phức tạp.

Từ khóa: Phân cụm mờ, Whale Optimization Algorithm, Fuzzy C-Means.

1. GIỚI THIỆU

Phân cụm dữ liệu là một trong những kỹ thuật quan trọng trong khai phá dữ liệu, giúp nhóm các đối tượng có đặc điểm tương đồng vào cùng một cụm. Dữ liệu trong tài chính và y tế đều có tính chất không rạch ròi, trong đó một đối tượng có thể thuộc nhiều nhóm khác nhau với mức độ thành viên nhất định, thay vì chỉ được gán vào một nhóm cố định. Trong lĩnh vực tài chính, điều này đặc biệt quan trọng trong phân tích hành vi khách hàng, đánh giá rủi ro tín dụng, quản lý danh mục đầu tư và phát hiện gian lận, nơi ranh giới giữa các phân khúc không rõ ràng. Tương tự, trong lĩnh vực y tế, một bệnh nhân có thể mắc nhiều bệnh đồng thời hoặc có triệu chứng thuộc nhiều nhóm bệnh khác nhau, khiến việc phân loại dứt khoát trở nên khó khăn. Vì vậy, gom cụm mờ (Fuzzy Clustering) trở thành một phương pháp hiệu quả để mô hình hóa dữ liệu phức tạp, giúp tăng độ chính xác và linh hoạt trong phân tích.

Fuzzy C-Means (FCM) là một trong những thuật toán gom cụm mờ phổ biến, được áp dụng rộng rãi trong phân khúc khách hàng, nhận diện mô hình chi tiêu, quản lý rủi ro tài chính, cũng như hỗ trợ chẩn đoán bệnh, phân tích dữ liệu y tế và hình ảnh y khoa. Chẳng hạn, tài liệu [1] đề xuất sử dụng FCM để xây dựng chỉ số đánh giá tài chính mờ, giúp hỗ trợ nhà đầu tư đưa ra quyết định trong ngắn hạn. Tuy nhiên, phương pháp này vẫn còn một số hạn chế như dễ bị nhiễu tác động và dễ rơi vào cực tiểu cục bộ, làm giảm tính ổn định và độ tin cậy khi triển khai trên dữ liệu tài chính thực tế.

Nhằm khắc phục những hạn chế này, nhiều nghiên cứu đã đề xuất các phiên bản nâng cấp của FCM. Trong đó, tài liệu [2] đã đề xuất thuật toán Multiple Fuzzy C-Means (M-FCM) nhằm cải thiện độ chính xác trong phân cụm dữ liệu. M-FCM cho thấy hiệu quả cao hơn so với FCM truyền thống nhưng vẫn gặp hạn chế về tốc độ hội tụ và khả năng xử lý dữ liệu nhiễu.

Một số nghiên cứu đã đề xuất kết hợp FCM với các thuật toán tối ưu hóa metaheuristic, trong đó có những phương pháp mô phỏng hành vi săn mồi của loài sói xám (*Canis lupus*). Tài liệu [3] đã chứng minh rằng sự kết hợp giữa thuật toán Grey Wolf Optimizer (GWO) và FCM giúp cải thiện đáng kể hiệu suất tối ưu hóa cụm dữ liệu.

Trong nghiên cứu này, bài báo đề xuất phương pháp kết hợp Fuzzy C-Means (FCM) với thuật toán Cá Voi (Whale Optimization Algorithm - WOA) nhằm tối ưu hóa vị trí tâm cụm, WOA mô phỏng hành vi săn mồi của cá voi lưng gù thông qua ba cơ chế chính: vây quanh con mồi, tấn công và tìm kiếm con mồi. Nhờ đó, thuật toán có khả năng điều chỉnh vị trí cụm hiệu quả hơn, giúp cải thiện độ chính xác phân nhóm trong các tập dữ liệu phức tạp và không rạch ròi.

2. GOM CỤM MỜ

2.1. Giới thiệu

Gom cụm mờ (*Fuzzy Clustering*) là một phương pháp phân cụm dữ liệu trong học máy và khai phá dữ liệu, cho phép một điểm dữ liệu có thể thuộc về nhiều cụm với các mức độ khác nhau thay vì bị ràng buộc vào một cụm duy nhất như trong phân cụm cứng (*Hard Clustering*). Nhờ đó, phương pháp này giúp tăng tính linh hoạt trong phân tích dữ liệu, đặc biệt là khi làm việc với dữ liệu có đặc điểm chồng lấn hoặc không rõ ràng[4].

Ví dụ về ứng dụng phân cụm mờ trong chẩn đoán bệnh lý hô hấp. Giả sử chúng ta có một tập hợp gồm ba bệnh nhân: $X = \{BN_1, BN_2, BN_3\}$ và ba cụm mờ tương ứng với các nhóm bệnh: G_1 (Hen suyễn), G_2 (Viêm phế quản), và G_3 (COVID-19).

Ma trận trọng số mờ U thể hiện mức độ "thuộc về" của từng bệnh nhân đối với mỗi nhóm bệnh như sau:

$$U = \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{pmatrix}$$

BN_1 có mức độ thuộc về nhóm Hen suyễn là 0.1, Viêm phế quản là 0.7, và COVID-19 là 0.2. Điều này cho thấy BN_1 có biểu hiện triệu chứng nghiêng nhiều về viêm phế quản, nhưng vẫn có dấu hiệu liên quan đến hai nhóm còn lại.

Tương tự, BN_2 có khả năng cao nhất mắc Hen suyễn (0.5), còn BN_3 lại có mức độ liên quan đến COVID-19 cao nhất (0.6).

Phân cụm mờ cho phép bác sĩ nhìn nhận bệnh nhân trong trạng thái không hoàn toàn thuộc về một chẩn đoán đơn nhất, mà thay vào đó đánh giá xác suất bệnh nhân có thể biểu hiện đồng thời triệu chứng của nhiều loại bệnh. Điều này rất hữu ích trong bối cảnh triệu chứng bệnh hô hấp thường chồng lấn và khó phân biệt rõ ràng, giúp đưa ra quyết định lâm sàng toàn diện và chính xác hơn.

2.2. Thuật toán Fuzzy C-Means

Một trong những thuật toán phân cụm mờ phổ biến nhất là Fuzzy C-Means (FCM) [5]. Tuy nhiên, để áp dụng hiệu quả FCM vào thực tế, cần hiểu rõ nguyên lý hoạt động của thuật toán. Dưới đây là trình bày chi tiết về thuật toán FCM.

Bước 1. Nhập dữ liệu đầu vào: Tập dữ liệu cần phân cụm, số cụm (C) và tham số mờ (m).

Bước 2. Khởi tạo ma trận độ thuộc U: Giá trị ban đầu của ma trận độ thành viên được gán ngẫu nhiên sao cho tổng độ thành viên của mỗi điểm dữ liệu bằng 1.

Bước 3. Tính toán tâm cụm: Sử dụng công thức:

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m \cdot x_i}{\sum_{i=1}^n u_{ij}^m} \quad (1)$$

Trong đó v_j là tâm cụm thứ j , u_{ij} là độ thành viên của điểm dữ liệu x_i đối với cụm j .

Bước 4. Cập nhật ma trận độ thuộc U:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

Bước 5. Kiểm tra điều kiện hội tụ: Nếu sự thay đổi của ma trận U hoặc tâm cụm nhỏ hơn một ngưỡng ϵ , thuật toán dừng. Nếu không, quay lại bước 3.

Bước 6. Kết thúc thuật toán: Xuất kết quả phân cụm.

Mặc dù FCM là một phương pháp phân cụm hiệu quả, nhưng vẫn tồn tại một số hạn chế, bao gồm độ nhạy cao với nhiễu, dễ rơi vào cực tiểu cục bộ, và khả năng phân tách kém khi dữ liệu có sự chồng lấn. Những yếu tố này có thể làm giảm độ chính xác của mô hình, ảnh hưởng đến hiệu suất phân cụm trong thực tế. Để khắc phục những nhược điểm này, nghiên cứu đề xuất tích hợp thuật toán WOA và FCM nhằm cải thiện chất lượng phân cụm. Phân tiếp theo trình bày chi tiết về thuật toán WOA

3. THUẬT TOÁN WHALE OPTIMIZATION ALGORITHM - WOA

Thuật toán WOA là một thuật toán tối ưu metaheuristic được đề xuất bởi Mirjalili và cộng sự vào năm 2016. WOA mô phỏng hành vi săn mồi của cá voi lưng gù, đặc biệt là chiến lược “Bubble-Net Attacking” khi săn mồi. Thuật toán này được sử dụng rộng rãi trong tối ưu hóa hàm số, điều chỉnh tham số mô hình máy học và cải thiện hiệu suất phân cụm dữ liệu[5].

Cơ chế hoạt động của WOA: Thuật toán WOA mô phỏng hành vi săn mồi của cá voi lưng gù và hoạt động dựa trên ba cơ chế chính:

1. Encircling Prey

- Cá voi nhận diện vị trí con mồi tốt nhất và di chuyển theo nó.
- Cập nhật vị trí cá voi bằng cách điều chỉnh khoảng cách so với con mồi.

2. Bubble-Net Attacking

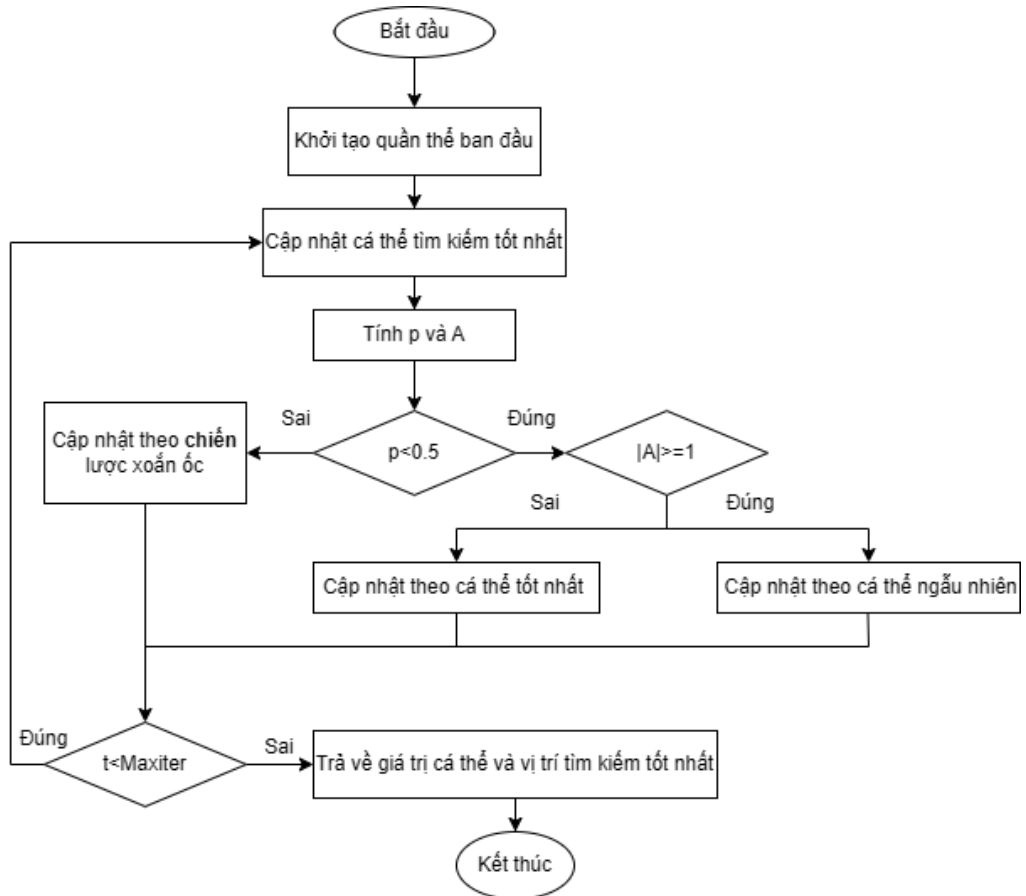
- Cá voi thực hiện chuyển động xoắn ốc để tiếp cận con mồi bằng cách thu hẹp phạm vi tìm kiếm.
- Điều chỉnh vị trí theo một quỹ đạo xoắn, giúp tăng khả năng khai thác (exploitation).

3. Search for Prey

- Khi chưa xác định con mồi tốt nhất, cá voi di chuyển ngẫu nhiên trong không gian tìm kiếm.

- Cơ chế này giúp thuật toán có khả năng thăm dò (exploration) và tránh rơi vào cực tiểu cục bộ.

Lưu đồ thuật toán WOA được biểu diễn dưới hình 1:



Hình 1. Lưu đồ thuật toán WOA

Mỗi cá thể trong quần thể biểu diễn một lời giải và được khởi tạo một cách ngẫu nhiên trong không gian tìm kiếm. Sau bước khởi tạo, thuật toán tiến hành cập nhật cá thể tốt nhất và tính toán các tham số điều khiển như A và p .

- Tham số A được sử dụng để điều chỉnh khoảng cách giữa cá thể hiện tại và cá thể mục tiêu, dựa trên cơ chế săn mồi đặc trưng của thuật toán.
- Tham số p là một giá trị ngẫu nhiên trong khoảng $[0,1]$, được sử dụng để quyết định chiến lược di chuyển của cá voi.

Dựa vào giá trị của A và p , thuật toán sẽ lựa chọn một trong ba chiến lược: bao vây con mồi, thăm dò không gian tìm kiếm hoặc di chuyển theo quỹ đạo xoắn ốc. Quá trình này được lặp lại cho đến khi đạt số vòng lặp tối đa, sau đó thuật toán trả về lời giải tối ưu thu được.

WOA có nhiều ưu điểm nổi bật, bao gồm khả năng cân bằng hiệu quả giữa khai thác (exploitation) và thăm dò (exploration), giúp quá trình tìm kiếm lời giải đạt hiệu suất cao. Ngoài ra, thuật toán này có cơ chế tìm kiếm động, giúp tránh rơi vào cực tiểu cục bộ, từ đó cải thiện chất lượng tối ưu. Nhờ những đặc điểm này, WOA được ứng dụng rộng rãi trong các bài toán tối ưu hóa, chẳng hạn như phân cụm dữ liệu, huấn luyện mạng nơ-ron nhân tạo và điều chỉnh tham số mô hình.

4. THUẬT TOÁN WHALE OPTIMIZATION ALGORITHM - FUZZY C-MEANS (WOA-FCM)

4.1. Giới thiệu

WOA-FCM là thuật toán được phát triển bằng cách tích hợp giữa thuật toán WOA) và FCM. Sự kết hợp này tận dụng khả năng tìm kiếm toàn cục hiệu quả của WOA để tinh chỉnh các tham số trong FCM, từ đó cải thiện chất lượng phân cụm và tăng độ chính xác. So với phương pháp FCM truyền thống, thuật toán WOA-FCM mang lại kết quả ổn định hơn, khả năng kháng nhiễu tốt hơn, tránh được hiện tượng mắc kẹt tại cực tiểu cục bộ, và đặc biệt hiệu quả khi xử lý các tập dữ liệu có đặc điểm chồng lấn.

4.2. Hàm mục tiêu cho thuật toán WOA-FCM

Trong thuật toán WOA-FCM, mục tiêu chính là tối ưu hóa quá trình phân cụm bằng cách xác định vị trí tối ưu của các tâm cụm. Điều này giúp giảm tổng sai số giữa các điểm dữ liệu và tâm cụm, cải thiện hiệu quả phân cụm so với phương pháp FCM truyền thống.

Trong thuật toán WOA-FCM, WOA đảm nhận vai trò tối ưu hóa vị trí các tâm cụm nhằm giảm giá trị hàm mục tiêu. Hàm mục tiêu được định nghĩa như sau:

$$J(U, V) = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m d^2(x_i, v_j) \quad (3)$$

Trong đó:

N : Số lượng điểm dữ liệu.

C : Số cụm.

u_{ij} : Mức độ thành viên (fuzzy membership) của điểm x_i trong cụm j

m (thường $m=2$): Hệ số fuzziness, điều chỉnh độ mờ của phân cụm.

v_j : Tọa độ của tâm cụm j (vector cần tối ưu bởi WOA).

$d^2(x_i, v_j)$: Khoảng cách Euclidean bình phương giữa điểm dữ liệu x_i và tâm cụm

v_j : $d^2(x_i, v_j) = \|x_i - v_j\|^2$

4.3. Quy trình hoạt động của thuật toán WOA-FCM

WOA-FCM hoạt động dựa trên hai giai đoạn chính:

- **Giai đoạn 1:** Khởi tạo cá thể trong quần thể
 - + Mỗi cá thể trong WOA đại diện cho một tập hợp các trọng số hoặc tâm cụm ban đầu của FCM.
 - + Quần thể được khởi tạo ngẫu nhiên dựa trên miền giá trị của dữ liệu.
- **Giai đoạn 2:** Cập nhật vị trí cá thể bằng WOA
 - + Các cá thể cập nhật vị trí dựa trên cơ chế săn mồi của cá voi, bao gồm bao vây con mồi, tạo vòng xoắn và tìm kiếm con mồi mới.
 - + Hàm mục tiêu được thiết lập để đánh giá chất lượng phân cụm.
 - + Cơ chế tìm kiếm động giúp tránh rơi vào cực tiểu cục bộ và cải thiện độ hội tụ.
- **Giai đoạn 3:** Phân cụm dữ liệu bằng FCM

- + Sử dụng tham số tối ưu do WOA tìm được để phân cụm dữ liệu với FCM.
- + Lặp lại quá trình này cho đến khi thuật toán hội tụ hoặc đạt điều kiện dừng.

4.4. Thuật toán WOA-FCM

a. Thuật toán cập nhật vị trí cá voi bằng WOA

Trong thuật toán Whale Optimization Algorithm (WOA), quá trình cập nhật vị trí cá voi đóng vai trò quan trọng trong việc tìm kiếm và tối ưu hóa nghiệm. Mỗi cá voi trong quần thể đại diện cho một lời giải tiềm năng, và thuật toán sẽ liên tục điều chỉnh vị trí của chúng để tiến gần hơn đến điểm tối ưu.

Thuật toán Update_Whale_Pos

Đầu vào:

- V_i : Vị trí hiện tại của cá voi (tập hợp các tâm cụm)
- $Best_Whale$: Cá voi có giá trị hàm mục tiêu tốt nhất hiện tại
- $Search_Agents$: Danh sách các cá voi trong quần thể
- a : Hệ số giảm dần từ 2 đến 0

Đầu ra:

- $V_{i_mới}$: Vị trí cập nhật của cá voi

Function Update_Whale_Pos ($V_i, Best_Whale, Search_Agents, a$)

Begin:

- (1) Tạo số ngẫu nhiên p trong khoảng $[0,1]$
- (2) Tạo hai số ngẫu nhiên r_1, r_2 trong khoảng $[0,1]$
- (3) Tính toán hai vector:
 $A = 2 * a * r_1 - a$
 $C = 2 * r_2$
- (4) Nếu $p < 0.5$ thì: // Chọn bao vây con mồi hoặc tìm kiếm con mồi mới
Nếu $|A| < 1$ thì: // Bao vây con mồi
Tính $D = |C * Best_Whale - V_i|$
Cập nhật vị trí: $V_{i_mới} = Best_Whale - A * D$
Ngược lại: // Tìm kiếm con mồi mới
Chọn ngẫu nhiên một cá voi X_rand từ $Search_Agents$
Tính $D = |C * X_rand - V_i|$
Cập nhật vị trí: $V_{i_mới} = X_rand - A * D$
- (5) Ngược lại ($p \geq 0.5$): // Tạo vòng xoắn
Tính khoảng cách $D' = |Best_Whale - V_i|$
Tạo số ngẫu nhiên l trong khoảng $[-1,1]$
Cập nhật vị trí: $V_{i_mới} = D' * \exp(b * l) * \cos(2 * \pi * l) + Best_Whale$
- (6) Trả về $V_{i_mới}$

End.

b. Thuật toán WOA-FCM

Để áp dụng thuật toán WOA-FCM vào bài toán phân cụm, trước tiên, chúng ta cần xác định các thông tin đầu vào quan trọng. Dưới đây là các tham số và dữ liệu cần thiết cho thuật toán:

Dữ liệu đầu vào:

- Tập dữ liệu $X = \{x_1, x_2, \dots, x_N\}$
 - + Gồm N điểm dữ liệu trong không gian D chiều.
 - + Mỗi điểm dữ liệu x_i là một vector đặc trưng $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$
- Số cụm C: Số lượng nhóm cần phân chia dữ liệu.
- Hệ số fuzziness m: Điều chỉnh mức độ mờ của phân cụm (thường chọn $m=2$).
- Số lượng cá voi N_w : Số lượng cá thể trong quần thể WOA.
- Số vòng lặp tối đa $MaxIter$: Giới hạn số lần lặp của thuật toán.
- Ngưỡng hội tụ ϵ : Điều kiện dừng dựa trên sự thay đổi nhỏ của hàm mục tiêu.

Dữ liệu đầu ra:

- Tập hợp tâm cụm tối ưu V^* :
 - + Gồm C tâm cụm $V^* = \{v_1, v_2, \dots, v_C\}$
 - + Mỗi tâm cụm v_j có tọa độ trong không gian D chiều.
- Ma trận thành viên $U^* = [u_{ij}]$
 - + Thể hiện mức độ điểm dữ liệu x_i thuộc vào cụm j .
 - + Giá trị u_{ij} nằm trong khoảng $[0, 1]$ và tổng các giá trị trong mỗi hàng bằng 1.
- Nhãn cụm của các điểm dữ liệu:
 - + Mỗi điểm x_i được gán vào cụm có giá trị u_{ij} cao nhất.

Function WOA_FCM(X, C, m, N_w , $MaxIter$, ϵ)

Input:

(1) Khởi tạo quần thể cá voi:

- Tạo N_w cá thể, mỗi cá thể là một tập hợp tâm cụm $V = \{v_1, v_2, \dots, v_C\}$
- Khởi tạo ngẫu nhiên các vị trí tâm cụm trong miền giá trị của X
- Đánh giá hàm mục tiêu $J(U, V)$ cho mỗi cá voi

(2) Xác định cá voi tốt nhất $Best_Whale$ có giá trị J nhỏ nhất

(3) Lặp qua từng vòng lặp $t = 1$ đến $MaxIter$:

a. Cập nhật tham số WOA: a giảm dần từ 2 đến 0

b. Duyệt qua từng cá voi i trong quần thể:

- Gọi hàm cập nhật vị trí:

$$V_{i_mới} = Update_Whale_Pos(V_i, Best_Whale, Search_Agents, a)$$

- Kiểm tra ràng buộc và điều chỉnh $V_{i_mới}$ nếu cần

- Tính toán $J(U, V_{i_mới})$

- Nếu $J(U, V_{i_mới}) < J(U, V_i)$, cập nhật $V_i = V_{i_mới}$

c. Cập nhật $Best_Whale$ nếu có cá voi tốt hơn

d. Kiểm tra điều kiện dừng:

$$\text{Nếu } |J(t) - J(t-1)| < \epsilon \text{ thì dừng}$$

(4) Áp dụng thuật toán FCM với tâm cụm $V^* = Best_Whale$ để tìm ma trận thành

viên U^*

(5) Xác định nhãn cụm dựa vào giá trị lớn nhất của ma trận U^*

Return V^* , U^* , Labels

5. TÍNH HỘI TỤ CỦA THUẬT TOÁN WOA-FCM

Thuật toán WOA-FCM là một phương pháp lai được phát triển bằng cách tích hợp giữa thuật toán tối ưu cá voi WOA và thuật toán phân cụm mờ Fuzzy C-Means (FCM). Sự kết hợp này tận dụng khả năng tìm kiếm toàn cục mạnh mẽ của WOA nhằm tối ưu hóa các tham số ban đầu trong quá trình phân cụm của FCM, qua đó cải thiện độ chính xác và ổn định của kết quả phân cụm. Tính hội tụ của WOA-FCM được đảm bảo bởi đặc điểm lặp cập nhật thích nghi của thuật toán WOA, cho phép các cá thể liên tục di chuyển về vùng tìm kiếm có độ thích nghi cao hơn. Nhờ đó, thuật toán có khả năng thoát khỏi các điểm cực tiểu cục bộ – một hạn chế phổ biến trong FCM truyền thống vốn phụ thuộc mạnh vào giá trị khởi tạo ban đầu. Thêm vào đó, việc sử dụng WOA giúp mở rộng không gian tìm kiếm và định hướng cập nhật vị trí các tâm cụm một cách linh hoạt, làm tăng khả năng hội tụ tới nghiệm tối ưu toàn cục. Kết quả là thuật toán WOA-FCM hội tụ nhanh hơn, ổn định hơn, và kháng nhiễu tốt hơn, đặc biệt hiệu quả khi áp dụng cho các tập dữ liệu có cấu trúc phức tạp hoặc các cụm chồng lấn.

6. KẾT QUẢ

Trong nghiên cứu này, phương pháp được áp dụng trên một tập dữ liệu mô phỏng gồm 600 khách hàng, được xây dựng nhằm phản ánh các đặc điểm hành vi tiêu dùng phổ biến trong bối cảnh kinh doanh dịch vụ tại Việt Nam, dựa trên tham khảo từ các nghiên cứu trước và dữ liệu nhân khẩu học thực tiễn [6], [7], [8]. Các khách hàng trong tập dữ liệu được phân thành ba phân khúc hành vi tiêu dùng chính, bao gồm: Cụm 1 – nhóm chi tiêu cao; Cụm 2 – nhóm tiêu dùng ổn định; và Cụm 3 – nhóm khách hàng có xu hướng chi tiêu tiết kiệm. Dữ liệu bao gồm các đặc trưng đầu vào định lượng và định tính phản ánh hành vi tiêu dùng thực tế, như: thu nhập hàng tháng (triệu VND), chi tiêu trung bình hàng tháng (triệu VND), tuổi, giới tính (0: nữ, 1: nam), và mức độ hài lòng với dịch vụ được đánh giá theo thang điểm từ 1 đến 5, như được trình bày trong Bảng 1.

Bảng 1. Dữ liệu phân tích hành vi khách hàng

Khách hàng	Thu nhập	Chi tiêu	Tuổi	Giới tính	Độ hài lòng
KH 01	32.48357	19.30868	53	1	4
KH 02	33.23844	27.61515	27	1	3
KH 03	28.82923	18.82932	50	0	4
KH 04	37.89606	23.83717	22	0	2
KH 05	27.65263	22.7128	52	1	3
KH 06	27.68291	17.67135	51	1	4
KH 07	31.20981	10.4336	50	0	2
KH 08	21.37541	17.18856	38	1	3
KH 09	24.93584	21.57124	34	1	4
KH 10	25.45988	12.93848	39	1	2

Dựa trên dữ liệu đặc trưng của khách hàng được trình bày trong Bảng 1, thuật toán WOA-FCM được áp dụng để phân cụm và xác định mức độ thuộc của từng khách hàng vào

các nhóm hành vi tiêu dùng. Kết quả được thể hiện trong Bảng 2, cho thấy mỗi khách hàng có thể đồng thời thuộc về nhiều cụm với các mức độ khác nhau, phản ánh bản chất mờ của hành vi tiêu dùng trong thực tế.

Bảng 2. Mức độ thuộc của khách hàng vào từng cụm hành vi tiêu dùng

Khách hàng	Cụm 1	Cụm 2	Cụm 3
KH 01	0.2	0.7	0.1
KH 02	0.6	0.3	0.1
KH 03	0.1	0.8	0.1
KH 04	0.7	0.2	0.1
KH 05	0.1	0.2	0.7
KH 06	0.2	0.6	0.2
KH 07	0.1	0.2	0.7
KH 08	0.2	0.7	0.1
KH 09	0.6	0.3	0.1
KH 10	0.1	0.3	0.6

WOA-FCM giúp cải thiện độ phân biệt giữa các cụm nhờ khả năng tìm kiếm toàn cục, từ đó giảm hiện tượng chồng lấn và tăng độ ổn định. Cụ thể, mô hình đạt Silhouette Score = 0.538, cao hơn so với FCM truyền thống (0.492), chứng minh khả năng phân tách cụm tốt hơn. Phương pháp này đặc biệt hiệu quả ở những vùng dữ liệu giao thoa, đồng thời hỗ trợ các ứng dụng như cá nhân hóa sản phẩm, tối ưu chiến dịch marketing và dự đoán xu hướng tiêu dùng. Nhìn chung, WOA-FCM cho thấy ưu thế rõ rệt về hiệu suất phân cụm và khả năng thích ứng với dữ liệu thực tế so với phương pháp FCM thông thường.

7. KẾT LUẬN

Nghiên cứu đã đề xuất mô hình phân cụm mờ WOA-FCM nhằm tối ưu hóa vị trí tâm cụm và cải thiện độ chính xác phân cụm. Nhờ khả năng tìm kiếm toàn cục của thuật toán WOA, mô hình giúp giảm chồng lấn cụm, hạn chế nhiễu và khắc phục điểm yếu của FCM truyền thống. Kết quả thực nghiệm cho thấy WOA-FCM đạt hiệu quả phân loại cao hơn, phản ánh rõ bản chất mờ trong dữ liệu và tiềm năng ứng dụng trong các hệ thống hỗ trợ ra quyết định. Phương pháp này đặc biệt phù hợp với các bài toán thực tế có yếu tố không chắc chắn, như phân khúc hành vi tiêu dùng hoặc đánh giá mức độ hài lòng. Trong tương lai, hướng nghiên cứu sẽ tiếp tục mở rộng kết hợp WOA-FCM với các kỹ thuật tối ưu hóa khác để nâng cao hiệu suất và độ chính xác phân cụm.

TÀI LIỆU THAM KHẢO

1. Gokten S., Kilic E., and Uluc A. – Using fuzzy c-means clustering algorithm in financial health scoring. *Audit Financiar*, 15 (149) (2017) 548–558.
2. Wu Y., Duan H., and Du S. – Multiple fuzzy c-means clustering algorithm in medical diagnosis. *Technology and Health Care*, 23 (Suppl. 2) (2015) S519–S527.
3. Mohammadian-Khoshnoud M., Soltanian A.R., Dehghan A., and Farhadian M. – Optimization of fuzzy c-means (FCM) clustering in cytology image segmentation using the gray wolf algorithm. *BMC Molecular and Cell Biology*, 23 (9) (2022).

4. Bezdek J.C. – Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press (1981).
5. Mirjalili S., and Lewis A. – The Whale Optimization Algorithm. *Advances in Engineering Software*, 95 (2016) 51–67.
6. Kaggle – Customer Personality Analysis Dataset
7. Tổng cục Thống kê Việt Nam (GSO) – Hệ thống chỉ tiêu thống kê thị trường tiêu dùng Việt Nam
8. Berry M.J.A., and Linoff G.S. – *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 3rd ed. Wiley (2011).

ABSTRACT

OPTIMIZING FUZZY CLUSTERING WITH THE WHALE OPTIMIZATION ALGORITHM

Nguyen Thi Thuy Trang¹, Phan Thi Ngoc Mai^{1,*}

¹*Ho Chi Minh City University of Industry and Trade*

*Email: maiphantn@huit.edu.vn

This study proposes an improved fuzzy clustering method by integrating the Whale Optimization Algorithm (WOA) to optimize cluster center positions, thereby enhancing clustering quality and reducing overlap among data groups. Leveraging WOA's global search capability—based on the humpback whale's hunting behavior—the approach enables Fuzzy C-Means (FCM) to identify more accurate cluster centers and mitigates the risk of local minima. Experimental results demonstrate that the WOA-FCM model not only improves clustering accuracy but also increases practical applicability, especially in complex data classification problems.

Keywords: Fuzzy Clustering, Whale Optimization Algorithm, Fuzzy C-Means.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thị Thùy Trang

Số điện thoại liên hệ: 0936522047

ỨNG DỤNG CÁC CÔNG CỤ TÍCH HỢP AI TRONG THIẾT KẾ WEBSITE: XU HƯỚNG VÀ TIỀM NĂNG PHÁT TRIỂN

Lâm Thị Họa Mi^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: milth@huit.edu.vn

Ngày nhận bài: 26/3/2025 ; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Trong thời đại phát triển mạnh mẽ về công nghệ và kỹ thuật số, trí tuệ nhân tạo (AI) đang ngày càng xác định vai trò quan trọng trong nhiều lĩnh vực, đặc biệt trong thiết kế ứng dụng website. Nghiên cứu này tập trung giới thiệu các công cụ tích hợp AI hiện đại đã và đang được sử dụng trong thiết kế website, đồng thời phân tích lợi ích, thách thức và tiềm năng phát triển của chúng. Qua đó, cung cấp cái nhìn tổng quan và định hướng cho các nhà thiết kế, giáo dục và doanh nghiệp trong việc ứng dụng AI để tối ưu hóa trải nghiệm người dùng và tăng hiệu suất của ứng dụng website.

Từ khóa: thiết kế website thông minh, tự động hóa thiết kế website, thiết kế website.

1. GIỚI THIỆU

Trí tuệ nhân tạo (AI) đang tạo ra sự chuyển đổi mạnh mẽ trong lĩnh vực thiết kế website, giúp tối ưu hóa trải nghiệm người dùng (UX), hiệu suất SEO và cá nhân hóa nội dung [1, 2]. Trước đây, việc tạo ra website đòi hỏi sự tham gia của cả nhà thiết kế đồ họa và lập trình viên, gây tốn kém chi phí và thời gian. Tuy nhiên, các công cụ tích hợp AI đã thay đổi hoàn toàn cách tiếp cận này bằng cách tự động hóa hầu hết các bước thiết kế, từ bố cục, màu sắc đến tối ưu hóa nội dung. AI không chỉ tăng tốc độ thiết kế mà còn cá nhân hóa trải nghiệm người dùng thông qua phân tích dữ liệu hành vi, giúp doanh nghiệp nâng cao hiệu quả kinh doanh và sự gắn kết của khách hàng.



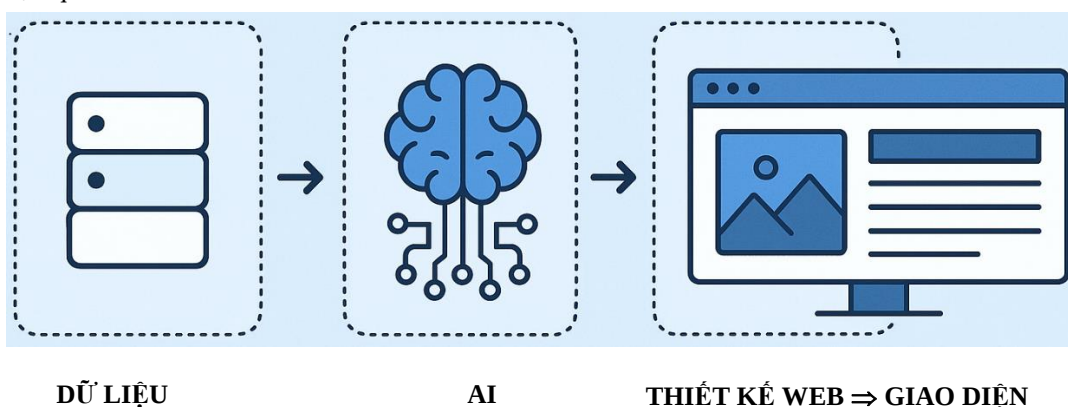
Hình 1.1. Thiết kế website bằng AI

Dù mang lại nhiều lợi ích, AI trong thiết kế website cũng đối mặt với thách thức về tính sáng tạo, bảo mật dữ liệu và phụ thuộc vào công nghệ. Bài báo này sẽ giới thiệu chi tiết các công cụ tích hợp AI phổ biến trong thiết kế website, phân tích lợi ích, thách thức cũng như dự báo xu hướng phát triển của công nghệ này trong tương lai. Qua đó, cung cấp một cái nhìn tổng quan và định hướng cho các nhà thiết kế, giáo dục và doanh nghiệp trong việc khai thác hiệu quả sức mạnh của AI trong lĩnh vực thiết kế website.

2. KHÁI NIỆM VÀ VAI TRÒ CỦA AI TRONG THIẾT KẾ WEBSITE

2.1 Khái niệm về AI trong thiết kế website

Trí tuệ nhân tạo (AI) trong thiết kế website là việc sử dụng các thuật toán máy học (Machine Learning) và xử lý ngôn ngữ tự nhiên (Natural Language Processing) để tự động hóa và tối ưu hóa quá trình thiết kế giao diện người dùng (UI) và trải nghiệm người dùng (UX). Thay vì phải lập trình và thiết kế thủ công, các công cụ AI có thể phân tích dữ liệu người dùng, nhận diện xu hướng thiết kế, và tạo ra các giao diện phù hợp một cách nhanh chóng và hiệu quả.



Hình 2.1. Nguyên lý hoạt động của AI trong thiết kế web

AI trong thiết kế website không chỉ dừng lại ở việc tạo ra giao diện thẩm mỹ mà còn bao gồm khả năng cá nhân hóa nội dung, tối ưu hóa SEO, và thậm chí hỗ trợ viết nội dung phù hợp với đối tượng khách hàng mục tiêu. Các công cụ AI hiện đại có thể tự động điều chỉnh bố cục trang web dựa trên hành vi người dùng, cải thiện khả năng truy cập và tăng cường trải nghiệm cá nhân hóa.

2.2. Vai trò của AI trong thiết kế website

AI đang thay đổi cách tiếp cận truyền thống trong thiết kế website [3, 4, 5], mang lại nhiều lợi ích và vai trò quan trọng như sau:

Tự động hóa quy trình thiết kế: AI tự động chọn màu sắc, bố cục và kiểu chữ, giúp tiết kiệm thời gian và công sức, đặc biệt cho người không có kinh nghiệm lập trình.

Cá nhân hóa trải nghiệm người dùng (UX): AI phân tích hành vi người dùng để đưa ra đề xuất cá nhân hóa, giúp tăng tỷ lệ chuyển đổi và giữ chân khách hàng lâu hơn.

Tối ưu hóa SEO và nội dung: AI phân tích từ khóa, tạo nội dung chuẩn SEO và đề xuất cải thiện thẻ meta, hình ảnh để tăng thứ hạng tìm kiếm.

Phân tích dữ liệu và cải tiến: AI phân tích dữ liệu lớn để hiểu hành vi người dùng, từ đó đề xuất điều chỉnh thiết kế và nội dung hiệu quả hơn.

Tạo nội dung đa phương tiện: AI có thể tạo logo, video quảng cáo và thiết kế đồ họa phù hợp với xu hướng thị trường.

Hỗ trợ thử nghiệm A/B: AI tự động phân tích hiệu suất các phiên bản trang web, giúp chọn ra thiết kế hiệu quả nhất.

Bảo mật và hiệu suất: AI phát hiện nguy cơ bảo mật và tối ưu hóa hiệu suất trang web bằng cách tự động điều chỉnh tải trang và mã nguồn.

Nhờ những lợi ích này, AI đang trở thành công cụ không thể thiếu trong thiết kế website hiện đại.

2.3. Tầm quan trọng của AI trong xu hướng thiết kế hiện đại

AI đang mở ra một kỷ nguyên mới trong thiết kế website, không chỉ giúp tối ưu hóa quy trình thiết kế mà còn mang lại sự cá nhân hóa và trải nghiệm người dùng vượt trội. Nhờ vào khả năng phân tích dữ liệu mạnh mẽ, AI giúp các doanh nghiệp hiểu rõ hơn về khách hàng, từ đó xây dựng chiến lược nội dung và thiết kế giao diện phù hợp hơn.

Trong tương lai, AI không chỉ dừng lại ở việc hỗ trợ thiết kế mà còn đóng vai trò như một nhà thiết kế ảo, có khả năng tự học và sáng tạo ra các thiết kế độc đáo phù hợp với xu hướng thị trường. Do đó, việc nắm bắt và ứng dụng AI trong thiết kế website sẽ là một lợi thế cạnh tranh lớn cho các doanh nghiệp trong kỷ nguyên số hóa.

3. CÁC CÔNG CỤ TÍCH HỢP AI NỔI BẬT TRONG THIẾT KẾ WEBSITE

Trong những năm gần đây, nhiều công cụ tích hợp AI đã ra đời, giúp đơn giản hóa và tự động hóa quy trình thiết kế website. Dưới đây là một số công cụ nổi bật đang được sử dụng rộng rãi trên thị trường:

3.1. Wix ADI (Artificial Design Intelligence)

Wix ADI [6] là công cụ tiên phong sử dụng AI để tự động tạo website một cách nhanh chóng. Người dùng chỉ cần trả lời một số câu hỏi cơ bản về lĩnh vực kinh doanh và phong cách thiết kế mong muốn, Wix ADI sẽ tự động sắp xếp bố cục, lựa chọn màu sắc và nội dung phù hợp.

Tính năng chính của Wix ADI có thể tự động tạo bố cục và giao diện dựa trên thông tin đầu vào của người dùng, tích hợp trình chỉnh sửa kéo-thả (drag-and-drop) dễ sử dụng và cá nhân hóa nội dung dựa trên thông tin doanh nghiệp.

Sử dụng Wix ADI có thể giúp người dùng tiết kiệm thời gian tạo ra một website nhanh chóng và dễ sử dụng cho người không có kỹ năng lập trình với giao diện thân thiện và đa dạng mẫu thiết kế có sẵn.

Tuy nhiên, Wix ADI lại bị giới hạn về khả năng tùy chỉnh sâu và linh hoạt trong thiết kế phức tạp và các tùy chọn nâng cao bị hạn chế so với việc thiết kế thủ công bằng mã nguồn mở.

3.2. The Grid

The Grid [7] là công cụ tiên phong trong việc sử dụng AI để tự động hóa toàn bộ quá trình thiết kế website, sử dụng AI tên Molly để tự động hóa toàn bộ quá trình thiết kế website bằng cách tự học từ hành vi người dùng và điều chỉnh bố cục, màu sắc và phông chữ phù hợp.

Bên cạnh đó, the Grid tự động điều chỉnh bố cục và màu sắc dựa trên nội dung tải lên, đảm bảo thiết kế hài hòa và nhất quán. Công cụ này sử dụng AI tự học để cải thiện giao diện theo thời gian và tích hợp chức năng thương mại điện tử, blog, tự động hóa gần như toàn bộ quy trình thiết kế, giúp tiết kiệm thời gian và thích ứng với nhu cầu người dùng.

Tuy nhiên, the Grid thiếu khả năng chỉnh sửa thủ công chi tiết, không phù hợp cho nhà thiết kế chuyên nghiệp muốn kiểm soát giao diện. UX/UI của công cụ này cũng cần cải thiện để đáp ứng thị hiếu người dùng đa dạng.

3.3. Designs.ai

Designs.ai [8] là nền tảng sáng tạo sử dụng AI để thiết kế logo, video, bài viết và đồ họa, giúp doanh nghiệp tạo nội dung số nhanh chóng và chuyên nghiệp mà không cần kỹ năng thiết kế chuyên sâu.

Nổi bật với công cụ tạo logo tự động dựa trên tên thương hiệu và ngành nghề, Video Maker ghép video từ clip và hình ảnh có sẵn, cùng Copywriter AI hỗ trợ viết nội dung quảng cáo và blog, Designs.ai giúp tiết kiệm thời gian và đồng bộ hóa nội dung tiếp thị trên nhiều kênh truyền thông.

Tuy nhiên, chất lượng thiết kế tự động vẫn chưa đạt mức chuyên nghiệp cao và hạn chế trong tùy chỉnh cho các yêu cầu thiết kế phức tạp, thiếu sự sáng tạo và độc đáo cho các chiến dịch đặc biệt.

Ngoài ra, còn những công cụ khác người dùng có thể tìm hiểu sâu hơn chẳng hạn như: Bookmark with AIDA (Artificial Intelligence Design Assistant) hay Jasper AI (Trước đây là Jarvis AI), ... Các công cụ này đều mang lại nhiều lợi ích như tiết kiệm thời gian, chi phí và tối ưu hóa trải nghiệm người dùng. Tuy nhiên, mỗi công cụ đều có điểm mạnh và điểm yếu riêng, phù hợp với từng nhu cầu và đối tượng sử dụng khác nhau. Để lựa chọn công cụ phù hợp, cần xác định rõ mục tiêu thiết kế, quy mô dự án và yêu cầu kỹ thuật cụ thể.

4. LỢI ÍCH VÀ THÁCH THỨC KHI ỨNG DỤNG AI TRONG THIẾT KẾ WEBSITE

4.1. Lợi ích của việc ứng dụng AI trong thiết kế website

Tăng tốc độ và hiệu quả thiết kế

AI giúp tự động hóa hầu hết các bước trong quá trình thiết kế, từ việc tạo bố cục giao diện, lựa chọn màu sắc, đến việc tối ưu hóa nội dung và hình ảnh. Nhờ vậy, thời gian thiết kế website được rút ngắn đáng kể, giúp doanh nghiệp nhanh chóng đưa sản phẩm ra thị trường. Ví dụ, công cụ Wix ADI có thể tạo một website hoàn chỉnh chỉ trong vài phút dựa trên thông tin đầu vào của người dùng.

Tiết kiệm chi phí và nguồn lực

Trước đây, việc thiết kế website yêu cầu sự tham gia của nhiều chuyên gia như designer, lập trình viên, và chuyên viên SEO. Tuy nhiên, với AI, quy trình này có thể được tự động hóa và đơn giản hóa, giúp tiết kiệm chi phí thuê nhân sự và giảm thiểu các lỗi thiết kế thủ công. Các doanh nghiệp nhỏ và startup có thể sử dụng các công cụ như Firedrop hoặc Bookmark để tự thiết kế website một cách dễ dàng mà không cần thuê đội ngũ chuyên nghiệp.

Cá nhân hóa trải nghiệm người dùng (UX)

AI có khả năng phân tích dữ liệu người dùng và hành vi truy cập để tạo ra các trải nghiệm cá nhân hóa. Ví dụ, các công cụ AI có thể tự động gợi ý sản phẩm dựa trên lịch sử mua sắm

của khách hàng hoặc điều chỉnh bố cục trang web sao cho phù hợp với sở thích cá nhân. Điều này giúp tăng cường sự gắn kết và hài lòng của người dùng, từ đó nâng cao tỷ lệ chuyển đổi.

Tối ưu hóa SEO và nội dung thông minh

AI không chỉ giúp tối ưu hóa từ khóa và thẻ meta mà còn hỗ trợ tạo nội dung tự động chuẩn SEO. Các công cụ như Jasper AI và Writesonic giúp tạo nội dung bài viết, mô tả sản phẩm và email marketing dựa trên từ khóa mục tiêu, giúp website đạt thứ hạng cao trên công cụ tìm kiếm.

Phân tích dữ liệu và cải tiến liên tục

AI có khả năng phân tích dữ liệu lớn (Big Data) để hiểu rõ hơn về hành vi người dùng và xu hướng thị trường. Từ đó, nó có thể đề xuất các cải tiến cho thiết kế giao diện và nội dung một cách linh hoạt và liên tục, giúp website luôn phù hợp với nhu cầu của người dùng.

Cải thiện hiệu suất và bảo mật

AI có thể tối ưu hóa hiệu suất trang web bằng cách tự động điều chỉnh tốc độ tải trang, quản lý lưu lượng truy cập và phát hiện các lỗ hổng bảo mật. Điều này giúp website hoạt động ổn định và an toàn hơn trước các mối đe dọa từ bên ngoài.

4.2. Thách thức khi ứng dụng AI trong thiết kế website

Hạn chế về tính sáng tạo và độc đáo

Mặc dù AI có thể tự động hóa nhiều khâu trong thiết kế, nhưng khả năng sáng tạo và phá cách trong thiết kế vẫn còn hạn chế. Các giao diện do AI tạo ra thường dựa trên mẫu có sẵn và xu hướng phổ biến, dẫn đến sự thiếu độc đáo và dễ bị trùng lặp. Điều này có thể làm giảm tính nhận diện thương hiệu của doanh nghiệp trên môi trường số.

Phụ thuộc vào dữ liệu và thuật toán

Hiệu quả của AI phụ thuộc lớn vào chất lượng dữ liệu và thuật toán được sử dụng. Nếu dữ liệu đầu vào không đủ hoặc không chính xác, AI có thể đưa ra các gợi ý thiết kế không phù hợp hoặc sai lệch. Hơn nữa, việc cập nhật và cải tiến thuật toán AI đòi hỏi chi phí và công sức đáng kể từ phía nhà phát triển.

Thiếu linh hoạt trong tùy chỉnh nâng cao

Các công cụ AI thường cung cấp các mẫu thiết kế và tùy chỉnh cơ bản. Tuy nhiên, khi cần thực hiện các yêu cầu phức tạp hoặc độc đáo, AI vẫn chưa thể đáp ứng một cách linh hoạt như việc thiết kế thủ công bởi các chuyên gia. Điều này gây khó khăn cho các doanh nghiệp có yêu cầu đặc thù về thiết kế và chức năng trang web.

Vấn đề bảo mật và quyền riêng tư dữ liệu

AI cần thu thập và phân tích dữ liệu người dùng để cá nhân hóa trải nghiệm, điều này đặt ra vấn đề về quyền riêng tư và bảo mật dữ liệu. Nếu không quản lý tốt, thông tin cá nhân của người dùng có thể bị lạm dụng hoặc rò rỉ. Điều này đòi hỏi các doanh nghiệp phải tuân thủ chặt chẽ các quy định về bảo mật dữ liệu như GDPR hay CCPA.

Phụ thuộc vào công nghệ và chi phí đầu tư ban đầu

Việc triển khai AI trong thiết kế website yêu cầu hệ thống hạ tầng công nghệ mạnh mẽ và chi phí đầu tư ban đầu khá cao, đặc biệt là với các doanh nghiệp nhỏ và startup. Bên cạnh đó, cần có đội ngũ kỹ thuật có kỹ năng chuyên sâu để quản lý và vận hành các công cụ AI một cách hiệu quả.

Khả năng thay thế công việc của con người

Một trong những thách thức lớn nhất khi ứng dụng AI là nguy cơ thay thế công việc của các nhà thiết kế và lập trình viên truyền thống. Điều này có thể gây ra sự lo ngại về mất việc làm và làm giảm đi yếu tố sáng tạo trong ngành thiết kế website.

Ứng dụng AI trong thiết kế website mang lại nhiều lợi ích to lớn về tốc độ, hiệu quả và khả năng cá nhân hóa trải nghiệm người dùng. Tuy nhiên, bên cạnh những tiện ích đó, vẫn tồn tại nhiều thách thức liên quan đến tính sáng tạo, bảo mật dữ liệu và phụ thuộc vào công nghệ.

Để ứng dụng AI hiệu quả, các doanh nghiệp cần cân nhắc kỹ lưỡng giữa lợi ích và thách thức, đồng thời kết hợp giữa AI và sự sáng tạo của con người để tạo ra các trang web độc đáo và hiệu quả cao. Việc không ngừng cập nhật xu hướng công nghệ, đào tạo đội ngũ nhân sự và đảm bảo quyền riêng tư dữ liệu sẽ là chìa khóa giúp doanh nghiệp tận dụng tối đa sức mạnh của AI trong thiết kế website.

5. XU HƯỚNG PHÁT TRIỂN VÀ ỨNG DỤNG TƯƠNG LAI

AI đang ngày càng đóng vai trò quan trọng trong lĩnh vực thiết kế website, mang lại nhiều lợi ích và tiềm năng phát triển mạnh mẽ trong tương lai. Dưới đây là các xu hướng phát triển và ứng dụng AI trong thiết kế website có thể trở thành hiện thực trong những năm tới:

5.1. Thiết kế tự động hóa hoàn toàn (Fully Automated Web Design)

Trong tương lai, các công cụ AI có thể tự động hóa hoàn toàn quá trình thiết kế website mà không cần sự can thiệp của con người. Với sự phát triển của AI tự học (Machine Learning) và AI tạo sinh (Generative AI), việc tạo ra các giao diện độc đáo và cá nhân hóa sẽ trở nên dễ dàng hơn bao giờ hết.

5.2. Thiết kế dựa trên giọng nói (Voice-Driven Design)

Voice UI (Voice User Interface) đang dần trở thành xu hướng phổ biến nhờ sự phát triển của trợ lý ảo như Alexa, Google Assistant và Siri. Trong tương lai, AI sẽ giúp người dùng tạo và chỉnh sửa website bằng cách ra lệnh bằng giọng nói một cách trực quan và tiện lợi hơn.

5.3. Thiết kế đáp ứng cảm xúc (Emotion-Responsive Design)

AI ngày càng có khả năng nhận diện cảm xúc thông qua giọng nói, nét mặt hoặc hành vi người dùng. Trong tương lai, các website sẽ có thể tự động điều chỉnh giao diện, nội dung và màu sắc dựa trên cảm xúc của người truy cập.

5.4. Sáng tạo nội dung tự động và thông minh hơn

Generative AI đang ngày càng phát triển mạnh mẽ, mở ra khả năng tạo nội dung tự động một cách tự nhiên và phong phú hơn. Các công cụ AI như Jasper AI, ChatGPT hay Writesonic sẽ ngày càng thông minh và sáng tạo hơn trong việc tạo ra các nội dung đa dạng như bài viết blog, kịch bản video, quảng cáo và email marketing.

5.5. Phát triển công cụ AI mã nguồn mở và hợp tác sáng tạo

Xu hướng phát triển công cụ AI mã nguồn mở sẽ ngày càng phổ biến, giúp cộng đồng nhà phát triển và nhà thiết kế hợp tác sáng tạo mạnh mẽ hơn. Điều này sẽ tạo ra một hệ sinh thái đa dạng các công cụ AI tùy chỉnh và dễ dàng tích hợp vào quy trình thiết kế website.

5.6. Tích hợp AI vào các nền tảng CMS và thương mại điện tử

Trong tương lai, AI sẽ được tích hợp sâu hơn vào các nền tảng CMS (Content Management System) và thương mại điện tử như WordPress, Shopify, và Magento. Điều này giúp tự động hóa việc quản lý nội dung, tối ưu hóa SEO, và cá nhân hóa trải nghiệm người dùng trên quy mô lớn.

Xu hướng phát triển và ứng dụng AI trong thiết kế website đang mở ra nhiều tiềm năng đột phá về tốc độ, hiệu quả và khả năng cá nhân hóa trải nghiệm người dùng. Tuy nhiên, để tận dụng tối đa các xu hướng này, doanh nghiệp cần không ngừng cập nhật công nghệ, đào tạo nhân sự và đảm bảo các yếu tố bảo mật dữ liệu.

Trong tương lai gần, AI không chỉ đóng vai trò hỗ trợ mà còn trở thành người đồng hành sáng tạo cùng các nhà thiết kế, giúp tạo ra các trang web độc đáo, hấp dẫn và phù hợp với nhu cầu ngày càng cao của người dùng.

6. KẾT LUẬN

AI đang cách mạng hóa thiết kế website bằng cách tự động hóa quy trình, tối ưu hóa nội dung và cá nhân hóa trải nghiệm người dùng, giúp tăng tốc độ, hiệu quả và sự sáng tạo. Nghiên cứu này giới thiệu các công cụ AI nổi bật như Wix ADI, Bookmark và The Grid, đồng thời phân tích lợi ích và thách thức khi ứng dụng AI, bao gồm tiết kiệm chi phí nhưng đối mặt với vấn đề bảo mật và hạn chế về tính sáng tạo.

Để tận dụng tiềm năng AI, nhà thiết kế, giáo dục hay doanh nghiệp cần cân bằng giữa công nghệ và sáng tạo con người, đảm bảo bảo mật dữ liệu. AI không chỉ hỗ trợ mà còn đồng hành cùng nhà thiết kế, mở ra cơ hội phát triển mới cho ngành thiết kế website trong thời đại số hóa.

TÀI LIỆU THAM KHẢO

1. Sonali Suryakant Jadhav, Sonali Sagar Gholve, "The Impact of AI on Web Development", vol. 3, no. 8, pp. 07–12, Aug. 2024, „ IJSRMST, vol. 3, no. 8, p. 07–12, Aug 2024.
2. André Costa, Firmino Silva, José Joaquim Moreira, "Towards an AI-Driven User Interface Design for Web Applications," *Procedia Computer Science*, vol. 237, pp. 179-186, 2024.
3. N. Upadhyaya, "Artificial Intelligence in Web Development: Enhancing Automation, Personalization, and Decision-Making," *IJARSCIT*, vol. 4, no. 1, pp. 534-540, 2024.
4. E. S. Kapoor, "The Impact of Artificial Intelligence on Modern Website Design," *IJFMR*, vol. 6, no. 3, pp. 1-7, 2024.
5. Ikechukwu Innocent Umeh, Kobimdi Cordelia Umeh, "A Comparative Analysis of AI System Development Tools for Improved Outcomes," *IJSMIT*, vol. 11, no. 1, pp. 1-20, 2025.
6. K. Goldstein, "WixBlog," WixBlog, 22 1 2025. [Online]. Available: <https://www.wix.com/blog/wix-artificial-design-intelligence>. [Accessed 20 3 2025].
7. M. Rhodes, "WIRED," WIRED, 9 10 2014. [Online]. Available: <https://www.wired.com/2014/10/publishing-tool-builds-websites-powered-ai/>. [Accessed 20 3 2025].
8. S. Mircevska, "Cybernews," Cybernews, 6 3 2025. [Online]. Available: <https://cybernews.com/ai-tools/designs-ai-review/>. [Accessed 20 3 2025].

ABSTRACT

THE INTEGRATION OF AI TOOLS IN WEBSITE DESIGN: TRENDS AND DEVELOPMENT POTENTIAL

Lam Thi Hoa Mi^{1,*}

Ho Chi Minh City University of Industry and Trade

*Email: milth@huit.edu.vn

In today's rapidly advancing technological landscape, Artificial Intelligence (AI) is becoming increasingly vital in various fields, particularly website application design. This study aims to introduce modern AI integration tools currently used in website design and to analyze their benefits, challenges, and potential for development. By doing so, it provides designers, educators, and businesses with insights and guidance on how to leverage AI to optimize user experience and enhance the performance of website applications.

Keywords: Smart website design, Automation in web design, Website design.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Lâm Thị Hoa Mi

Số điện thoại liên hệ: 0904422030

AI – GIẢI PHÁP TỐI ƯU HÓA WEBSITE

Dương Thị Mộng Thùy^{1*}

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: thuydtm@huit.edu.vn

Ngày nhận bài: 12/05/2025 ; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Tối ưu hóa website tự động bằng AI (SEO AI) đang là công nghệ đột phá giúp tối ưu hóa website một cách hiệu quả và tiết kiệm thời gian. Với sự trợ giúp của trí tuệ nhân tạo (AI), SEO tự động thực hiện nhiều tác vụ phức tạp để cải thiện thứ hạng tìm kiếm hơn so với SEO truyền thống do con người thực hiện. Cách tự động phân tích, tối ưu và dự đoán xu hướng của AI ứng dụng vào SEO giúp website tăng thứ hạng [1] trên Google.

Từ khóa: Tối ưu hóa website, trí tuệ nhân tạo, thứ hạng tìm kiếm.

1. MỞ ĐẦU

Trong thời đại số hóa, website không chỉ là bộ mặt của doanh nghiệp trên internet mà còn là công cụ quan trọng giúp tiếp cận khách hàng, xây dựng thương hiệu và thúc đẩy doanh số. Một website hoạt động hiệu quả không chỉ thu hút người dùng mà còn giúp doanh nghiệp tăng cường khả năng cạnh tranh trên thị trường trực tuyến đầy sôi động. Tuy nhiên, với sự cạnh tranh khốc liệt giữa hàng triệu trang web trên toàn cầu, việc làm thế nào để website của bạn nổi bật, thu hút người dùng và đạt thứ hạng cao trên các công cụ tìm kiếm như Google vẫn là một thách thức lớn.

Khi hành vi người dùng ngày càng thay đổi và yêu cầu về trải nghiệm trực tuyến trở nên khắt khe hơn, việc tối ưu hóa website không còn đơn thuần là một tùy chọn mà đã trở thành một yếu tố sống còn đối với mọi doanh nghiệp. Người dùng không chỉ mong muốn một trang web có giao diện bắt mắt mà còn yêu cầu tốc độ truy cập nhanh, nội dung hữu ích, trải nghiệm mượt mà và đặc biệt là bảo mật cao. Một trang web tải chậm hay có nội dung không phù hợp có thể khiến khách hàng rời đi chỉ sau vài giây, làm giảm tỷ lệ chuyển đổi và ảnh hưởng tiêu cực đến doanh thu. Vì vậy, tối ưu hóa website không chỉ giúp cải thiện hiệu suất mà còn đóng vai trò quan trọng trong việc giữ chân khách hàng, nâng cao uy tín thương hiệu và tối đa hóa lợi nhuận.

Trước đây, tối ưu website là một quá trình phân tích thủ công, đòi hỏi chuyên môn kỹ thuật cao và liên tục cập nhật theo các thuật toán tìm kiếm mới nhất. Các chuyên gia SEO và quản trị web phải dành rất nhiều thời gian để thu thập dữ liệu, nghiên cứu từ khóa, tối ưu nội dung, xây dựng liên kết và kiểm tra hiệu suất trang web. Tuy nhiên, phương pháp thủ công

này không chỉ tốn nhiều công sức mà còn dễ mắc phải sai sót, đặc biệt là khi các thuật toán tìm kiếm thay đổi liên tục và ngày càng phức tạp.

Nhưng với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), việc tối ưu hóa website đã bước sang một kỷ nguyên mới – tự động, thông minh và hiệu quả hơn bao giờ hết. AI không chỉ giúp cải thiện hiệu suất trang web mà còn tối ưu hóa SEO, nâng cao trải nghiệm người dùng và tối đa hóa tỷ lệ chuyển đổi thông qua các công nghệ tiên tiến như phân tích dữ liệu, học máy (Machine Learning), xử lý ngôn ngữ tự nhiên (NLP) và tự động tối ưu nội dung. Cụ thể, AI có thể tự động thu thập và phân tích dữ liệu từ hàng triệu tìm kiếm trên Google, xác định các xu hướng mới nhất, đề xuất từ khóa tiềm năng và tối ưu nội dung sao cho phù hợp với nhu cầu của người dùng. Không chỉ dừng lại ở SEO, AI còn hỗ trợ tăng tốc độ tải trang, cá nhân hóa trải nghiệm người dùng, cải thiện thiết kế giao diện và phát hiện các vấn đề bảo mật để đảm bảo website luôn hoạt động ổn định và an toàn. Hơn nữa, AI giúp doanh nghiệp giảm thiểu sự phụ thuộc vào các phương pháp SEO truyền thống, đồng thời cung cấp các báo cáo và đề xuất tối ưu theo thời gian thực, giúp tiết kiệm thời gian và chi phí đáng kể. Nhờ đó, các doanh nghiệp có thể tập trung hơn vào việc xây dựng nội dung chất lượng, nâng cao dịch vụ và tối ưu chiến lược marketing.

Từ SEO thông minh, cá nhân hóa trải nghiệm khách hàng đến bảo mật website và phân tích dữ liệu chuyên sâu, AI đang mở ra những cơ hội mới để các doanh nghiệp nâng tầm nền tảng số của mình và đạt được lợi thế cạnh tranh trong thị trường số hóa. Vậy AI đang ứng dụng vào việc tối ưu website như thế nào? Những lợi ích cụ thể mà AI mang lại là gì? Và làm thế nào để tận dụng công nghệ này một cách hiệu quả nhất? Hãy cùng khám phá trong bài viết dưới đây!

2. SEO TỰ ĐỘNG VÀ CÁC CÔNG CỤ AI

2.1 Tính năng trong SEO tự động

Để tối ưu hóa website thì chuyên gia SEO phải tốn rất nhiều thời gian nghiên cứu từ khóa, tìm ra các mẫu và xu hướng theo từng nhóm người dùng trong kho dữ liệu lớn. Khi có sự thay đổi về thuật toán tìm kiếm, SEO phải nhanh chóng cập nhật, điều chỉnh chiến lược để đáp ứng yêu cầu mới. Để tiết kiệm chi phí, thời gian và nâng cao hiệu quả trong SEO, các doanh nghiệp sử dụng các tính năng trong công cụ SEO tự động.

2.1.1 Tự động nghiên cứu và đề xuất từ khóa

Một trong những tính năng quan trọng nhất của SEO tự động là khả năng tự động nghiên cứu và đề xuất các từ khóa phù hợp với nội dung website. Công cụ SEO tự động sẽ phân tích các yếu tố như xu hướng tìm kiếm của người dùng, mức độ cạnh tranh của từ khóa, mật độ từ khóa trên các trang web đối thủ, ngữ cảnh xuất hiện của từ khóa. Từ đó, công cụ sẽ đề xuất danh sách các từ khóa có tiềm năng cao mà doanh nghiệp nên ưu tiên tối ưu hóa.

Ưu điểm:

- Tiết kiệm thời gian nghiên cứu từ khóa thủ công
- Đề xuất được nhiều từ khóa có lưu lượng dữ liệu (traffic) tốt hơn
- Cập nhật xu hướng từ khóa nhanh chóng theo thời gian thực

Như vậy, với khả năng tự động nghiên cứu và đề xuất từ khóa, SEO tự động đã tiết kiệm được rất nhiều thời gian và công sức cho chuyên gia SEO.

2.1.2. Tự động gom nhóm từ khóa

Sau khi có danh sách từ khóa, công cụ SEO tự động sẽ tự động phân loại, gom nhóm các từ khóa dựa trên nội dung và chủ đề có liên quan. Quy trình gom nhóm từ khóa bao gồm phân tích ngữ nghĩa của từng từ khóa, so sánh và tìm ra mối liên hệ giữa các từ khóa, nhóm các từ khóa có chủ đề giống nhau, đặt tên và mô tả cho từng nhóm từ khóa, sắp xếp các nhóm từ khóa theo mức độ quan trọng.

Ưu điểm:

- Có cái nhìn tổng quan hơn về cơ cấu từ khóa
- Xác định được các chủ đề then chốt cần tối ưu hóa
- Thấy rõ mối liên hệ giữa các nhóm từ khóa khác nhau
- Có thể ưu tiên tối ưu hóa các nhóm từ khóa quan trọng hơn

Nhờ khả năng phân nhóm từ khóa tự động, người dùng có thể nhanh chóng xác định được chiến lược tối ưu hóa từ khóa cho website một cách khoa học và hệ thống hơn.

2.1.3. Tự động tạo dàn bài có chiều sâu

Sau khi nghiên cứu từ khóa, công cụ SEO tự động có thể tự động tạo ra dàn bài chi tiết cho từng từ khóa hoặc chủ đề. Quy trình tự động tạo dàn bài: xác định chủ đề chính và phụ của bài viết, nghiên cứu các khía cạnh, hướng tiếp cận với chủ đề, xác định các đối tượng, sự kiện liên quan, đề xuất các tiêu đề phụ cho mỗi khía cạnh của chủ đề, sắp xếp các tiêu đề phụ theo trình tự logic

Ưu điểm:

- Tiết kiệm thời gian xây dựng dàn bài cho các bài viết dài
- Tự động tạo được dàn bài chi tiết, đa chiều với nội dung phong phú
- Bài viết có cấu trúc rõ ràng, dễ đọc và dễ hiểu hơn

Nhờ khả năng tự động tạo dàn bài, AI có thể hỗ trợ rất tốt cho công việc viết bài chi tiết, đảm bảo độ sâu và tính logic cho bài viết.

2.1.4. Tự động viết bài chất lượng, tối ưu SEO

Ngoài tạo dàn bài, nhiều công cụ SEO tự động còn có thể tự động sinh ra nội dung bài viết dựa trên dàn bài đã có. Quy trình viết bài tự động bao gồm: phân tích chủ đề, từ khóa và dàn bài chi tiết, thu thập thông tin, dữ liệu liên quan từ nhiều nguồn, xử lý ngôn ngữ tự nhiên để sinh ra câu văn hoàn chỉnh, kiểm tra và chỉnh sửa ngữ pháp, ngữ nghĩa cho chuẩn xác, tự động tối ưu hóa bài viết về cấu trúc và từ khóa

Ưu điểm:

- Tiết kiệm thời gian so với viết bài thủ công
- Tạo ra được bài viết chất lượng cao, có độ dài lớn
- Tự động tối ưu SEO ngay trong quá trình viết bài

Với khả năng tự động viết bài, các công cụ SEO tự động thực sự trở thành “cộng sự viên” đắc lực, giúp tiết kiệm thời gian và công sức cho chuyên gia.

2.1.5. Tự động chèn backlink nội bộ

Backlink nội bộ (internal link) là liên kết giữa các trang/bài viết trong cùng một website. Việc tạo liên kết nội bộ giữa các bài viết có liên quan sẽ giúp truyền tải thương hiệu và thứ hạng tốt. Công cụ SEO tự động có thể tự động chèn các backlink nội bộ vào bài viết một cách hợp lý dựa trên nội dung. Quy trình tự động chèn backlink nội bộ gồm phân tích nội dung bài viết và từ khóa chính, tìm các bài viết khác có liên quan đến chủ đề tương tự, lựa chọn 1-2 bài viết phù hợp nhất để chèn link, chèn link vào vị trí thích hợp trong bài viết hiện tại, tự động lặp lại quy trình cho tất cả bài viết trên website

Ưu điểm:

- Tiết kiệm thời gian chèn link thủ công
- Backlink nội bộ chất lượng và liên quan chặt chẽ về nội dung
- SEO tự động có thể chèn hàng loạt backlink vào nhiều bài viết

Nhờ tự động hóa, việc tạo liên kết nội bộ trở nên nhanh chóng và hiệu quả hơn rất nhiều. Đây là một trong những tính năng hữu ích nhất của công cụ SEO tự động.

2.1.6. Tự động liên kết với các bài có nội dung tương tự trên website

Ngoài chèn backlink nội bộ, SEO tự động còn có thể tự động liên kết các bài viết có chủ đề tương tự với nhau trên cùng một website. Quy trình liên kết bài viết tự động bao gồm phân tích nội dung và chủ đề của từng bài viết, so sánh và tìm ra các bài viết có chủ đề, từ khóa liên quan, đề xuất liên kết phù hợp từ bài A đến bài B và ngược lại, chèn liên kết vào vị trí thích hợp trong mỗi bài viết, lặp lại quy trình để tạo liên kết cho tất cả các bài viết có liên quan

Ưu điểm:

- Kết nối chặt chẽ giữa các bài viết liên quan chủ đề
- Người đọc dễ dàng chuyển đổi qua lại giữa các bài viết
- SEO tự động tiết kiệm thời gian so với làm thủ công

Nhờ liên kết chặt chẽ giữa các bài viết, website sẽ có cấu trúc liên kết logic và thân thiện với người đọc hơn.

2.1.7. Tự động thêm ảnh và video phù hợp vào bài

Ảnh và video có vai trò quan trọng trong việc cải thiện trải nghiệm người dùng, tăng tương tác và thời gian lưu trữ trên trang. Do đó, việc tự động thêm hình ảnh, video minh họa phù hợp vào bài viết là một tính năng hữu ích của SEO tự động. Quy trình thêm ảnh, video tự động gồm phân tích nội dung bài viết để xác định chủ đề chính, tìm kiếm trên các kho ảnh, video stock có bản quyền phù hợp, lọc và chọn ra ảnh/video phù hợp với nội dung, chèn ảnh/video vào vị trí thích hợp trong bài viết, tự động lặp lại cho tất cả bài viết cần hình ảnh minh họa

Ưu điểm:

- Tiết kiệm thời gian tìm kiếm và chèn ảnh video
- Ảnh, video luôn phù hợp với nội dung bài viết
- Nâng cao trải nghiệm người dùng, thời gian lưu trữ trang

Nhờ SEO tự động, việc bổ sung hình ảnh và video minh họa cho bài viết trở nên dễ dàng và nhanh chóng hơn rất nhiều.

2.1.8. Tự động tối ưu SEO cho bài viết hoặc sản phẩm

Ngoài viết bài, công cụ SEO tự động còn có thể tự động tối ưu hóa bài viết để có thứ hạng cao hơn. Các tác vụ tối ưu SEO tự động bao gồm tối ưu tiêu đề và URL dựa trên từ khóa, tự động chèn từ khóa vào nội dung, kiểm tra và khắc phục sự trùng lặp nội dung, tối ưu hóa thẻ tiêu đề H1, H2, alt cho hình ảnh, tự động tăng liên kết nội bộ đến bài viết.

Ưu điểm:

- Tiết kiệm thời gian tối ưu SEO thủ công
- Tự động điều chỉnh khi có sự thay đổi về thuật toán tìm kiếm
- Giúp bài viết/sản phẩm dễ dàng đạt thứ hạng cao hơn

Với khả năng tự động tối ưu hóa bài viết, AI đã trở thành công cụ đắc lực để doanh nghiệp nâng cao thứ hạng trang web một cách dễ dàng và hiệu quả.

2.1.9. Tự động đăng bài lên các mạng xã hội để tạo backlink

Công cụ SEO tự động cũng có thể tự động đăng tải bài viết lên các mạng xã hội như Facebook, Twitter... nhằm tạo backlink có giá trị về mặt SEO. Quy trình đăng bài tự động bao gồm chọn bài viết phù hợp để chia sẻ lên mạng xã hội, tự động tạo mô tả ngắn gọn, hấp dẫn cho bài viết, đăng bài lên fanpage, group, tài khoản mạng xã hội, dùng các tool tự động tương tác, tăng engagement cho bài viết, lập lại định kỳ để duy trì sự hiện diện thường xuyên.

Ưu điểm:

- Tiết kiệm thời gian so với đăng bài thủ công
- Tự động hóa hoàn toàn quy trình đăng bài lên mạng xã hội
- Tạo lượng lớn backlink chất lượng từ các nguồn uy tín

Với tính năng này, SEO tự động giúp doanh nghiệp dễ dàng mở rộng phạm vi tiếp cận khách hàng qua nhiều kênh mạng xã hội khác nhau.

2.2 AI trong SEO tự động

AI đang thay đổi cách làm SEO truyền thống, giúp website tăng thứ hạng trên Google nhanh hơn bằng cách tự động phân tích, tối ưu và dự đoán xu hướng.

2.2.1. AI tự động phân tích và nghiên cứu từ khóa

AI có thể quét hàng triệu truy vấn để xác định từ khóa tiềm năng có độ cạnh tranh thấp nhưng traffic cao, dự đoán xu hướng tìm kiếm trong tương lai, giúp website đón đầu nhu cầu người dùng [4].

Công cụ AI phổ biến: Google Bard, ChatGPT, SurferSEO, SEMrush AI, Ahrefs AI.

2.2.2. Viết và tối ưu nội dung chuẩn SEO bằng AI

AI tạo nội dung chuẩn SEO [5], tự động chèn từ khóa, tối ưu heading, meta description.

Công cụ AI như GPT-4, Jasper AI, Writesonic, Copy.ai giúp viết bài chuẩn ngữ cảnh, tránh nhồi nhét từ khóa.

2.2.3. Tối Ưu On-Page SEO Bằng AI

AI tự động kiểm tra và đề xuất tối ưu theo tốc độ tải trang (cải thiện Core Web Vitals), cấu trúc nội dung (Heading, Schema, Internal link), tối ưu hình ảnh (giảm dung lượng, thêm ALT text).

Công cụ hỗ trợ: RankMath AI, Yoast SEO AI, NeuronWriter.

2.2.4. Xây Dựng Liên Kết (Backlink) Thông Minh

AI phân tích đối thủ và đề xuất backlink chất lượng, gợi ý guest post hoặc các trang có thể hợp tác để xây dựng backlink tự nhiên. AI tự động kiểm tra và báo cáo liên kết hỏng (broken link) để tránh ảnh hưởng SEO.

2.2.5. Phân Tích & Tối Ưu Trải Nghiệm Người Dùng (UX/UI)

AI theo dõi hành vi người dùng trên website theo các phương diện thời gian ở lại trang, tỷ lệ thoát, số lần nhấp. Từ đó, AI đưa ra dự đoán và đề xuất cải tiến giao diện, điều hướng, bố cục nội dung.

Công cụ phổ biến: Hotjar AI, Google Analytics 4 (GA4), Microsoft Clarity.

2.2.6. SEO Tự Động với AI Chatbot & Voice Search

AI giúp tối ưu nội dung cho tìm kiếm bằng giọng nói (Voice SEO), Chatbot AI giúp giữ chân người dùng lâu hơn, tăng thời gian người dùng ở lại trang lâu hơn

Công cụ: Google Assistant, Alexa SEO, ChatGPT API cho chatbot.

2.2.7. AI Dự Đoán & Theo Dõi Thứ Hạng SEO

AI dự đoán xu hướng thuật toán Google, giúp website đón đầu thay đổi, theo dõi thứ hạng từ khóa theo thời gian thực và tự động đề xuất cải thiện.

Công cụ: SEMrush Sensor, Google Search Console AI, SurferSEO [3].

AI đang giúp tự động hóa và tối ưu SEO nhanh hơn, thông minh hơn. Không chỉ giúp tiết kiệm thời gian mà còn nâng cao khả năng cạnh tranh trên công cụ tìm kiếm.

3. ỨNG DỤNG MÔ HÌNH NLP VÀO SEO TỰ ĐỘNG

Xử lý ngôn ngữ tự nhiên (NLP - Natural Language Processing) là một lĩnh vực của trí tuệ nhân tạo (AI) giúp máy tính hiểu, phân tích và tạo nội dung giống con người. Trong tối ưu hóa website, NLP đóng vai trò quan trọng trong việc phân tích nội dung, tối ưu từ khóa và cải thiện chất lượng bài viết, giúp website có thứ hạng cao hơn trên công cụ tìm kiếm.

NLP giúp đánh giá độ liên quan giữa nội dung và từ khóa mục tiêu, xác định từ khóa chính (primary keywords) và từ khóa phụ (LSI - Latent Semantic Indexing) để tối ưu bài viết một cách tự nhiên, tránh nhồi nhét từ khóa (keyword stuffing).

NLP giúp hiểu ý định tìm kiếm (search intent) của người dùng, đảm bảo bài viết cung cấp câu trả lời chính xác theo mục đích tìm kiếm (thông tin, thương mại, điều hướng, giao dịch), gợi ý từ khóa liên quan, giúp bài viết bao quát nội dung tốt hơn, tăng cơ hội xếp hạng cao trên Google.

NLP phát hiện câu văn khó hiểu, dài dòng hoặc thiếu logic để đề xuất cải tiến khả năng đọc hiểu và trải nghiệm người dùng. NLP điều chỉnh cấu trúc bài viết, sử dụng tiêu đề, đoạn văn, danh sách để tăng tính dễ đọc.

3.1. Các Mô Hình NLP Nổi Bật Trong Tối Ưu Nội Dung

Google BERT [2] (Bidirectional Encoder Representations from Transformers) giúp Google hiểu ngữ cảnh và mối quan hệ giữa các từ khóa trong một câu, tập trung vào ý định tìm kiếm, giúp website được xếp hạng dựa trên nội dung hữu ích thay vì chỉ dựa vào mật độ từ khóa.

GPT-4 (Generative Pre-trained Transformer 4) là mô hình AI tiên tiến có khả năng viết, tóm tắt, tối ưu và phân tích nội dung hiệu quả, hỗ trợ sáng tạo nội dung chất lượng cao, phù hợp với xu hướng tìm kiếm.

Surfer SEO [6] & Clearscope [7] (AI SEO NLP Tools) phân tích nội dung của các bài viết top đầu Google để đề xuất cải thiện từ khóa, cấu trúc bài viết và mức độ chuyên sâu của nội dung.

3.2. Cách Ứng Dụng NLP Để Tối Ưu Bài Viết Website

- ✓ Bước 1: Phân tích từ khóa và ý định tìm kiếm
 - Sử dụng NLP để tìm từ khóa phù hợp với mục đích tìm kiếm của người dùng.
 - Nhận diện các từ khóa tiềm năng giúp bài viết bao quát chủ đề rộng hơn.
- ✓ Bước 2: Viết nội dung theo tiêu chuẩn NLP
 - Dùng từ khóa một cách tự nhiên, không nhồi nhét.
 - Sử dụng câu văn ngắn gọn, dễ đọc, đảm bảo trải nghiệm người dùng tốt.
- ✓ Bước 3: Kiểm tra và tối ưu bài viết bằng công cụ NLP
 - Dùng Surfer SEO, Clearscope, Grammarly AI để phân tích chất lượng nội dung.
 - Điều chỉnh mật độ từ khóa, cấu trúc tiêu đề, liên kết nội bộ để tăng khả năng xếp hạng.

Ứng dụng NLP trong tối ưu hóa website giúp cải thiện nội dung, tối ưu từ khóa và nâng cao trải nghiệm người dùng. Việc sử dụng NLP cùng với AI không chỉ giúp doanh nghiệp tăng thứ hạng SEO mà còn xây dựng nội dung chất lượng cao, phù hợp với nhu cầu tìm kiếm của khách hàng.

4. KẾT LUẬN

Trí tuệ nhân tạo (AI) đang cách mạng hóa quá trình tối ưu hóa website, giúp doanh nghiệp nâng cao hiệu suất, cải thiện thứ hạng SEO và mang lại trải nghiệm người dùng tốt hơn. Từ việc phân tích dữ liệu thông minh, tối ưu từ khóa, cải thiện nội dung, đến tăng tốc độ tải trang và bảo mật web, AI đã chứng minh vai trò quan trọng trong việc nâng tầm nền tảng số của doanh nghiệp.

Thay vì những phương pháp tối ưu thủ công tốn nhiều thời gian và công sức, AI mang đến sự tự động hóa, chính xác và hiệu quả vượt trội. Với khả năng học hỏi từ dữ liệu thực tế, dự đoán xu hướng và đưa ra các đề xuất tối ưu liên tục, AI giúp website không chỉ đạt thứ hạng cao trên công cụ tìm kiếm mà còn giữ chân khách hàng, tăng tỷ lệ chuyển đổi và tối đa hóa lợi nhuận.

Trong bối cảnh cạnh tranh ngày càng gay gắt trên không gian số, việc ứng dụng AI vào tối ưu hóa website không còn là một xu hướng, mà là một bước tiến tất yếu. Doanh nghiệp nào biết khai thác AI đúng cách sẽ có lợi thế vượt trội, tạo ra sự khác biệt và phát triển bền vững trong thời đại số hóa.

TÀI LIỆU THAM KHẢO

1. Google Search Central. (2024). How Google Search Works: Ranking Results with AI & Machine Learning.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics.
3. Moz. (2024). Artificial Intelligence and the Future of SEO.
4. SEMrush. (2023). How AI is Transforming Keyword Research and SEO Strategies.
5. <https://aznet.vn/seo-ai.html>
6. Surfer SEO. (2024). AI-Powered Content Optimization & On-Page SEO.
7. Clearscope. (2023). Using AI & NLP for Smarter SEO Content Strategies.

ABSTRACT

AI - SEARCH ENGINE OPTIMIZATION SOLUTION

Author: Duong Thi Mong Thuy

Ho Chi Minh City University of Industry and Trade

Email: thuydtm@huit.edu.vn

Automated Website Optimization with AI (SEO AI) is a groundbreaking technology that optimizes websites efficiently while saving time. With the assistance of Artificial Intelligence (AI), automated SEO performs complex tasks to improve search rankings more effectively than traditional human-executed SEO. AI's ability to automatically analyze, optimize, and predict trends in SEO applications helps websites rank higher on Google.

Keywords: Website Optimization, Artificial Intelligence, Search Rankings

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Dương Thị Mộng Thùy

Số điện thoại liên hệ: 0909900545

XÂY DỰNG HỆ THỐNG KHUYẾN NGHỊ CHỌN MỸ PHẨM DỰA TRÊN MÔ HÌNH HỌC MÁY

Nguyễn Thị Bích Ngân^{1,*}, Nguyễn Thị Bích Tuyền¹, Đinh Thị Tâm¹

² Khoa Công nghệ thông tin, Trường Công Thương Thành phố Hồ Chí Minh

*Email: nganntb@huit.edu.vn, 2001203004@hufi.edu.vn, tamdt@huit.edu.vn

Ngày nhận bài 28/03/2025 ; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Hiện nay, mỹ phẩm có vô số loại sản phẩm từ bình dân đến hàng cao cấp, với rất nhiều công dụng khác nhau dành cho nhiều đối tượng người dùng. Bên cạnh đó, với sự phát triển của các hệ thống thương mại điện tử, người dùng rất dễ dàng tiếp cận các sản phẩm này. Vậy một vấn đề đặt ra là làm sao giúp cho người dùng có thể tìm được các sản phẩm mỹ phẩm phù hợp với đặc điểm cá nhân, và hỗ trợ người dùng ra quyết định phù hợp với nhu cầu của họ. Bài báo này sẽ tập trung vào việc nghiên cứu các mô hình học máy K-Means Clustering và Support Vector Machine (SVM) để huấn luyện dữ liệu đưa ra quyết định, từ đó khuyến nghị, tư vấn gợi ý các sản phẩm mỹ phẩm phù hợp với người tiêu dùng ở Việt Nam. Nghiên cứu cũng xây dựng một ứng dụng website để minh họa ứng dụng khuyến nghị sản phẩm mỹ phẩm dựa trên kết quả học máy đã huấn luyện được. Bộ dữ liệu thực nghiệm trong nghiên cứu này là bộ dữ liệu về mỹ phẩm lấy từ hệ thống Sephora¹ (hệ thống chuỗi cửa hàng bán lẻ mỹ phẩm và sản phẩm làm đẹp nổi tiếng trên toàn thế giới) đã cho kết quả thực nghiệm đạt độ chính xác 92.4%.

Từ khóa: học máy, K-Means Clustering, Support Vector Machine, hệ thống khuyến nghị tư vấn, mỹ phẩm

1. MỞ ĐẦU

1.1. Đặt vấn đề

Xã hội tiến bộ, cuộc sống con người được nâng cao, công nghiệp mỹ phẩm phát triển,... và nhiều lý do khác đã làm thúc đẩy nhu cầu làm đẹp, sử dụng mỹ phẩm của mọi người tăng cao. Cụ thể, trong những năm gần đây, ngành công nghiệp mỹ phẩm đã chứng kiến sự phát triển vượt bậc với sự ra đời của hàng loạt sản phẩm từ bình dân đến cao cấp của rất nhiều nhãn hàng hoặc các thương hiệu nổi tiếng trên thế giới; người tiêu dùng không chỉ quan tâm đến việc làm đẹp mà còn đặc biệt chú trọng đến việc chăm sóc và bảo vệ sức khỏe của làn da; việc sử dụng mỹ phẩm không chỉ dành cho phái đẹp, người trẻ tuổi mà còn trở

¹ <https://www.sephora.com>

nên nhu cầu thiết yếu của phái mạnh và cả người lớn tuổi. Từ những lý do cung và cầu của xã hội nói trên, sự phong phú và đa dạng của các sản phẩm mỹ phẩm là tất yếu. Tuy nhiên nó cũng tạo ra một thách thức lớn cho người tiêu dùng trong việc lựa chọn sản phẩm phù hợp với loại da và nhu cầu cá nhân. Hiện nay, các hệ thống thương mại điện tử phát triển không chỉ trên thế giới mà ở Việt Nam cũng rất mạnh mẽ, người tiêu dùng có thể mua bất kỳ sản phẩm mỹ phẩm nào một cách dễ dàng.

Song song với sự phát triển của các hệ thống thương mại điện tử, ở nước ngoài, các hệ thống khuyến nghị tư vấn chăm sóc khách hàng ứng dụng học máy đã và đang phát triển mạnh mẽ. Chúng ta có thể kể đến các hệ thống phổ biến và nổi tiếng trong lĩnh vực mỹ phẩm sau:

- Sephora (Mỹ), chuỗi bán lẻ mỹ phẩm quốc tế, đã triển khai hệ thống "Virtual Artist"² sử dụng thực tế tăng cường (Augmented Reality - AR) và trí tuệ nhân tạo (Artificial Intelligence – AI) để tư vấn mỹ phẩm (son môi, phấn nền, kẻ mắt). Hệ thống cho phép người dùng tải ảnh tự chụp (selfie) lên ứng dụng, hệ thống phân tích đặc điểm khuôn mặt (màu da, hình dạng) bằng thị giác máy tính. Sau đó, hệ thống dùng thuật toán học máy kết hợp với dữ liệu từ hàng triệu khách hàng để đề xuất sản phẩm phù hợp [1].

- L'Oréal (Pháp) phát triển Perso, một thiết bị AI tại nhà kết hợp hệ thống khuyến nghị để tạo mỹ phẩm cá nhân hóa (kem dưỡng, son môi, kem nền). Hệ thống sử dụng cảm biến đo lường tình trạng da (độ ẩm, lỗ chân lông) và môi trường (độ ẩm, tia UV). Sau đó dùng thuật toán học sâu phân tích dữ liệu và đưa ra quyết định gợi ý sản phẩm theo nhu cầu tức thời (real time) cho người dùng [2].

- Ứng dụng ModiFace (Canada - thuộc L'Oréal) sử dụng công nghệ AR được L'Oréal mua lại, ứng dụng trong tư vấn mỹ phẩm qua ứng dụng di động và gương thông minh tại cửa hàng. Ứng dụng phân tích hình ảnh thời gian thực để tư vấn sản phẩm dựa trên màu da, tuổi tác, và xu hướng làm đẹp. Dữ liệu của hệ thống được kết hợp từ dữ liệu người dùng với cơ sở dữ liệu sản phẩm toàn cầu [3].

Ở Việt Nam, sàn thương mại điện tử Shopee sử dụng hệ thống khuyến nghị dựa trên hành vi người dùng và học máy để đề xuất sản phẩm, bao gồm mỹ phẩm. Hệ thống này phân tích lịch sử tìm kiếm, mua sắm, và đánh giá của người dùng để đưa ra gợi ý cá nhân hóa. Hệ thống sử dụng thuật toán lọc cộng tác (Collaborative Filtering) kết hợp với học sâu để dự đoán sở thích mà gợi ý sản phẩm cho người dùng. Dữ liệu dùng cho hệ thống được tích hợp từ dữ liệu từ các chương trình khuyến mãi (flash sale, mã giảm giá) để tăng tính hấp dẫn [4].

Tuy nhiên, vấn đề đặt ra là các nền tảng tư vấn khuyến nghị này ở Việt Nam vẫn chưa có phổ biến, đặc biệt là các hệ thống tìm kiếm dạng khuyến nghị, tư vấn tự động để đề xuất cho khách hàng những sản phẩm phù hợp với loại da và những yếu tố khác (như giới tính, độ tuổi, màu da, màu tóc, đặc điểm da...) một cách có cơ sở khoa học.

Xuất phát từ nguyên nhân này, nhóm chúng tôi đã nghiên cứu và đề xuất ứng dụng học máy để xây dựng hệ thống khuyến nghị tư vấn mỹ phẩm cho người Việt, hệ thống này có thể được tích hợp vào các trang thương mại điện tử mua bán mỹ phẩm.

1.2. Các nghiên cứu liên quan

Qua quá trình khảo sát các nghiên cứu liên quan vấn đề này ở trong và ngoài nước, chúng tôi ghi nhận được như sau:

² <https://play.google.com/store/apps/details?id=com.sephora.digital&hl=en>

Năm 2022, Fadly và cộng sự [5] đã có nghiên cứu đánh giá của khách hàng trên trang web của Sephora, bài nghiên cứu đề cập phương pháp thực hiện quá trình thu thập dữ liệu và xử lý dữ liệu từ trang Sephora. Sử dụng thuật toán học máy, huấn luyện mô hình Naive Bayes. Kết quả cho thấy mô hình đạt độ chính xác 94.7% sau khi áp dụng kỹ thuật cross-validation, giúp Sephora hiểu rõ hơn về cảm nhận của khách hàng đối với sản phẩm và cải thiện chiến lược kinh doanh.

Năm 2022, trong một nghiên cứu của tác giả Liew Yi Kei sử dụng tập dữ liệu được thu thập từ trang Sephora, tập trung vào việc phát triển một chatbot tư vấn sản phẩm chăm sóc da dựa trên phân tích cảm xúc từ đánh giá người dùng thực tế và nhận diện loại da bằng CNN (Convolutional Neural Network). Chatbot được xây dựng trên nền tảng Telegram, sử dụng mô hình DistilBERT để phân tích cảm xúc và mô hình Factorization Machine để đề xuất sản phẩm, nhằm mục đích cung cấp khuyến nghị sản phẩm cá nhân hóa và tiện lợi hơn so với các phương pháp truyền thống. Nghiên cứu này trình bày chi tiết về phương pháp luận, triển khai hệ thống, đánh giá hiệu quả và so sánh với các chatbot tương tự khác mang lại hiệu quả cao trên tập dữ liệu thu thập từ Sephora về mỹ phẩm với các thuộc tính như tên sản phẩm, rating (xếp hạng) màu mắt, màu da, màu tóc, đặc điểm nổi bật,... [6]

Năm 2023, lĩnh vực da liễu có nhiều tiến bộ qua việc sử dụng trí tuệ nhân tạo, tập trung cải thiện da vì mục đích thẩm mỹ, Pat Vatiwutipong và cộng sự đã trích dẫn các bài báo được xuất bản từ năm 2018 đến năm 2023, trong đó có tổng cộng 63 ấn phẩm được coi là có liên quan dựa trên các tiêu chí đưa vào đã được thiết lập, được chia thành 5 loại dựa trên lĩnh vực sử dụng. Đó là phát triển sản phẩm mỹ phẩm, đánh giá da, chuẩn đoán tình trạng da, khuyến nghị điều trị, và dự đoán kết quả điều trị. Bài viết này không chỉ cung cấp những hiểu biết sâu sắc có giá trị cho các nhà nghiên cứu muốn khám phá các lĩnh vực mới liên quan đến y học thẩm mỹ mà còn cung cấp hướng dẫn áp dụng cho những người đang tìm cách triển khai công nghệ AI và các kỹ thuật học máy, học sâu để giải quyết những thách thức trong thế giới thực trong dịch vụ thẩm mỹ [7].

Năm 2022, Muskan Chaurasia và cộng sự đã sử dụng các kỹ thuật phân tích văn bản, xử lý ngôn ngữ tự nhiên cho các thành phần mỹ phẩm để phù hợp với loại da, gán giá trị nhị phân cho các thành phần có trong mỗi sản phẩm, sử dụng tính toán độ tương đồng Cosine để đề xuất 5 sản phẩm đầu có tính tương tự mang tính cá nhân hoá người dùng. Tuy nhiên, hạn chế của hệ thống là mong muốn người dùng nhập vào sản phẩm ban đầu làm đầu vào, người dùng phải tìm kiếm riêng từng loại sản phẩm và không có đánh giá độ chính xác cho hệ thống [8].

Khi khảo sát ở Việt Nam, chúng tôi thấy có ít các nghiên cứu liên quan vấn đề này. Năm 2016, một trong những nghiên cứu dành cho nam giới trong lĩnh vực mỹ phẩm, trong đó có nghiên cứu của Mai Ngọc Khương and Hoàng Thị Mỹ Duyên [9] cho thấy mức độ chú ý đến sức khỏe làn da, tuổi tác, quá trình lão hoá, hình ảnh bản thân, nhận thức về sử dụng các sản phẩm chăm sóc của nam giới ngày càng cao. Từ đó có căn cứ để góp phần cho quyết định mua sản phẩm chăm sóc da của nam giới tại Việt Nam, trong bối cảnh thị trường sản phẩm chăm sóc da nam ngày càng phát triển như hiện nay. Trong bài nghiên cứu của tác giả Nguyễn Thái Nghe [10] giới thiệu sơ lược về hệ thống gợi ý và những vấn đề liên quan, các nhóm kỹ thuật hiện đang được ứng dụng trong các hệ thống khuyến nghị, từ đó trình bày chi

tiết một vài kỹ thuật cho kết quả dự đoán tin cậy - như kỹ thuật phân rã ma trận (matrix factorization) và những biến thể của nó. Bên cạnh đó, tác giả cũng giới thiệu các lĩnh vực ứng dụng chính, các giải thuật chính của Recommender Systems (nhóm giải thuật lọc cộng tác, nhóm giải thuật lọc trên nội dung, nhóm lai ghép, nhóm kỹ thuật không cá nhân hoá) trong các công trình kèm theo là các kết quả thực nghiệm và hệ thống minh họa cho từng lĩnh vực. Đồng thời, nêu ra các phương pháp đánh giá độ tin cậy của giải thuật đã đề cập, đánh giá hiệu quả của hệ thống gợi ý qua các độ đo: Precision, Recall, Accuracy.

Trên thực tế, không khẳng định rõ việc chọn lựa phương pháp nào là phù hợp nhất, nhưng khi sử dụng hệ thống hỗ trợ tư vấn, gợi ý cho người dùng, ưu tiên cá nhân hoá đặc điểm riêng của khách hàng để khách hàng hiểu rõ đặc tính của riêng mình sẽ mang lại trải nghiệm tốt cho người dùng hơn. Đối với các hệ thống khuyến nghị tư vấn, một trong những khuyết điểm của phương pháp lọc theo nội dung (Content based filtering - CBF) là khó khăn trong việc thu thập thông tin, trong khi phần lớn các mô hình dựa trên lọc cộng tác chỉ cần các thông tin (user id, item id, feedback) là có thể hoạt động tốt.

Sau khi tìm hiểu những nghiên cứu trên, kết hợp với bộ dữ liệu thu thập được từ trang Sephora, đây là một bộ dữ liệu phong phú và đáng tin cậy đã được các tác giả công bố từ các công trình nghiên cứu [5,6], chúng tôi đã nghiên cứu thực hiện tạo hệ thống khuyến nghị dựa trên nội dung sản phẩm mỹ phẩm là thành phần hoạt chất của mỹ phẩm và công dụng khuyến nghị của sản phẩm từ bộ dữ liệu, sử dụng phương pháp lọc theo nội dung để đánh giá sự tương đồng về thành phần của mỹ phẩm. Hệ thống cung cấp tính năng lọc dựa trên nội dung dựa vào việc người dùng nhập sản phẩm ban đầu để đề xuất mang tính cá nhân hoá hơn.

2. ĐỀ XUẤT MÔ HÌNH

Trong nghiên cứu này, chúng tôi nghiên cứu và đề xuất phương pháp học máy kết hợp giữa kỹ thuật gom cụm K-Means Clustering, phân lớp SVM (Support Vector Machine) và lọc dựa trên nội dung văn bản Content-based Filtering (CBF).

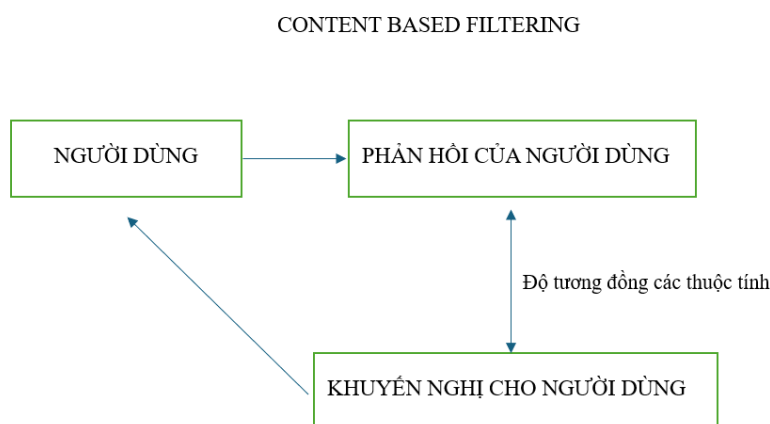
Support Vector Machine - SVM là một thuật toán phân loại mạnh mẽ được sử dụng rộng rãi trong các bài toán phân loại và hồi quy. SVM tìm cách xác định một siêu phẳng tối ưu để phân chia dữ liệu thành các lớp khác nhau. Mục tiêu của SVM là tìm ra siêu phẳng mà khoảng cách (margin) giữa nó và các điểm dữ liệu gần nhất thuộc các lớp khác nhau là lớn nhất. Các điểm dữ liệu gần với siêu phẳng này được gọi là các "vector hỗ trợ" (support vectors), chúng đóng vai trò quyết định trong việc xác định vị trí của siêu phẳng [11].

Với thuật toán *K-Means Clustering*, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn) [12].

Content-Based Recommendation System: Gợi ý các sản phẩm dựa vào hồ sơ (profiles) của người dùng hoặc dựa vào nội dung/thuộc tính (attributes) của những sản phẩm (thành

phần sản phẩm, đặc điểm nổi bật) tương tự như sản phẩm mà người dùng đã chọn. Hình 1 minh họa luồng của hệ thống lọc dựa theo nội dung.

Bài toán xây dựng hệ thống khuyến nghị có một phần liên quan đến bài toán lọc theo nội dung dựa trên dữ liệu đầu vào mà người dùng cung cấp. Vì vậy mục tiêu của nghiên cứu này bao gồm:



Hình 1. Minh họa luồng của hệ thống Content Based Filtering

- Xử lý và phân tích dữ liệu cho phù hợp với yêu cầu của hệ thống. Áp dụng phương pháp lọc dựa trên nội dung Content-based Filtering, sử dụng thuật toán K-Means Clustering và SVM để phân cụm và phân lớp sản phẩm theo đặc điểm người dùng.
- Tính hiệu suất và độ chính xác: Mục tiêu là đạt được độ chính xác cao khi áp dụng mô hình đề xuất sử dụng thuật toán SVM để phân lớp sản phẩm theo đặc điểm người dùng, so sánh giữa các thuật toán. Quá trình thực nghiệm trên bộ dữ liệu đặc trưng đã chuẩn hóa từ các nghiên cứu đã được công bố.
- Hiển thị chi tiết thông tin về các thành phần trong sản phẩm cho người dùng, trực quan hoá giao diện người dùng.

Từ mục tiêu trên, chúng tôi xây dựng hệ thống khuyến nghị tư vấn với kiến trúc hệ thống hoạt động như sau (thể hiện trong Hình 2):

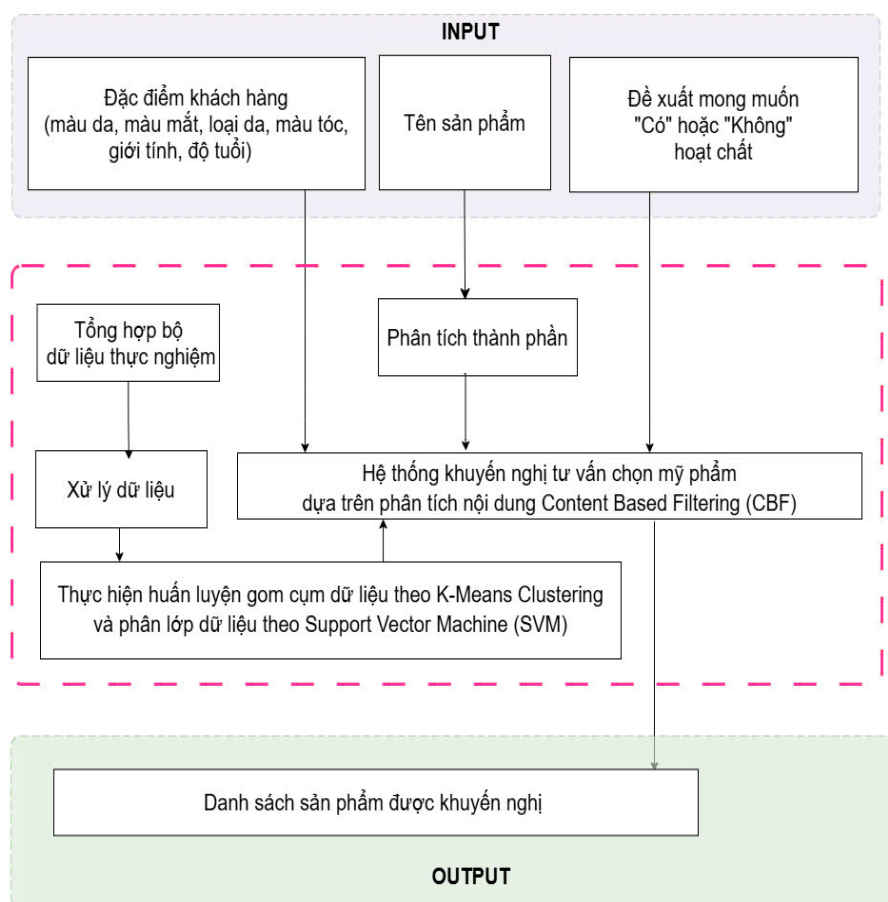
(1). Dữ liệu đầu vào từ người dùng cung cấp có thể là một trong ba dạng:

- + Đặc điểm người dùng (màu da, màu mắt, loại da, màu tóc, giới tính, độ tuổi).
- + Tên sản phẩm mỹ phẩm.
- + Đề xuất mong muốn có hoặc không các hoạt chất cụ thể.

(2). Ngoài ra hệ thống có bước tổng hợp và xử lý dữ liệu để tạo dữ liệu huấn luyện cho mô hình học máy. Quá trình huấn luyện bao gồm gom cụm dữ liệu theo K-Means Clustering và phân lớp dữ liệu theo SVM.

(3) Sau khi tiếp nhận dữ liệu đầu vào, tùy theo dạng của dữ liệu mà hệ thống sẽ phân tích thành phần hoặc chuyển trực tiếp đến phần phân tích lọc dựa trên nội dung CBF. Chức năng phân tích nội dung dựa trên kết quả huấn luyện của quá trình gom cụm và phân lớp trước đó để đưa ra kết quả phù hợp.

(4) Cuối cùng, hệ thống đưa ra danh sách các sản phẩm khuyến nghị dựa trên yêu cầu đầu vào của người dùng.



Hình 2. Kiến trúc đề xuất của hệ thống khuyến nghị tư vấn mỹ phẩm.

Từ kiến trúc chung của hệ thống khuyến nghị tư vấn chọn mỹ phẩm đã được mô tả ở trên, để huấn luyện dữ liệu, nhằm tối ưu hoá độ chính xác trong việc thực hiện khuyến nghị, chúng tôi xây dựng mô hình xử lý huấn luyện dữ liệu để đưa ra khuyến nghị sản phẩm theo hai thuật toán là K-means Clustering và SVM như trong hình 3.

2.1. Mô hình 1:

- Bước 1: Làm sạch dữ liệu bỏ các dòng dữ liệu không chính xác, sai định dạng.
- Bước 2: Phân cụm dữ liệu theo từng đặc điểm khách hàng với các đánh giá trên 4 sao. Từ mẫu dữ liệu đánh giá sản phẩm của khách hàng, chúng tôi phân cụm cho các đặc

điểm khách hàng và điểm đánh giá khách hàng (rating), như: phân cụm theo màu da, loại da, màu mắt, màu tóc, giới tính, độ tuổi và đánh giá của từng thuộc tính.

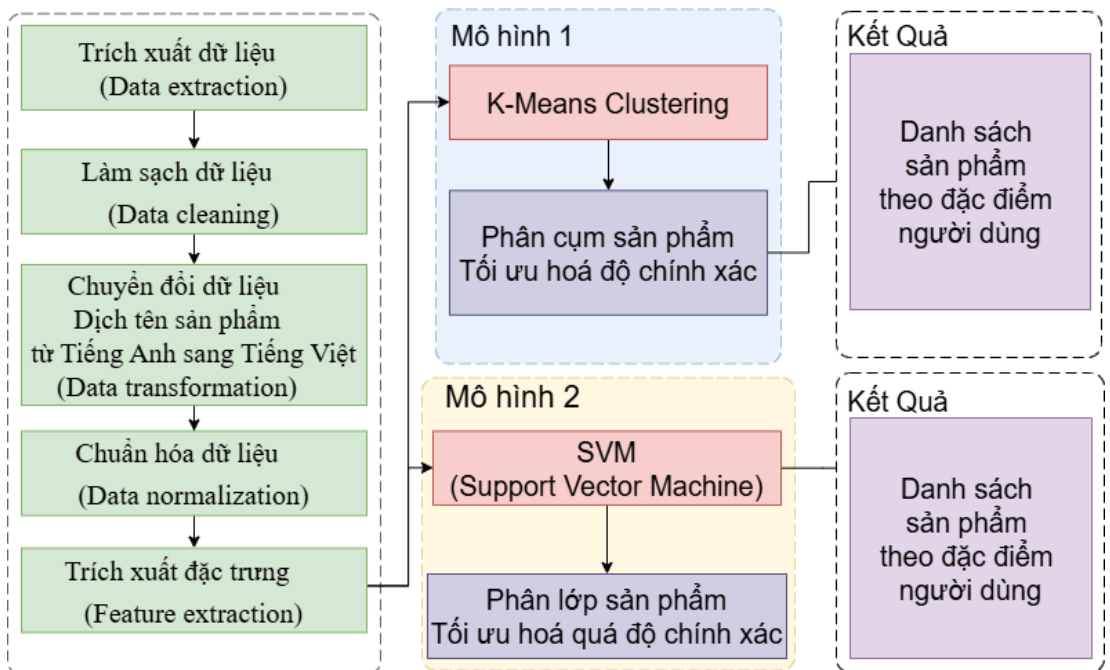
- Bước 3: Cấu hình các tham số:

Với thuật toán K-means Clustering sẽ cấu hình như sau:

+ Chọn dữ liệu ban đầu: Đầu tiên, số cụm (k) mà tập dữ liệu sẽ được chia ra theo các giá trị của đặc điểm người dùng. Trung tâm của các cụm được đặt sao cho cách xa nhau nhất có thể nhằm tăng độ chính xác của kết quả.

+ Tính khoảng cách: Thuật toán sau đó tính khoảng cách từ mỗi đối tượng trong tập dữ liệu đến từng cụm. Tọa độ nhỏ nhất sẽ xác định đối tượng thuộc về cụm nào.

+ Tái tính toán trung tâm cụm: Trung tâm của cụm được tính lại dựa trên trung bình tọa độ của tất cả các đối tượng trong cụm. Quá trình này lặp lại từ bước đầu tiên, nhưng với trung tâm cụm mới đã được tái tính toán. Các lần lặp sẽ tiếp diễn cho đến khi đạt được điều kiện dừng, khi trung tâm cụm không còn thay đổi hoặc thay đổi không đáng kể giữa các lần lặp.



Hình 3. Mô hình học máy đề xuất cho quá trình xử lý huấn luyện dữ liệu để đưa ra khuyến nghị sản phẩm theo hai thuật toán là K-means Clustering và SVM

2.2. Mô hình 2:

- Bước 1: Làm sạch dữ liệu bỏ các dòng dữ liệu không chính xác, sai định dạng.
- Bước 2: Chỉnh sửa phân lớp dữ liệu theo từng đặc điểm khách hàng.
- Bước 3: Cấu hình tham số:

Phân lớp dữ liệu bằng thuật toán SVM:

- + Mô hình này phân lớp sản phẩm theo đánh giá của khách hàng với các đặc điểm khách hàng. Lớp này được thiết kế từ các đặc điểm khách hàng có các thuộc tính sau: Sản phẩm, màu da, màu mắt, loại da, màu tóc, giới tính, độ tuổi.
- + Để có các giá trị của các tham số tối ưu nhất thì sẽ chạy vòng lặp thay đổi giá trị của các tham số với độ chính xác cao nhất.

3. MÔ TẢ VÀ XỬ LÝ DỮ LIỆU

3.1. Mô tả dữ liệu

Dữ liệu thực nghiệm của nghiên cứu này sử dụng từ nghiên cứu của Fadly và cộng sự [5] kết hợp tổng hợp dữ liệu của dự án “Sephora EDA and Sentiment Analysis using Pytorch” của Muladi Muhamad và cộng sự được công bố trên Kaggle [15]. Dữ liệu gồm có 8.494 sản phẩm bao gồm tên sản phẩm, thương hiệu, giá cả, thành phần, xếp hạng, với một số thông tin khác và 1.301.136 dòng đánh giá sản phẩm của khách hàng chứa thông tin đặc điểm như màu da, loại da, màu tóc màu mắt, thời gian đánh giá. Trong số nhiều danh mục sản phẩm chăm sóc cá nhân, tập trung vào các sản phẩm chăm sóc da. Bộ dữ liệu sản phẩm gồm các thông tin như: loại sản phẩm, tên sản phẩm, thương hiệu, giá cả, thứ hạng, loại da, thành phần hoá học và đặc điểm nổi bật của từng sản phẩm.

3.2. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu gồm các bước chính sau:

- Loại bỏ giá trị thiếu: điều chỉnh các dữ liệu thiếu, và loại trừ các dữ liệu dễ gây nhiễu để đảm bảo rằng dữ liệu sạch và đáng tin cậy.
- Dịch tên sản phẩm từ Tiếng Anh sang Tiếng Việt, sử dụng thư viện “googletrans” của Python.
- Chọn các từ gốc trong danh sách thành phần của từng sản phẩm trước khi đạt mục tiêu cuối cùng là so sánh thành phần. Tạo danh sách các thành phần của mỹ phẩm trong cơ sở dữ liệu. Hệ thống sử dụng mã hoá nhị phân. Điều này giúp mô hình máy học làm việc với dữ liệu số thay vì dữ liệu văn bản, tăng tốc độ xử lý và hiệu quả huấn luyện của mô hình.
- Xây dựng ma trận về thành phần của mỹ phẩm. Tổng số mục mỹ phẩm trong dữ liệu được biểu thị thành các hàng trong ma trận, tổng số thành phần được thể hiện bằng các cột trong ma trận. Nếu một thành phần xuất hiện trong một sản phẩm mỹ phẩm thì giá trị sẽ là 1, nếu không thì là 0 (ví dụ minh họa ở bảng 1).

Bảng 1. Minh họa ma trận thành phần của danh sách mỹ phẩm

	Water	Propanediol	Glycerine	EthyleneGlycol	Betaine	...
OLAY 7in1 Total Effects	1	0	1	0	0	
Garnier Water Gel	1	1	0	1	1	

...

- Vector hoá tất cả sản phẩm theo thành phần hoá học. Tạo ra đồ thị trên mặt phẳng hai chiều. Khoảng cách giữa các điểm sẽ cho thấy điểm tương đồng giữa các sản phẩm làm đẹp trong dữ liệu. Dữ liệu được vector hoá thành tọa độ mặt phẳng. Tính toán độ tương đồng bằng cách tính độ tương đồng Cosine Similarity.

- Bước tiếp theo là chia tách dữ liệu thành tập huấn luyện và tập kiểm tra nhằm đảm bảo rằng mô hình được đánh giá một cách chính xác. Trong nghiên cứu này, dữ liệu lần lượt được chia theo tỷ lệ 50/50, 60/40, 70/30, 80/20, trong đó phần lớn dữ liệu được sử dụng để huấn luyện mô hình và phần còn lại được sử dụng để kiểm tra hiệu suất của mô hình. Các mẫu dữ liệu được sắp xếp ngẫu nhiên trước khi chia để đảm bảo rằng cả hai tập dữ liệu đều có phân bố tương tự về các đặc trưng và nhãn. Tập huấn luyện được sử dụng để xây dựng và tối ưu hóa các tham số của mô hình, trong khi tập kiểm tra hoàn toàn độc lập với tập huấn luyện và được sử dụng để đánh giá hiệu suất của mô hình sau khi quá trình huấn luyện hoàn tất.

4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Môi trường thực nghiệm

Các thực nghiệm của nghiên cứu này được thực hiện trên máy tính với cấu hình như bảng 2

Bảng 2. Cấu hình máy thực thi thực nghiệm của nghiên cứu

Phần cứng	Cấu hình
Loại máy	Laptop – Dell latitude
Hệ điều hành	Windows 11 Pro
CPU	Intel core i5 CPU 2.21 GHz
RAM	8 GB
Ổ Cứng	SSD 250 GB
GPU	Không

4.2. Đánh giá kết quả

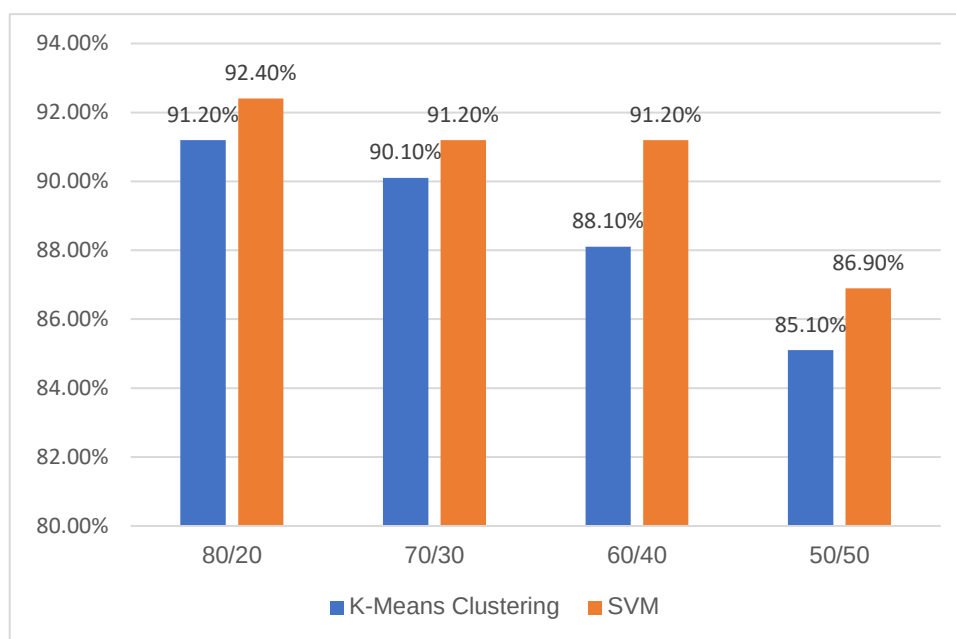
Bảng 3. Kết quả thực nghiệm theo các tỷ lệ chia dữ liệu của 2 thuật toán K-Means Clustering và SVM

Dữ liệu huấn luyện (%)	Dữ liệu thử nghiệm (%)	Độ chính xác của K-Means Clustering (%)	Độ chính xác của SVM (%)
80	20	91.2	92.4
70	30	90.1	91.2
60	40	88.1	91.2
50	50	85.1	86.9

Bộ dữ liệu được chia thành hai phần với tỷ lệ 80-20, trong đó 80% được sử dụng để huấn luyện và 20% dành cho thử nghiệm. Trong quá trình thực nghiệm trên tập dữ liệu huấn luyện, các thuật toán sẽ được đánh giá dựa trên các chỉ số đo lường hiệu suất và độ chính xác. Kết quả thu được sẽ được so sánh để xác định thuật toán hoạt động tốt nhất. Dựa vào kết

qua so sánh này, chúng ta có thể lựa chọn thuật toán phù hợp nhất để đưa ra gợi ý sản phẩm tối ưu.

Khi chạy thực nghiệm trên ứng dụng thì thuật toán SVM theo tỉ lệ 80:20 là mang lại độ chính xác cao nhất: 92.4%.



Hình 4. Biểu đồ thể hiện kết quả thực nghiệm theo các tỷ lệ chia dữ liệu của 2 thuật toán K-Means Clustering và SVM

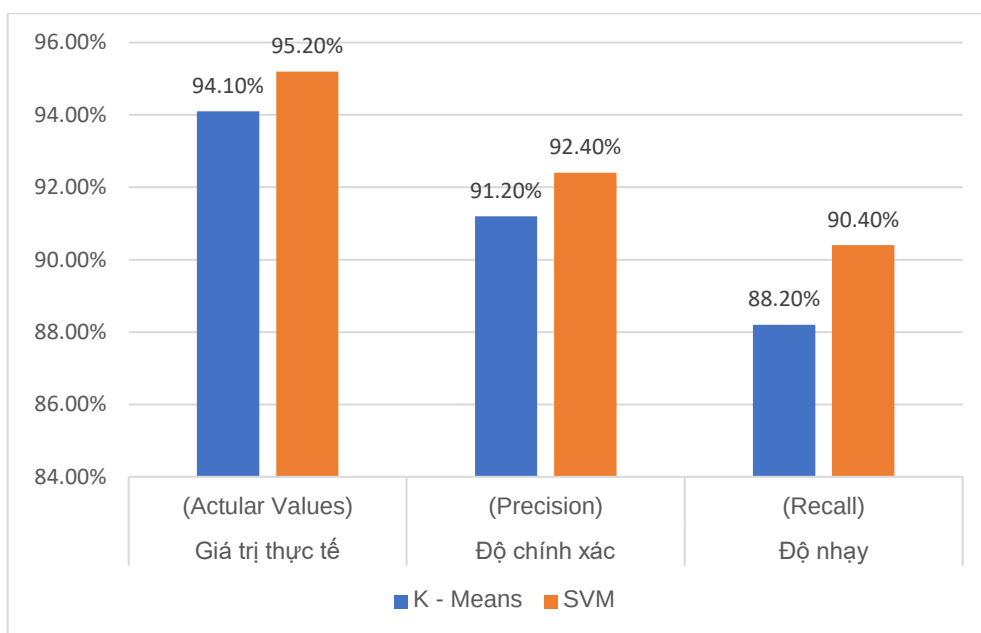
Sau khi thực nghiệm, thuật toán SVM mang lại độ chính xác cao nhất: 92.4%, độ nhạy là 90.4%.

Bảng 4. So sánh các độ đo trên kết quả của 2 thuật toán K-Means Clustering và SVM

Thuật toán	Giá trị thực tế (Actual Values)	Độ chính xác (Precision)	Độ nhạy (Recall)
K - Means	94.10%	91.20%	88.20%
SVM	95.20%	92.40%	90.40%

Hệ thống sử dụng hai thuật toán K-Means Clustering và SVM để huấn luyện dữ liệu. Kết quả cho thấy thuật toán SVM đạt độ chính xác 92.4% và độ nhạy 90.4%. Do đó, hệ thống đề xuất đã áp dụng SVM để cập nhật kết quả phân lớp dữ liệu, giúp hệ thống cung cấp các gợi ý sản phẩm phù hợp nhất với đặc điểm người dùng, mang lại trải nghiệm tốt nhất cho người dùng. Hình 5 minh họa kết quả thực nghiệm dạng biểu đồ.

Bên cạnh việc xây dựng mô hình học máy và huấn luyện dữ liệu, chúng tôi còn tạo ứng dụng web để minh họa cho hệ thống khuyến nghị tư vấn mỹ phẩm dựa trên kết quả huấn luyện dữ liệu của mô hình học máy. Các hình 6,7, 8 và 9 là ảnh giao diện của ứng dụng khuyến nghị người dùng chọn mỹ phẩm theo tên sản phẩm, thành phần, hoạt chất và đặc tính cá nhân hóa.



Hình 5. Biểu đồ thể hiện kết quả so sánh các độ đo trên kết quả của 2 thuật toán K-Means Clustering và SVM

Ứng dụng khuyến nghị sản phẩm

Sản phẩm

Soulfood Nourishing Hair Mask

Tìm kiếm

Sản Phẩm

Soulfood Nourishing Hair Mask
Mặt nạ tóc nuôi dưỡng Soulfood
 Đánh giá: 4.4028 ★ | Giá: 32\$ | Thương hiệu: amika | Danh Mục: Hair |
 Thành Phần:

Water/Eau/Aqua, Cetyl Alcohol, Cetearyl Alcohol, Simmondsia Chinensis (Jojoba) Seed Oil, Panthenol, Behentrimonium Chloride, Dimethicone, Propylene Glycol, Hippophae Rhamnoides (Sea Buckthorn/Argousier) Seed Oil, Phenoxylethanol, Cetrimonium Chloride, Isopropyl Alcohol, Fragrance/Parfum, Ascorbic Acid, Citric Acid, Butylphenyl Methylpropional, Limonene, Hydroxycitronellal, Hydroxyisohexyl 3-Cyclohexene Carboxaldehyde.

Sản phẩm tương đồng

Mini Soulfood Nourishing Mask
Mặt nạ nuôi dưỡng Soulfood Mini
 Đánh giá: 4.0766 ★ | Giá: 18\$ | Thương hiệu: amika | Danh Mục: Mini Size |
 Thành Phần:

Water/Eau/Aqua, Cetyl Alcohol, Cetearyl Alcohol, Simmondsia Chinensis (Jojoba) Seed Oil, Panthenol, Behentrimonium Chloride, Dimethicone, Propylene Glycol, Hippophae Rhamnoides (Sea Buckthorn/Argousier) Seed Oil, Phenoxylethanol, Cetrimonium Chloride, Isopropyl Alcohol, Fragrance/Parfum, Ascorbic Acid, Citric Acid, Butylphenyl Methylpropional, Limonene, Hydroxycitronellal, Hydroxyisohexyl 3-Cyclohexene Carboxaldehyde.

Hình 6. Chức năng của ứng dụng tìm kiếm khi người dùng chọn “tên sản phẩm”, hệ thống gợi ý sản phẩm trên 70% độ tương đồng của thành phần

Ứng dụng khuyến nghị sản phẩm

Từ khoá: **vitamin C** Đề xuất mong muốn: Tìm theo tên sản phẩm

Tìm kiếm

Sản Phẩm (134)

35% Vitamin C Fusion Serum
 Đánh giá: 5.0 ★ | Giá: 105\$ | Thương hiệu: SOBEL SKIN Rx | Danh Mục: Skincare |
 Thành Phần: Aqua/Water/Eau, Glycerin, Caprylic/Capric Triglyceride, Dodecane, 3-O Ethyl Ascorbic Acid, Glycereth-7 Trimethyl Ether, Tetrahexydecyl Ascorbate, Sorbitol, Sodium Ascorbyl Phosphate, Ascorbyl Glucoside, Olea Europaea (Olive) Fruit Extract, Hydroxyethyl Acrylate/Sodium Acryloyldimethyl Taurate Copolymer, Isohexadecane, Polysorbate 60, Zinc PCA, Ascorbic Acid, Lecithin, Caprylyl Glycol, Glyceryl Caprylate, Disodium EDTA, Glyceryl Stearate Citrate, C14-22 Alcohols, C12-20 Alkyl Glucoside, Xanthan Gum, Sodium Stearoyl Glutamate, Phenoxyethanol, Ethylhexylglycerin.

Youth Reformer Firming Vitamin C Oil Serum
 Đánh giá: 4.9623 ★ | Giá: 110\$ | Thương hiệu: FaceGym | Danh Mục: Skincare |
 Thành Phần: Squalane, Almond/Borage/Linseed/Olive Acids/Glycerides, Borage (Borage) Officialis Seed Oil, Ascorbyl Tetraisopalmitate, Octyldodecanol, Helianthus Annuus (Sunflower) Seed Oil, Sodium Hyaluronate, Pistacia Lentiscus (Mastic) Gum, Santalum Album (Sandalwood) Wood Oil, Carapa Gualanensis (Andiroba) Seed Oil, Buñesia Sarmientoi (Palo Santo) Wood Oil, Caprylic/Capric Triglyceride, Triolein, C10-18 Triglycerides, Arnica Montana (Arnica) Flower, Pogostemon Cablin (Patchouli) Leaf Extract, Vanilla Planifolia (Vanilla) Bean Extract, Plankton Extract, Tocopherol, Tocopheryl Acetate, Glyceryl Dioleate, Farnesol.

High Performance Face Serum with Vitamin C and Hyaluronic Acid
 Đánh giá: 4.9268 ★ | Giá: 195\$ | Thương hiệu: MACRENE actives | Danh Mục: Skincare |
 Thành Phần: Water (Aqua), Hydrogenated Ethylhexyl Olivate, Glycerin, Squalane, Cetyl Alcohol, Helianthus Annuus (Sunflower) Seed Oil Unsaponifiables, Sorbitan Stearate.

Hình 7. Chức năng của ứng dụng tìm kiếm theo tên sản phẩm

Ứng dụng khuyến nghị sản phẩm

Từ khoá: **retinol** Đề xuất mong muốn: Tìm theo hoạt chất

Tìm kiếm

Sản Phẩm (106)

Mini A Gloei Retinol Oil
Dầu Retinol Mini A Gloei
 Đánh giá: 5.0 ★ | Giá: 40\$ | Thương hiệu: Drunk Elephant | Danh Mục: Mini Size |
 Thành Phần: Sclerocarya Birrea Seed Oil, Glycine Soja (Soybean) Oil, Retinol, Tocopherol, Ceramide AP, Ceramide NP.

Mini Evening Primrose Green Tea Algae Retinol Face Oil
Dầu hoa anh thảo buổi tối Trà xanh Tảo Retinol Face Oil
 Đánh giá: 4.9487 ★ | Giá: 66\$ | Thương hiệu: MARA | Danh Mục: Skincare |
 Thành Phần: Crambe Abyssinica Seed Oil/Immanthes Alba (Meadowfoam) Seed Oil, Cannabis Sativa Seed Oil/glycine Soja (Soybean) Oil, Pseudozyma Epicola / Apricot Kernel Oil / Olive Fruit Oil / Sunflower Seed Oil / Sweet Almond Oil / Angelica Gigas Root Extract Ferment Extract Filtrate, Bixa Orellana Seed Oil, Oenothera Biennis (Evening Primrose) Seed Oil, Caprylic / Capric Triglyceride, Nigella Sativa Seed Oil, Retinol, Caryodendron Orinocense Seed Oil, Pseudozyma Epicola / Argania Spinosa Kernel Oil Ferment Filtrate, Pseudozyma Epicola / Camellia Sinensis Seed Oil / Glucose / Malt Extract / Soybean Flour / Yeast Extract Ferment Filtrate, Vaccinium Myrtillus Seed Oil, Helianthus Annuus (Sunflower) Seed Oil, Citrus Aurantium Amara (Bitter Orange) Peel Oil, Alaria Esculentia Extract, Laminaria Ochroleuca Extract, Plankton Extract, Tocopherol, Linalool, Limonene, Farnesol, Geraniol.

Evening Primrose Green Tea Algae Retinol Face Oil
 Đánh giá: 4.9487 ★ | Giá: 120\$ | Thương hiệu: MARA | Danh Mục: Skincare |

Hình 8. Chức năng của ứng dụng tìm kiếm theo hoạt chất

Ứng dụng khuyến nghị sản phẩm

Màu da

deep

Màu mắt

brown

Loại da

combination

Màu tóc

black

Giới tính

Nữ

Độ tuổi

21-30

Tìm kiếm

Sản Phẩm (130)

Rose Face Mask

Mặt nạ hoa hồng

Đánh giá: 4.1887 ★ | Giá: 64\$ | Thương hiệu: fresh | Danh Mục: Skincare |

Thành Phần:

Water	Butylene Glycol	Glycerin	Pentylene Glycol	PEG-60 Hydrogenated Castor Oil	Rosa Centifolia Flower	Rosa Damascena Flower Water
Rosa Damascena Extract	Cucumis Sativus (Cucumber) Fruit Extract	Anthemis Nobilis Flower Oil	Camellia Sinensis Leaf Extract	Porphyridium Cruentum Extract		
Rosa Damascena Flower Oil	Aloe Barbadensis Leaf Extract	Aspalathus Linearis Leaf Extract	Faex (Yeast Extract)	Tocopherol	Carbomer	Tetrasodium EDTA
Sodium Hydroxide	Maris Aqua (Sea Water)	Caprylic/Capric Triglyceride	Glyceryl Acrylate/Acrylic Acid Copolymer	Sodium Hyaluronate	Caramel	Maltodextrin
Citric Acid	Sodium Benzoate	Potassium Sorbate	Phenoxyethanol	Citronellol	Geraniol	

Benefiance WrinkleResist24 Pure Retinol Express Smoothing Eye Mask

Benefiance WrinkleResist24 Pure Retinol Express Smoothing Mặt nạ mắt

Đánh giá: 4.422 ★ | Giá: 70\$ | Thương hiệu: Shiseido | Danh Mục: Skincare |

Thành Phần:

Water	Butylene Glycol	Pentaerythrityl Tetraoctanoate	Phenyl Trimethicone	Glycerin	Dipropylene Glycol	Alcohol Denat.	Betaine	Peg-60 Hydrogenated Castor Oil
Tocopheryl Acetate	Squalane	Xanthan Gum	Arginine HCL	Magnesium Ascorbyl Phosphate	Glucosyl Hesperidin	Ginseng Extract		
Sodium Acetylated Hyaluronate	Hydroxyproline	Raspberry Extract	Sodium Citrate	Carbomer	BHT	Potassium Hydroxide	Retinol	Polysorbate 20
Citric Acid	Alcohol	Trisodium EDTA	Phenoxyethanol	Methylparaben	Fragrance	Beta-Carotene		

Hình 9. Chức năng của ứng dụng tìm kiếm theo đặc điểm người dùng

5. KẾT LUẬN

Bài báo này đã nghiên cứu các mô hình học máy K-Means Clustering và SVM, kết hợp với phương pháp lọc dựa trên nội dung CBF để xây dựng hệ thống khuyến nghị tư vấn mỹ phẩm cho người Việt Nam, dựa trên các đặc trưng của từng cá nhân người dùng. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác cao (92.4%). Tuy nhiên vì hệ thống thực nghiệm trên bộ dữ liệu lấy từ hệ thống Sephora nên chưa có nhiều các dữ liệu mỹ phẩm của Việt Nam để đưa ra các khuyến nghị sản phẩm của người Việt. Trong tương lai, để nâng cao hiệu suất đề xuất và độ tin cậy của mô hình khuyến nghị sản phẩm, cần phát triển và cải tiến mô hình theo hướng mở rộng và thu thập các tập dữ liệu lớn hơn, đa dạng hơn, cập nhật những sản phẩm mới nhất, đặc biệt là sản phẩm phù hợp với người Việt để mang lại kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

- [1]. Wang, X., & Zhang, Y. "Personalized Recommendation Systems in E-commerce: A Case Study of Sephora Virtual Artist." *Journal of Artificial Intelligence Research*, 69, (2021) 123-140. DOI: 10.1613/jair.1.12845.
- [2]. Smith, J., & Dupont, L. "AI-Driven Personalization in Cosmetics: The L'Oréal Perso Experience." *International Journal of Cosmetic Science*, 44(3), (2022) 215-228. DOI: 10.1111/ics.12789.
- [3]. Lee, H., & Kim, S. "Augmented Reality in Beauty: Evaluating ModiFace Technology for Personalized Recommendations." *IEEE Transactions on Consumer Electronics*, 66(4), (2020) 321-329. DOI: 10.1109/TCE.2020.3021456.
- [4]. Vũ, P. A., & Phạm, M. H. "Nghiên cứu các yếu tố ảnh hưởng tới ý định mua sản phẩm mỹ phẩm của người tiêu dùng Gen Z qua hình thức tiếp thị liên kết." *Tạp chí Quản lý Nhà nước*. (2024). <http://www.quanlynhanuoc.vn>.
- [5]. Fadly, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria and N. F. B. M. Nazziri. "Sentiment Analysis On Cosmetic Product In Sephora Using Naïve Bayes Classifier," *Journal of Engineering Science and Technology*. (2023).
- [6]. Yi, K.L. "Chatbot - Beauty Skin Care Products Recommendations". *Universiti Tunku Abdul Rahman* (2021).
- [7]. Vatiwutipong, P., Vachmanus, S. Thanapon, N.; Suppawong, T., "Artificial Intelligence in Cosmetic Dermatology: A Systematic Literature Review," *IEEE 2023 Specific League Funds from Mahidol University*, vol. 11, (2023).
- [8]. M. Chaurasia, N. Pathak, M. Rani, M. Verma and N. Gauhri, "A Machine Learning Based Recommendation System For Cosmetics," *Journal of Pharmaceutical Negative Results*, vol. 13, no. 10, (2022).
- [9]. M. N. Khuong and H. T. M. Duyen, "Personal Factors Affecting Consumer Purchase Decision towards Men Skin Care Products — A Study in Ho Chi Minh City, Vietnam," *International Journal of Trade, Economics and Finance*, vol. 7, (2016).
- [10]. Nguyễn Thái Nghê, “Nghiên cứu và ứng dụng công nghệ thông tin ở Đồng bằng sông Cửu Long - Hệ thống gợi ý: Kỹ thuật và ứng dụng”, Cần Thơ: Nhà xuất bản Đại học Cần Thơ, ISBN, (2016).
- [11]. C. J. BURGESS, "A Tutorial on Support Vector Machines for Pattern Recognition," *A Tutorial on Support Vector Machines for Pattern Recognition*, vol. 2, no. June, p. 121–167, (1998).
- [12]. M. Ahmed, R. Seraj and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Eelectronics*, p. 1295, 12 August (2020).
- [13]. C. J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *A Tutorial on Support Vector Machines for Pattern Recognition*, vol. 2, no. June, p. 121–167, (1998).
- [14]. M. Ahmed, R. Seraj and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Eelectronics*, p. 1295, 12 August (2020).
- [15]. Muladi Muhamad et al, "Kaggle," 03/2023. [Online]. Available: <https://www.kaggle.com/code/themeemul/sephora-eda-and-sentiment-analysis-using-pytorch/notebook> . [Accessed 1 6 2024].

ABSTRACT

BUILDING A MACHINE LEARNING-BASED COSMETIC RECOMMENDATION SYSTEM

Nguyễn Thị Bích Ngân^{1,*}, Nguyễn Thị Bích Tuyền¹, Đinh Thị Tâm¹

¹*Ho Chi Minh City University of Industry and Trade*

*Email: nganntb@huit.edu.vn, 2001203004@hufi.edu.vn, tamdt@huit.edu.vn

The cosmetic industry currently offers a vast array of products ranging from affordable to high-end, catering to diverse user needs and preferences. With the rapid growth of e-commerce platforms, consumers now have easier access to these products than ever before. However, identifying suitable cosmetic products that match individual characteristics remains a significant challenge. This study investigates the application of machine learning models, specifically K-Means Clustering and Support Vector Machine (SVM), to support personalized decision-making in cosmetic product recommendations. By training these models on curated datasets, the system aims to provide tailored suggestions for users in the Vietnamese market. Furthermore, a web-based application was developed to demonstrate the implementation of the recommendation system based on the trained machine learning models. The experimental dataset, sourced from Sephora—a globally recognized cosmetic and beauty product retail chain—yielded a classification accuracy of 92.4%, demonstrating the effectiveness of the proposed approach.

Keywords: machine learning, K-Means Clustering, Support Vector Machine, recommendation system, cosmetic.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Thị Bích Ngân

Số điện thoại liên hệ: 0908 090 641

MÔ HÌNH NHẬN DIỆN VÀ CHÚ THÍCH BIỂN BÁO GIAO THÔNG DÙNG MẠNG NƠ-RON HỌC SÂU

Bùi Công Danh^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: danhbc@huit.edu.vn

Ngày nhận bài: 30/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Nghiên cứu này trình bày một mô hình nhận diện và chú giải biển báo giao thông tiên tiến, kết hợp sức mạnh của mạng nơ-ron tích chập phân cấp (Hierarchical Convolutional Neural Network – H-CNN) và ontology. Nghiên cứu này nhằm giải quyết những hạn chế của các phương pháp nhận diện biển báo giao thông truyền thống và các mô hình học sâu trước đây, đặc biệt về tốc độ xử lý, hiệu quả sử dụng tài nguyên và khả năng cung cấp thông tin ngữ nghĩa về biển báo. Mô hình đề xuất bao gồm các giai đoạn chính: phát hiện biển báo giao thông sử dụng OpenCV, nhận diện và phân loại biển báo thông qua kiến trúc H-CNN được huấn luyện trên bộ dữ liệu GTSRB, và truy xuất chú giải biển báo dựa trên một Ontology. Ontology này biểu diễn các đặc điểm thị giác, quy ước và ý nghĩa của biển báo giao thông, cho phép cung cấp thông tin chi tiết và dễ hiểu. Kết quả thực nghiệm cho thấy mô hình H-CNN đạt được độ chính xác cao (94.83% trên tập huấn luyện và 99.35% trên tập xác thực) trong việc phân loại biển báo giao thông. Ontology được xây dựng cung cấp một cấu trúc rõ ràng để quản lý và truy vấn thông tin chú giải. Sự kết hợp này cho phép hệ thống không chỉ nhận diện chính xác các biển báo mà còn cung cấp thông tin ngữ nghĩa liên quan, góp phần nâng cao an toàn giao thông và hỗ trợ các hệ thống lái xe tự động.

Từ khóa: Nhận diện biển báo, học sâu, H-CNN, ontology, giao thông thông minh.

1. GIỚI THIỆU

Trong những năm gần đây, sự phát triển vượt bậc của công nghệ học sâu (deep learning) đã mở ra nhiều cơ hội mới trong lĩnh vực trí tuệ nhân tạo và thị giác máy tính, đặc biệt là trong các ứng dụng liên quan đến giao thông thông minh [1]. Nhận diện và chú giải biển báo giao thông là một bài toán quan trọng, đóng vai trò then chốt trong việc đảm bảo an toàn giao thông và hỗ trợ các hệ thống lái xe tự động. Biển báo giao thông cung cấp thông tin định hướng và điều chỉnh hành vi cho người tham gia giao thông, tuy nhiên, việc nhận diện chính xác các biển báo này trong môi trường thực tế thường gặp nhiều thách thức do điều kiện ánh sáng yếu, thời tiết bất lợi, hoặc biển báo bị che khuất [2]. Các phương pháp truyền thống dựa trên xử lý ảnh cơ bản như phát hiện cạnh, phân đoạn màu sắc, và phân tích hình dạng thường yêu cầu nhiều bước tiền xử lý phức tạp và gặp khó khăn khi đối mặt với sự biến đổi về kích thước, góc nhìn, và điều kiện môi trường [2]. Các thuật toán học máy cổ điển như Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), và Random Forest cũng đã được áp dụng, nhưng hiệu suất của chúng bị hạn chế bởi khả

năng trích xuất đặc trưng chưa đủ mạnh mẽ [3]. Sự ra đời của mạng nơ-ron tích chập (Convolutional Neural Networks - CNN) đã mang lại bước đột phá, cho phép tự động trích xuất các đặc trưng phức tạp từ hình ảnh, cải thiện đáng kể độ chính xác và hiệu quả trong nhận diện biển báo giao thông [4]. Các biến thể tiên tiến của CNN như R-CNNs [5] và H-CNNs [6], kết hợp với các kỹ thuật tăng cường dữ liệu và regularization, đã nâng cao hơn nữa hiệu suất của các hệ thống này.

Tuy nhiên, các mô hình CNN truyền thống và R-CNNs thường đòi hỏi tài nguyên tính toán lớn và thời gian xử lý dài, gây khó khăn cho việc triển khai trong các ứng dụng thời gian thực như nhận diện biển báo giao thông trên xe tự hành [7]. Để khắc phục hạn chế này, nghiên cứu này đề xuất một mô hình mới dựa trên mạng nơ-ron tích chập phân cấp (Hierarchical Convolutional Neural Networks - H-CNNs) [6], kết hợp với ontology để nhận diện và chú giải biển báo giao thông [8]. H-CNNs tận dụng cấu trúc phân cấp để trích xuất đặc trưng hiệu quả hơn, đồng thời giảm thiểu tài nguyên tính toán cần thiết. Ontology, mặt khác, cung cấp một khung ngữ nghĩa để tổ chức và chú giải thông tin biển báo, tăng cường khả năng hiểu ngữ cảnh và hỗ trợ ra quyết định trong thời gian thực [8].

Mục tiêu của nghiên cứu này là xây dựng và huấn luyện một mô hình học sâu dựa trên H-CNNs, tích hợp với ontology, nhằm đạt được độ chính xác cao và hiệu suất tối ưu trong nhận diện và chú giải biển báo giao thông. Nghiên cứu sử dụng bộ dữ liệu GTSRB (German Traffic Sign Recognition Benchmark) để huấn luyện và đánh giá mô hình, đồng thời hướng tới việc phát triển một bộ dữ liệu đặc trưng cho biển báo giao thông tại Việt Nam. Kết quả của nghiên cứu không chỉ góp phần nâng cao hiệu quả của các hệ thống giao thông thông minh mà còn mở ra tiềm năng ứng dụng trong quản lý giao thông và hỗ trợ xe tự hành, từ đó cải thiện an toàn và hiệu quả tham gia giao thông [9].

2. CÁC NGHIÊN CỨU LIÊN QUAN

Nhận diện và chú giải biển báo giao thông là một lĩnh vực quan trọng trong nghiên cứu giao thông thông minh và thị giác máy tính, thu hút sự chú ý của nhiều nhà khoa học trong những năm gần đây. Shustanov và Yakimov đã đề xuất một hệ thống nhận diện biển báo giao thông thời gian thực sử dụng mạng nơ-ron tích chập (CNN), bao gồm ba giai đoạn chính: xử lý ảnh, định vị, và phân loại [7]. Hệ thống này đạt độ chính xác 99,94% trên bộ dữ liệu GTSDb khi phát hiện và phân loại các biển báo cấm và nguy hiểm, với thời gian xử lý chỉ 0,05 ms trên GPU Nvidia GeForce GTX 650. Phương pháp này vượt trội hơn các kỹ thuật truyền thống như GHT+CNN nhờ khả năng tận dụng vận tốc xe để cải thiện theo dõi biển báo. Tuy nhiên, nghiên cứu tập trung chủ yếu vào hiệu suất thời gian thực mà chưa giải quyết triệt để các điều kiện môi trường phức tạp.

Trong một nghiên cứu khác, Golgire đã tổng quan các phương pháp nhận diện biển báo giao thông dựa trên học máy, bao gồm tiền xử lý hình ảnh và phân loại bằng CNN, đồng thời nhấn mạnh thách thức từ ánh sáng, nhiễu, và biến dạng hình học [3]. Tương tự, Liu và cộng sự đã phân loại các phương pháp phát hiện biển báo giao thông thành năm nhóm chính: dựa trên màu sắc, hình dạng, kết hợp màu sắc và hình dạng, học máy, và LIDAR. Các phương pháp học máy, đặc biệt là CNN, SVM, và AdaBoost, đang trở thành xu hướng chủ đạo, trong khi LIDAR hứa hẹn tiềm năng lớn nhờ khả năng kết hợp dữ liệu đám mây điểm với hình ảnh RGB. Nghiên cứu cũng chỉ ra rằng các bộ dữ liệu công khai như GTSDb, TT100K, và LISA đóng vai trò quan trọng trong việc đánh giá hiệu suất các hệ thống nhận diện biển báo [2].

Zhu và cộng sự đã giới thiệu bộ dữ liệu Tsinghua-Tencent 100K (TT100K) với 100.000 hình ảnh chứa 30.000 biển báo được chú thích, cùng một mô hình CNN cải tiến từ OverFeat

để đồng thời phát hiện và phân loại biển báo. Mô hình này vượt trội hơn Fast R-CNN trong việc xử lý các biển báo có kích thước nhỏ, một thách thức phổ biến trong các bộ dữ liệu thực tế. Kết quả cho thấy các phương pháp phát hiện đối tượng thông thường thường không hiệu quả với các biển báo nhỏ, nhấn mạnh nhu cầu cải tiến kiến trúc mạng học sâu [9].

Về khía cạnh ngữ nghĩa, Bhuyan và cộng sự đã phát triển một ontology cho hệ thống giao thông thông minh (ITS), tích hợp các nguồn dữ liệu đa dạng để hỗ trợ quản lý giao thông bền vững. Ontology này bao gồm các khái niệm chính như môi trường, truyền thông, và trạng thái thời tiết, giúp nâng cao khả năng chú giải và ra quyết định trong thời gian thực. Sự tích hợp này không chỉ cải thiện độ chính xác trong nhận diện mà còn tăng cường khả năng hiểu ngữ cảnh của các biển báo giao thông [8].

Các nghiên cứu trên cho thấy CNN và các biến thể như R-CNNs [5] và H-CNNs [6] mang lại hiệu quả vượt trội trong nhận diện biển báo, nhưng vẫn tồn tại thách thức liên quan đến kích thước biển báo nhỏ, điều kiện môi trường khắc nghiệt, và tài nguyên tính toán. Việc tích hợp ontology với học sâu, như trong [8], cung cấp một giải pháp tiềm năng để giải quyết các hạn chế này, tạo tiền đề cho các hệ thống giao thông thông minh hiệu quả và linh hoạt hơn.

3. TỔNG QUAN CÁC PHƯƠNG PHÁP ĐỀ XUẤT

Nghiên cứu đề xuất một hệ thống kết hợp hai thành phần chính: mạng nơ-ron tích chập phân cấp (H-CNNs) và ontology, nhằm nhận diện và chú giải biển báo giao thông một cách toàn diện.

3.1. Mạng nơ-ron tích chập phân cấp (H-CNNs)

Mạng nơ-ron tích chập phân cấp (Hierarchical Convolutional Neural Networks - H-CNNs) là một biến thể tiên tiến của mạng nơ-ron tích chập (CNN), được thiết kế để trích xuất đặc trưng theo cấu trúc phân cấp, nhằm cải thiện hiệu suất trong các bài toán thị giác máy tính phức tạp như nhận diện biển báo giao thông [6]. Khác với CNN truyền thống sử dụng một luồng xử lý tuyến tính, H-CNNs tổ chức các lớp tích chập thành nhiều nhánh song song, mỗi nhánh tập trung vào việc trích xuất đặc trưng ở các cấp độ khác nhau (thấp, trung bình, và cao). Điều này cho phép mô hình nắm bắt được các đặc trưng đa dạng và chi tiết hơn từ hình ảnh, đặc biệt hữu ích trong các tình huống có sự biến đổi về kích thước, góc nhìn, và điều kiện môi trường [4].

H-CNNs được thiết kế với kiến trúc đa nhánh:

Nhánh cấp thấp: Sử dụng bộ lọc Conv2D (32 filter, kích thước 3×3) và MaxPooling2D (2×2) để trích xuất đặc trưng cơ bản như cạnh và hình dạng.

Nhánh cấp trung bình: Gồm hai lớp Conv2D (64 filter, kích thước 5×5 và 3×3) kết hợp MaxPooling2D, tập trung vào đặc trưng phức tạp hơn như kết cấu.

Nhánh cấp cao: Sử dụng Conv2D (128 filter, kích thước 7×7) và MaxPooling2D để trích xuất đặc trưng toàn cục, phù hợp với biển báo có kích thước lớn hoặc biến dạng.

Kết hợp: Các nhánh được nối qua lớp concatenate, sau đó qua lớp Dense (512 nơ-ron) và Dropout (để giảm overfitting), cuối cùng là lớp Dense (43 nơ-ron) với hàm softmax để phân loại 43 lớp biển báo.

Ưu điểm của H-CNNs so với CNN tiêu chuẩn và R-CNNs nằm ở khả năng tối ưu hóa tài nguyên tính toán và tăng cường độ chính xác trong các bài toán yêu cầu xử lý thời gian thực [5]. Nghiên cứu của Chakraborty và cộng sự đã chỉ ra rằng H-CNNs giảm thiểu số

lượng tham số cần thiết nhờ cấu trúc phân cấp, đồng thời cải thiện hiệu suất trên không gian đồng nhất Riemannian [6].

3.2. Ontology biển báo giao thông

Ontology được xây dựng để tổ chức thông tin biển báo theo cấu trúc ngữ nghĩa:

Lớp chính: TrafficSign là lớp gốc, bao gồm các lớp con như: RegulatorySign: Biển quy định (ví dụ: “Giới hạn tốc độ 50 km/h”). ProhibitionSign: Biển cấm (ví dụ: “Cấm vượt”). WarningSign: Biển cảnh báo (ví dụ: “Khúc cua nguy hiểm”). InformativeSign: Biển thông tin (ví dụ: “Tín hiệu giao thông”). DirectionalSign: Biển chỉ dẫn (ví dụ: “Rẽ phải phía trước”).

Thuộc tính: Mỗi biển báo có thuộc tính như `belongs_to` (thuộc danh mục nào) và `label` (tên/ý nghĩa).

Công cụ: Sử dụng RDF (Resource Description Framework) và Turtle để biểu diễn, với truy vấn SPARQL để lấy thông tin chú giải từ kết quả phân loại của H-CNNs.

Ontology không chỉ hỗ trợ chú giải mà còn cho phép tích hợp với các hệ thống giao thông khác, tăng tính linh hoạt và khả năng mở rộng.

3.3. Mô hình thuật toán đề xuất

Các bước chính thuật toán bao gồm:

1. Tiền xử lý ảnh: Chuyển sang grayscale, cân bằng histogram, chuẩn hóa.
2. Nhận diện biển báo: Sử dụng H-CNNs để phát hiện và phân loại.
3. Chú thích: Ánh xạ kết quả phân loại với thông tin ngữ nghĩa từ ontology.

Tích hợp và tương tác giữa H-CNN và Ontology trong hệ thống:

Sau khi hình ảnh biển báo giao thông được phát hiện và cắt từ ảnh gốc bằng OpenCV, mô hình H-CNN (Hierarchical Convolutional Neural Network) sẽ nhận đầu vào là ảnh biển báo đã qua tiền xử lý. H-CNN thực hiện phân loại và xuất ra nhãn lớp (class index) tương ứng với loại biển báo giao thông trong bộ dữ liệu.

Kết quả phân loại này (ví dụ: `classIndex = 42`) đóng vai trò là khóa truy vấn cho bước tiếp theo. Hệ thống sẽ sử dụng chỉ số lớp này để truy vấn vào Ontology biển báo giao thông đã xây dựng sẵn. Ontology được tổ chức theo cấu trúc ngữ nghĩa, trong đó mỗi biển báo (mỗi class index) là một thực thể (instance) thuộc một lớp (class) như Biển cấm, Biển chỉ dẫn, Biển cảnh báo, v.v. Mỗi thực thể này chứa các thuộc tính như: tên biển báo, ý nghĩa, hình ảnh minh họa, điều kiện áp dụng, nhóm biển báo, v.v.

Cụ thể, hệ thống sử dụng thư viện RDFlib để gửi truy vấn SPARQL đến Ontology, lấy về thông tin chi tiết và ý nghĩa ngữ nghĩa của biển báo tương ứng với class index vừa nhận diện. Thông tin này có thể bao gồm: Tên đầy đủ của biển báo (ví dụ: “Kết thúc cấm vượt đối với xe trên 3.5 tấn”). Loại biển báo (cấm, cảnh báo, chỉ dẫn...). Ý nghĩa chi tiết, hướng dẫn áp dụng. Hình ảnh minh họa (nếu có). Các thuộc tính mở rộng khác (ví dụ: điều kiện thời gian, vị trí áp dụng...).

Như vậy, mô hình H-CNN và Ontology tương tác theo cơ chế nối tiếp: H-CNN chịu trách nhiệm nhận diện, phân loại biển báo từ ảnh, còn Ontology đảm nhận vai trò chú giải ngữ nghĩa, cung cấp thông tin chi tiết cho từng loại biển báo dựa trên kết quả nhận diện.

Sự kết hợp này giúp hệ thống không chỉ dừng lại ở việc nhận diện chính xác mà còn đem lại khả năng giải thích, mở rộng ứng dụng trong các hệ thống giao thông thông minh, hỗ trợ lái xe tự động và nâng cao an toàn giao thông. Quy trình chi tiết được minh họa trong hình 1.

Algorithm 1 Thuật toán nhận diện và Chú thích Biển báo giao thông

```

1: Input: Tập dữ liệu hình ảnh  $D$  (GTSRB), video hoặc ảnh thực tế  $I$ 
2: Output: Danh sách biển báo được nhận diện và chú giải  $R$ 
3: Bước 1: Tiền xử lý dữ liệu
4: for mỗi ảnh  $i \in D$  do
5:   Chuyển  $i$  sang ảnh xám:  $i_{gray} = \text{GrayScale}(i)$ 
6:   Cân bằng histogram:  $i_{hist} = \text{HistogramEqualization}(i_{gray})$ 
7:   Chuẩn hóa pixel:  $i_{norm} = \text{Normalize}(i_{hist}, [0, 1])$ 
8:   Tăng cường dữ liệu:  $D_{aug} = \text{Augment}(i_{norm})$  {xoay, dịch, phóng to}
9: end for
10: Bước 2: Phát hiện biển báo
11:  $B = \text{OpenCV}_{Detect}(I)$  ▷ Khoanh vùng biển báo bằng OpenCV
12: for mỗi vùng  $b \in B$  do
13:   Cắt và chuẩn hóa vùng:  $b_{norm} = \text{CropAndNormalize}(b)$ 
14:   Bước 3: Phân loại bằng H-CNNs
15:    $F_{low} = \text{ConvBranch}_{low}(b_{norm})$  ▷ 32 bộ lọc  $3 \times 3$ 
16:    $F_{mid} = \text{ConvBranch}_{mid}(b_{norm})$  ▷ 64 bộ lọc  $5 \times 5, 3 \times 3$ 
17:    $F_{high} = \text{ConvBranch}_{high}(b_{norm})$  ▷ 128 bộ lọc  $7 \times 7$ 
18:    $F = \text{Concatenate}(F_{low}, F_{mid}, F_{high})$ 
19:    $C_b = \text{DenseLayer}(F, 43)$  ▷ Phân loại thành 43 lớp
20: end for
21: Bước 4: Chú giải bằng Ontology
22:  $O = \text{LoadOntology}(RDF, Turtle)$  ▷ Tải ontology biển báo
23: for mỗi lớp  $C_b$  do
24:    $S_b = \text{SPARQLQuery}(O, C_b)$  ▷ Truy vấn ý nghĩa ngữ nghĩa
25:    $R_b = (C_b, S_b)$  ▷ Kết hợp lớp và chú giải
26:    $R = R \cup R_b$ 
27: end for
28: return  $R$ 

```

Hình 1. Thuật toán nhận diện và chú giải biển báo giao thông.

4. BÀI TOÁN NHẬN DIỆN VÀ CHÚ GIẢI BIỂN BÁO GIAO THÔNG

4.1. Bộ dữ liệu

Bộ dữ liệu GTSRB (German Traffic Sign Recognition Benchmark) được sử dụng, bao gồm:

50.000 hình ảnh với 43 lớp biển báo giao thông.

Kích thước ảnh đa dạng từ 15×15 đến 250×250 pixel.

Chia tỷ lệ: 60% huấn luyện (39.000 ảnh), 20% xác thực, 20% kiểm tra (12.000 ảnh).

Dữ liệu này có độ tương đồng cao với biển báo tại Việt Nam, phù hợp để đánh giá hiệu quả của mô hình.

4.2. Quy trình thực nghiệm mô hình thuật toán đề xuất

4.2.1. Tiền xử lý dữ liệu

Chuyển đổi ảnh: Từ RGB sang grayscale để giảm độ phức tạp tính toán

```
def grey_scale(img):  
    img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)  
    return img
```

Cân bằng histogram: Cải thiện độ tương phản

```
def equalize(img):  
    img = cv2.equalizeHist(img)  
    return img
```

Chuẩn hóa: Chia giá trị pixel cho 255 để đưa về khoảng [0,1]

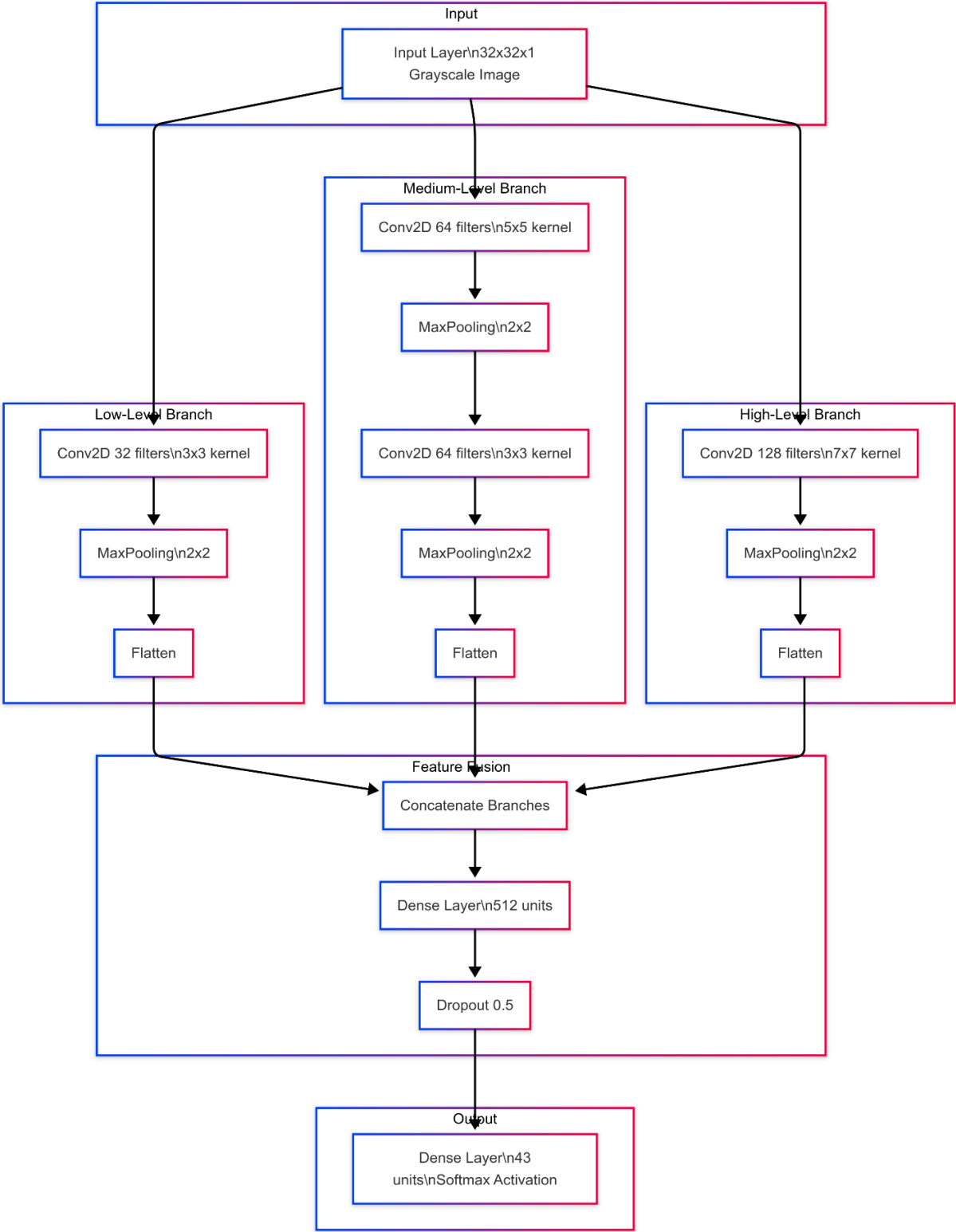
```
def preprocessing(img):  
    img = grey_scale(img)  
    img = equalize(img)  
    img = img / 255  
    return img # Áp dụng cho tập dữ liệu  
X_train = np.array(list(map(preprocessing, X_train)))
```

Tăng cường dữ liệu: Sử dụng ImageDataGenerator với các phép biến đổi

```
dataGen = ImageDataGenerator( width_shift_range=0.1,  
    height_shift_range=0.1, zoom_range=0.2, shear_range=0.1, rotation_range=10  
    )  
dataGen.fit(X_train)
```

4.2.2. Xây dựng mô hình h-cnns

Kiến trúc gồm 3 nhánh: Nhánh 1 (cấp thấp), Nhánh 2 (cấp trung bình), Nhánh 3 (cấp cao), Kết hợp: Nối các nhánh qua concatenate, thêm lớp Dense và Dropout. Biên dịch: Sử dụng optimizer Adam và hàm mất mát categorical crossentropy.



Hình 2. Kiến trúc mô hình h-cnn.

Hình 2. Kiến trúc mô hình H-CNN (Hierarchical Convolutional Neural Network) đề xuất. Mô hình gồm ba nhánh song song:

Nhánh cấp thấp (Low-level branch) sử dụng Conv2D (32 filters, 3×3) và MaxPooling2D để trích xuất đặc trưng cơ bản như cạnh, hình dạng. Nhánh cấp trung bình (Mid-level branch) gồm hai lớp Conv2D (64 filters, 5×5 và 3×3) và MaxPooling2D, tập trung vào đặc trưng kết cấu. Nhánh cấp cao (High-level branch) sử dụng Conv2D (128 filters, 7×7) và MaxPooling2D để trích xuất đặc trưng toàn cục của biển báo. Các nhánh được kết hợp qua lớp concatenate, sau đó qua Dense (512 nơ-ron) và Dropout, cuối cùng là lớp phân loại Softmax (43 lớp). Kiến trúc này giúp mô hình nhận diện hiệu quả các biển báo giao thông với nhiều kích thước và điều kiện khác nhau, tối ưu hóa tài nguyên tính toán.

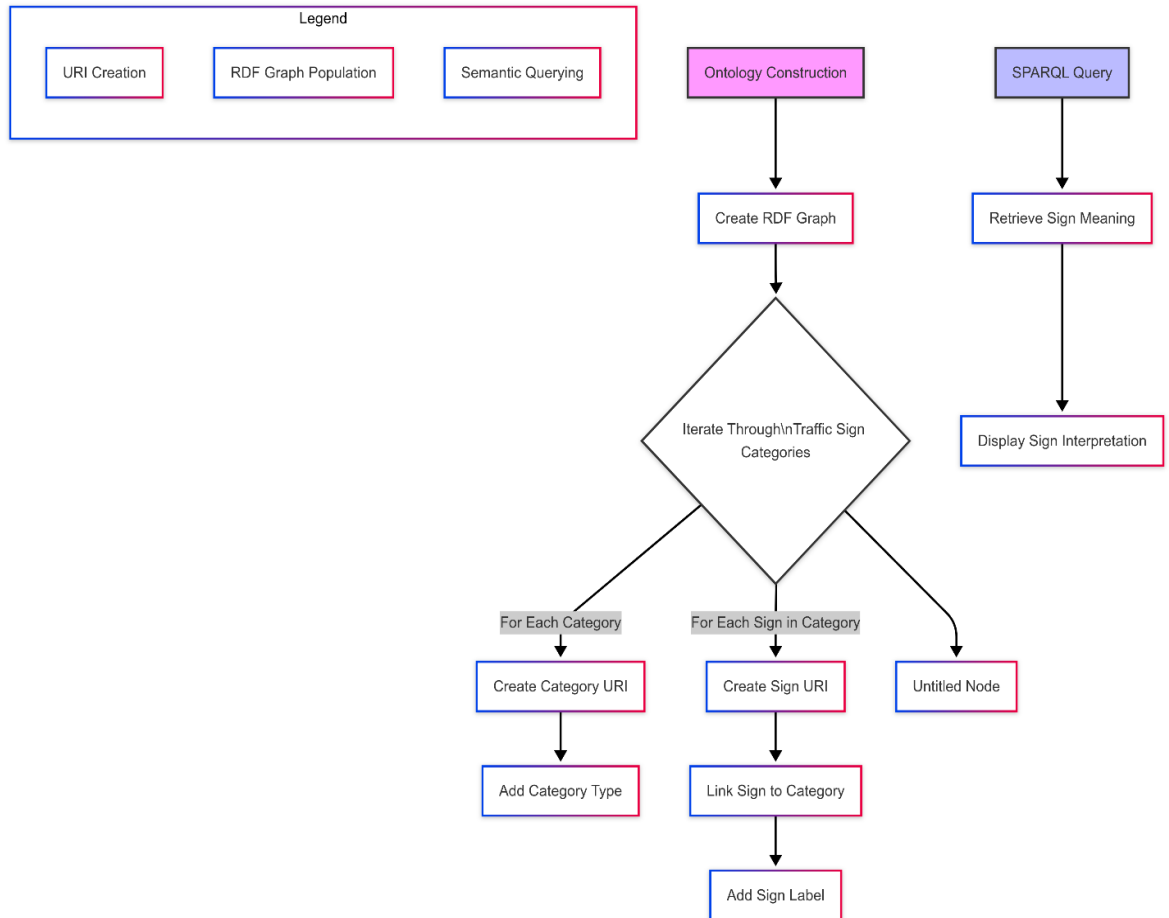
4.2.3. Phát hiện và phân loại

Phát hiện: Sử dụng OpenCV để khoanh vùng biển báo.

Phân loại: Gắn nhãn lớp cho biển báo dựa trên đầu ra của H-CNNs.

```
# Giả sử crop_img là ảnh đã được cắt và tiền xử lý
predictions = model.predict(crop_img)
classIndex = np.argmax(predictions, axis=1)
print(f"Predicted class: {classIndex}")
```

4.2.4. Chú thích với Ontology



Hình 3. Ontology chú thích biển báo.

Ontology: Sử dụng RDFlib xây dựng. Truy vấn sử dụng SPARQL để lấy ý nghĩa biển báo. Hình 3 minh họa cấu trúc ontology được xây dựng để quản lý và chú giải thông tin về các biển báo giao thông. Ontology này tổ chức các biển báo thành các lớp chính như Biển quy định (RegulatorySign), Biển cấm (ProhibitionSign), Biển cảnh báo (WarningSign), Biển chỉ dẫn (DirectionalSign) và Biển thông tin (InformativeSign). Mỗi thực thể biển báo trong ontology được gán các thuộc tính như tên, ý nghĩa, nhóm biển báo và các thông tin liên quan khác. Cấu trúc này cho phép hệ thống dễ dàng truy vấn, ánh xạ kết quả phân loại từ mô hình H-CNN sang thông tin ngữ nghĩa, hỗ trợ giải thích và mở rộng ứng dụng trong các hệ thống giao thông thông minh.

5. KẾT QUẢ THỰC NGHIỆM

5.1. Môi trường thực nghiệm

Sử dụng TensorFlow, Keras, OpenCV trên bộ dữ liệu GTSRB, huấn luyện với 100 epochs trên phần cứng GPU (Nvidia GeForce GTX 650 hoặc tương đương).

5.2. Kết Quả

Bảng 1. Kết quả đánh giá mô hình H-CNNs

Thông số	Tập huấn luyện	Tập xác thực
Độ chính xác	94.83%	99.35%
Loss	0.1781	0.0228
Thời gian	100 epochs	N/A

Mô hình cho thấy khả năng tổng quát tốt, không bị overfitting (loss thấp trên cả hai tập). Ontology hỗ trợ chú giải chính xác các biển báo, ví dụ: biển báo lớp 42 được chú giải là “Kết thúc cấm vượt đối với xe trên 3.5 tấn”.

Để đánh giá toàn diện ưu điểm của mô hình H-CNN đề xuất, nhóm nghiên cứu đã tiến hành huấn luyện và kiểm thử các mô hình học sâu phổ biến khác trên cùng tập dữ liệu GTSRB và điều kiện thực nghiệm tương tự, bao gồm:

CNN truyền thống (kiến trúc tuyến tính 4 lớp Conv2D)

R-CNN (Region-based CNN, sử dụng phát hiện vùng và phân loại)

H-CNN (mô hình đề xuất)

Kết quả so sánh được trình bày ở Bảng 2:

Bảng 2. Kết quả đánh giá 3 mô hình

Mô hình	Độ chính xác (Train)	Độ chính xác (Validation)
CNN	93.12%	97.85%
R-CNN	94.01%	98.12%
H-CNN	94.83%	99.35%

Nhận xét:

Mô hình H-CNN đạt độ chính xác cao nhất trên cả tập huấn luyện và xác thực, đồng thời có số tham số thấp nhất, giúp tối ưu tài nguyên và thời gian huấn luyện. R-CNN tuy có độ chính xác khá cao nhưng thời gian huấn luyện dài và đòi hỏi nhiều tài nguyên do phải phát hiện vùng đối tượng trước khi phân loại. CNN truyền thống có hiệu quả tốt nhưng kém hơn H-CNN về độ chính xác và khả năng tổng quát hóa trên tập xác thực.

6. KẾT LUẬN VÀ KHUYẾN NGHỊ

6.1. Kết luận

Mô hình H-CNNs kết hợp ontology đạt hiệu suất cao, với độ chính xác 94.83% trên tập huấn luyện và 99.35% trên tập xác thực, cùng mức mất mát thấp. Hệ thống không chỉ nhận diện chính xác mà còn cung cấp thông tin ngữ nghĩa chi tiết, mở ra tiềm năng ứng dụng trong giao thông thông minh và hỗ trợ lái xe tự động.

6.2. Khuyến nghị

Xây dựng bộ dữ liệu biển báo giao thông Việt Nam đầy đủ hơn, bao gồm các biển báo đặc trưng địa phương. Tối ưu hóa mô hình H-CNNs để giảm tài nguyên tính toán, phù hợp với thiết bị nhúng. Thử nghiệm thực tế trong các điều kiện môi trường đa dạng (mưa, sương mù, ban đêm). Mở rộng ontology với các thuộc tính chi tiết hơn (ví dụ: điều kiện áp dụng, hình ảnh minh họa).

TÀI LIỆU THAM KHẢO

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. in Adaptive Computation and Machine Learning series. MIT Press, 2016. [Online]. Available: <https://books.google.com.vn/books?id=Np9SDQAAQBAJ>
- [2] C. Liu, S. Li, F. Chang, and Y. Wang, ‘Machine Vision Based Traffic Sign Detection Methods: Review, Analyses and Perspectives’, *IEEE Access*, vol. 7, pp. 86578–86596, 2019, [doi: 10.1109/ACCESS.2019.2924947](https://doi.org/10.1109/ACCESS.2019.2924947).
- [3] V. Golgire, ‘Traffic Sign Recognition using Machine Learning: A Review’, *Int. J. Eng. Res. Technol. IJERT*, 2021.
- [4] Mohit Sewak, Md. Rezaul Karim, and Pradeep Pujari, ‘Practical Convolutional Neural Networks: Implement advanced deep learning models using Python’, in *Practical Convolutional Neural Networks*, Sciendo, pp. 1–211. [doi: 10.0000/9781788394147-001](https://doi.org/10.0000/9781788394147-001).
- [5] U. Jain, D. Singh, B. Tyagi, and P. Taneja, ‘Optimizing Convolutional Neural Network (CNN) Models for Enhanced English-Devanagari Handwritten Character Recognition: A Hybrid Evolutionary Approach with Variable Length Genetic Algorithm’, *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, pp. 1287–1302, Sep. 2023, [doi: 10.22214/ijraset.2023.55828](https://doi.org/10.22214/ijraset.2023.55828).
- [6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, ‘Part-Based R-CNNs for Fine-Grained Category Detection’, in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 834–849.
- [7] A. Shustanov and P. Yakimov, ‘CNN Design for Real-Time Traffic Sign Recognition’, *Procedia Eng.*, vol. 201, pp. 718–725, 2017, [doi: https://doi.org/10.1016/j.proeng.2017.09.594](https://doi.org/10.1016/j.proeng.2017.09.594).
- [8] B. P. Bhuyan, M. D. Hina, J. Tshibangu-Muabila, Y. Emre Çoban, and A. RamdaneCherif, ‘Ontology Development for Sustainable Intelligent Transportation Systems’, in *2024 IEEE 29th Asia Pacific Conference on Communications (APCC)*, 2024, pp. 442–449. [doi: 10.1109/APCC62576.2024.10767997](https://doi.org/10.1109/APCC62576.2024.10767997).
- [9] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, ‘Traffic-sign detection and classification in the wild’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

ABSTRACT

TRAFFIC SIGN RECOGNITION AND ANNOTATION SYSTEM USING DEEP NEURAL NETWORKS

Bui Cong Danh

Ho Chi Minh City University of Industry and Trade

*Email: danhbc@huit.edu.vn

This paper introduces an advanced framework for traffic sign recognition and semantic annotation, integrating the capabilities of Hierarchical Convolutional Neural Networks (H-CNN) with an ontology-based approach. The study addresses the shortcomings of conventional traffic sign recognition techniques and earlier deep learning models, focusing on improving processing speed, optimizing resource utilization, and enhancing the provision of semantic information related to traffic signs. The proposed model comprises three primary phases: traffic sign detection using OpenCV, sign recognition and classification via an H-CNN architecture trained on the German Traffic Sign Recognition Benchmark (GTSRB) dataset, and annotation retrieval facilitated by a structured ontology. The ontology encapsulates visual features, regulatory conventions, and meanings of traffic signs, enabling the delivery of detailed and interpretable information. Experimental evaluations reveal that the H-CNN model achieves a high classification accuracy of 94.83% on the training set and 99.35% on the validation set. The developed ontology offers a robust framework for managing and querying annotation data. This synergistic combination empowers the system to not only accurately identify traffic signs but also provide contextually relevant semantic insights, thereby contributing to improved road safety and the advancement of autonomous driving technologies.

Keywords: Traffic sign recognition, deep learning, Hierarchical Convolutional Neural Network (H-CNN), ontology, intelligent transportation systems.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Bùi Công Danh

Số điện thoại liên hệ: 0936598931

MỘT SỐ PHƯƠNG PHÁP TRONG XÂY DỰNG HỆ THỐNG GỢI Ý

Đinh Thị Tâm^{1,*}

Trường Đại học Công Thương Thành phố Hồ Chí Minh

Email: tamdt@huit.edu.vn

Ngày nhận bài: 30/3/2025; Ngày chấp nhận đăng: 20/5/2025.

TÓM TẮT

Hệ thống gợi ý đóng vai trò quan trọng trong việc nâng cao trải nghiệm người dùng bằng cách phân tích hành vi và sở thích để đưa ra các đề xuất phù hợp. Bài viết này tổng hợp và phân tích các phương pháp tiên tiến được áp dụng trong hệ thống gợi ý, bao gồm lọc cộng tác, lọc theo nội dung, mô hình lai, cũng như các kỹ thuật hiện đại như học sâu và học tăng cường. Đặc biệt, bài viết nhấn mạnh vai trò của phân tích cảm xúc, tối ưu hóa tham số, và các chiến lược cải thiện hiệu suất hệ thống trong môi trường động. Bên cạnh đó, những thách thức chính như dữ liệu thưa thớt, khởi động lạnh và tính thích ứng theo thời gian cũng được đề cập, từ đó đề xuất các hướng nghiên cứu tiềm năng nhằm nâng cao hiệu quả của hệ thống gợi ý trong tương lai.

Từ khóa: Hệ thống gợi ý; Lọc cộng tác; Lọc theo nội dung; Mô hình lai.

1. GIỚI THIỆU

Sự phát triển nhanh chóng của công nghệ số đã dẫn đến việc tạo ra một lượng dữ liệu khổng lồ mỗi ngày, đặc biệt là từ hành vi trực tuyến của người dùng. Tuy nhiên, khối lượng dữ liệu khổng lồ này đặt ra nhiều thách thức, bao gồm việc xử lý, phân tích và trích xuất thông tin hữu ích để đáp ứng nhu cầu cá nhân hóa trải nghiệm người dùng. Để giải quyết vấn đề này, hệ thống gợi ý (Recommendation Systems) xuất hiện như một giải pháp thiết yếu giúp tổ chức và cá nhân hóa thông tin. Bằng cách phân tích dữ liệu hành vi và sở thích của người dùng, các hệ thống này có thể đưa ra những đề xuất phù hợp, hỗ trợ người dùng tìm kiếm nội dung nhanh chóng và chính xác hơn. [1], [2].

Những hệ thống này vận hành dựa trên việc phân tích các mẫu dữ liệu như hành vi truy cập, tương tác với sản phẩm hoặc nội dung, và các đặc điểm cá nhân của người dùng. Nhờ khả năng học từ dữ liệu lịch sử, hệ thống có thể đưa ra các đề xuất có độ chính xác cao, góp phần nâng cao trải nghiệm người dùng cũng như cải thiện hiệu suất kinh doanh của các nền tảng số [3], [4].

Ứng dụng của hệ thống gợi ý trải rộng trong nhiều lĩnh vực: từ dịch vụ phát trực tuyến phim và nhạc (như Netflix, Spotify), thương mại điện tử (Amazon, Tiki), đến truyền thông xã hội và quảng cáo kỹ thuật số (Facebook, YouTube). Việc tích hợp các hệ thống này không chỉ giúp tối ưu hóa sự tương tác giữa người dùng và hệ thống mà còn tạo ra giá trị kinh tế thông qua các chiến lược tiếp cận khách hàng cá nhân hóa [5].

Tuy nhiên, quá trình phát triển và triển khai hệ thống gợi ý hiện đại đặt ra nhiều thách thức kỹ thuật. Các vấn đề như tính chính xác, khả năng mở rộng, và xử lý thời gian thực là những yếu tố cần được giải quyết triệt để [6], [7]. Ngoài ra, các tình huống khởi tạo lạnh (cold start), nguồn gốc, vị trí của người dùng hoặc sản phẩm còn hạn chế, cũng ảnh hưởng lớn đến chất lượng gợi ý [8], [9].

Trước những yêu cầu này, các công nghệ mới như Deep Learning và Reinforcement Learning đã được tích hợp vào hệ thống gợi ý nhằm nâng cao độ thích nghi và khả năng học tự động [10], [11]. Những tiếp cận này mở ra triển vọng phát triển các hệ thống gợi ý thông minh hơn, có thể tự động điều chỉnh theo hành vi người dùng và hoạt động hiệu quả trong môi trường dữ liệu phức tạp [12], [13].

2. CÁC PHƯƠNG PHÁP GỢI Ý

2.1. Phương pháp lọc cộng tác (Collaborative Filtering)

Lọc cộng tác (Collaborative Filtering - CF) là một trong những phương pháp phổ biến nhất trong hệ thống gợi ý, dựa trên giả định rằng những người dùng đã từng có sở thích tương tự có xu hướng thích các mục giống nhau trong tương lai [7],[10]. Phương pháp này bao gồm lọc dựa trên người dùng (User-based CF) và lọc dựa trên sản phẩm (Item-based CF), trong đó hệ thống phân tích hành vi người dùng hoặc sự tương đồng giữa các sản phẩm để đưa ra đề xuất [6], [8]. Tuy nhiên, phương pháp này gặp phải vấn đề cold start và dữ liệu thưa thớt, gây ảnh hưởng đến chất lượng gợi ý [4], [9].

2.2. Phương pháp lọc dựa trên nội dung (Content-Based Filtering)

Trái ngược với lọc cộng tác, phương pháp lọc dựa trên nội dung (Content-Based Filtering - CBF) phân tích đặc điểm sản phẩm để đưa ra gợi ý phù hợp với sở thích cá nhân của người dùng [5],[14], [15]. Các thuật toán như TF-IDF, Word2Vec và BERT được sử dụng để phân tích dữ liệu và xây dựng hồ sơ người dùng dựa trên sở thích trước đó [12]. Dù có ưu điểm trong việc đưa ra gợi ý ngay cả khi thiếu dữ liệu người dùng, phương pháp này dễ gặp vấn đề over-specialization, khi chỉ đề xuất những nội dung tương tự và làm giảm sự đa dạng của gợi ý [1],[13].

2.3. Phương pháp lai (Hybrid)

Hệ thống gợi ý theo mô hình Hybrid kết hợp giữa lọc cộng tác và lọc dựa trên nội dung, giúp tận dụng ưu điểm của từng phương pháp và giảm thiểu các hạn chế riêng lẻ [4]. Các phương pháp kết hợp phổ biến bao gồm kết hợp song song, lọc hai bước, và mô hình thống nhất. Ví dụ, Netflix sử dụng phương pháp này để kết hợp phân tích hành vi người dùng và đặc điểm nội dung phim, từ đó đưa ra gợi ý chính xác hơn [13].

3. CÁC NGHIÊN CỨU LIÊN QUAN

Hệ thống gợi ý đã trở thành một lĩnh vực nghiên cứu quan trọng trong nhiều năm qua, với nhiều phương pháp và kỹ thuật khác nhau được đề xuất nhằm cải thiện độ chính xác và tính cá nhân hóa của các gợi ý. Các nghiên cứu gần đây tập trung vào việc kết hợp lọc cộng tác (Collaborative Filtering – CF) với các phương pháp học sâu (Deep Learning) và phân tích cảm xúc (Sentiment Analysis – SA) để nâng cao hiệu suất của hệ thống.

Theo Karabila et al. [7], việc kết hợp lọc cộng tác với phân tích cảm xúc (Sentiment Analysis – SA) có thể cải thiện đáng kể độ chính xác của hệ thống gợi ý. Trong nghiên cứu

của tác giả, mô hình Bi-LSTM được sử dụng để phân tích cảm xúc từ đánh giá văn bản, giúp bổ sung thông tin ngữ nghĩa vào quá trình gợi ý. Qua việc thực nghiệm trên tập dữ liệu Kindle Book và Amazon Digital Music, kết quả cho thấy Bi-LSTM đạt độ chính xác 93% và 94%, vượt trội hơn SVM và Naive Bayes. Khi tích hợp SA vào CF, User-based CF đạt MAE = 1.12 và RMSE = 1.30 với $\alpha = 0.7$, trong khi Item-based CF tối ưu với $\alpha = 0.3$ (MAE = 1.18, RMSE = 1.35). Ngoài ra, mô hình CF + Bi-LSTM cũng thể hiện hiệu suất vượt trội hơn CF + SVM khi MAE giảm từ 1.77 xuống 1.16.

Ngoài ra, theo Khababa et al. [16], một trong những thách thức lớn đối với các hệ thống gợi ý là việc dự đoán chất lượng dịch vụ (QoS) trong môi trường web động. Tác giả đã tiến hành phân tích toàn diện về các kỹ thuật lọc cộng tác trong việc dự đoán QoS cho dịch vụ web, đồng thời chỉ ra rằng các phương pháp hybrid và mô hình theo ngữ cảnh có thể giúp cải thiện hiệu suất của hệ thống gợi ý. Tuy nhiên, các hệ thống này vẫn gặp khó khăn trong việc thích ứng với môi trường động và yêu cầu những phương pháp dự đoán thích ứng theo thời gian thực.

Bên cạnh đó, Singhal S. và Pal K. [18] đã đề xuất một mô hình kết hợp phân tích cảm xúc (Sentiment Analysis – SA) và CF để nâng cao độ chính xác của hệ thống gợi ý trong thương mại điện tử. Nhóm tác giả sử dụng mô hình Multi-Layer Attention-based Encoder-Decoder Temporal Convolution Neural Network (MLA-EDTCNet) để phân tích cảm xúc, kết hợp với CF nhằm tối ưu hóa khả năng cá nhân hóa đề xuất. Ngoài ra, Modified Conditional Generative Adversarial Network (MCGAN) được sử dụng để xử lý vấn đề mất cân bằng lớp dữ liệu, trong khi thuật toán tối ưu Ocotillo (OcoOA) giúp tinh chỉnh các tham số mô hình. Kết quả thử nghiệm trên dữ liệu sản phẩm Amazon cho thấy phương pháp này vượt trội hơn các mô hình truyền thống về độ chính xác, độ nhạy, độ chính xác và điểm F1-score.

Theo Pérez-Almaguer và cộng sự [5], các mô hình gợi ý nhóm dựa trên nội dung (Content-Based Group Recommender Systems – CB-GRS) được đánh giá qua ba phương pháp: mô hình dựa trên luật, mô hình học máy, và mô hình lai. Kết quả thực nghiệm cho thấy mô hình dựa trên luật có độ chính xác 64.3%, Personalization Score 52.1%, và Group Satisfaction 58.7%, nhưng thiếu khả năng thích ứng. Trong khi đó, mô hình học máy đạt hiệu suất cao hơn, với SVM có Precision 79.2%, Recall 76.3%, và F1-score 77.7%, vượt trội hơn Naïve Bayes (F1-score 70.2%) và Decision Trees (F1-score 73.1%). Trong nghiên cứu này, mô hình lai (Hybrid Models) kết hợp lọc cộng tác (CF) và lọc nội dung (CB Filtering) mang lại kết quả tốt nhất. Cụ thể, CF + CB Filtering đạt F1-score 80.4%, trong khi Bi-LSTM + CB Filtering đạt 84.9%, cùng với Personalization Score 82.4% và Group Satisfaction 85.3%, cho thấy khả năng cá nhân hóa và sự hài lòng của nhóm được cải thiện đáng kể.

Theo Xu Z. và cộng sự [15], hệ thống gợi ý bài báo dựa trên nội dung được cải tiến bằng cách phân tích hành vi tìm kiếm thông tin của người dùng đối với ba phần chính của bài báo: nền tảng, phương pháp và kết quả. Mô hình đề xuất sử dụng cấu trúc lọc dựa trên nội dung (CBF), trong đó mỗi phần văn bản được vector hóa và gán trọng số khác nhau tùy theo mức độ ưu tiên của người dùng. Kết quả thực nghiệm trên tập dữ liệu DBLP cho thấy phương pháp này đạt Recall@5 là 81.25% và MAP là 80.81%, vượt trội so với sáu mô hình baseline không phân tách nội dung. Nghiên cứu của nhóm tác giả cho thấy việc điều chỉnh trọng số theo hành vi tìm kiếm giúp cải thiện độ chính xác và mức độ cá nhân hóa của hệ thống, đặc biệt trong các tình huống người dùng có nhu cầu chuyên biệt như là chỉ quan tâm đến phần phương pháp hoặc kết quả nghiên cứu.

Tác giả Neve [14] đã trình bày phương pháp lọc nội dung trong hệ thống gợi ý tương hỗ (Reciprocal Recommender Systems) trong nghiên cứu Content-Based Filtering thuộc Reciprocal Recommender Systems. Tác giả đã đề cập đến hai nghiên cứu điển hình tiêu biểu đó là Recon (sử dụng dữ liệu phân loại để đưa ra gợi ý), và ImRec (dựa vào dữ liệu hình ảnh để thực hiện gợi ý). Nghiên cứu của Neve cũng đề cập đến các khía cạnh quan trọng như trích xuất dữ liệu và triển khai thuật toán lọc nội dung. Ngoài ra, tác giả đã phân tích ưu nhược điểm của các thuật toán lọc nội dung so với lọc cộng tác.

Theo tác giả Amany Sami và cộng sự [13], mô hình gợi ý lại HRS-IU-DL được đề xuất nhằm cải thiện độ chính xác và khả năng cá nhân hóa trong hệ thống gợi ý. Mô hình này kết hợp Lọc cộng tác (CF), Lọc cộng tác thần kinh (NCF), Lọc theo nội dung (CBF), và Mạng nơ-ron hồi quy (RNN) để khắc phục các hạn chế của phương pháp truyền thống như vấn đề dữ liệu thưa thớt và khởi động lạnh. Kết quả thực nghiệm trên bộ dữ liệu MovieLens 100k cho thấy HRS-IU-DL đạt $RMSE = 0.7723$, $MAE = 0.6018$, $Precision = 0.8127$, và $Recall = 0.7312$, vượt trội hơn so với các mô hình khác như SVD ($RMSE = 0.943$, $MAE = 0.744$, $Precision = 0.715$, $Recall = 0.620$) và NCF ($RMSE = 0.935$, $MAE = 0.738$, $Precision = 0.722$, $Recall = 0.628$). Đặc biệt, sau khi tối ưu hóa tham số, mô hình đề xuất giúp cải thiện đáng kể độ chính xác và độ phù hợp của gợi ý, đồng thời xử lý tốt hơn xu hướng thay đổi sở thích của người dùng theo thời gian. Nhìn chung, nghiên cứu này khẳng định rằng việc kết hợp học sâu và thuật toán gợi ý truyền thống trong một hệ thống lai có thể nâng cao chất lượng gợi ý đáng kể.

4. THÁCH THỨC VÀ HƯỚNG NGHIÊN CỨU TƯƠNG LAI

Dựa trên các nghiên cứu được phân tích, có thể thấy rằng hệ thống gợi ý đã có nhiều cải tiến đáng kể nhờ vào sự kết hợp giữa các phương pháp truyền thống như lọc cộng tác với các kỹ thuật tiên tiến như học sâu, phân tích cảm xúc và các mô hình lai. Bảng 1 dưới đây tổng hợp các nghiên cứu liên quan được trình bày ở phần trên, mô tả chi tiết về phương pháp, mô hình sử dụng, tập dữ liệu và kết quả chính của từng nghiên cứu.

Bảng 1. Tổng hợp các nghiên cứu về hệ thống gợi ý

Nghiên cứu	Phương pháp	Mô hình sử dụng	Tập dữ liệu	Kết quả chính
Karabila et al. [7]	SA + CF	Bi-LSTM, SVM, Naïve Bayes	Kindle Book, Amazon Digital Music	Bi-LSTM đạt 93%-94% chính xác. CF + SA cải thiện MAE và RMSE.
Khababa et al. [16]	CF	Hybrid CF, mô hình theo ngữ cảnh	Dịch vụ Web	Hybrid CF cải thiện dự đoán QoS nhưng gặp khó khăn trong môi trường động.
Singhal S. and Pal K [18]	SA + CF	MLA-EDTCNet, MCGAN, OcOA	Sản phẩm Amazon	Cải thiện độ chính xác, độ nhạy, Precision, F1-score so với CF truyền thống.
Pérez-Almaguer et al. [5]	CB-GRS	Luật, Học máy, Hybrid (CF + CB Filtering, Bi-LSTM + CB Filtering)	MovieLens 100 K. HetRec	Hybrid (Bi-LSTM + CB Filtering) đạt F1-score 84.9%, cải

Nghiên cứu	Phương pháp	Mô hình sử dụng	Tập dữ liệu	Kết quả chính
				thiện cá nhân hóa và sự hài lòng nhóm.
Xu Z. [15]	CBF	BERT, SPECTER	Digital Bibliography & Library Project (DBLP)	Recall@5 = 0.8125 và MAP = 0.8081, cao hơn các phương pháp baseline không phân tách nội dung.
Neve [14]	Hệ thống gợi ý tương hỗ	RECON (dữ liệu phân loại), ImRec (dữ liệu hình ảnh)	Dữ liệu dịch vụ hẹn hò trực tuyến Pairs	Phân tích ưu/nhược điểm của lọc nội dung so với lọc cộng tác.
Amany Sami et al. [13]	Hệ thống gợi ý lai	HRS-IU-DL (CF, NCF, CBF, RNN)	MovieLens 100k	RMSE = 0.7723, MAE = 0.6018, cải thiện độ chính xác và khả năng cá nhân hóa.

Bảng 1 tóm tắt những đóng góp quan trọng từ các nghiên cứu khác nhau, làm nổi bật những cải tiến trong hệ thống gợi ý thông qua việc áp dụng các phương pháp học sâu và kỹ thuật lai. Những kết quả đạt được khẳng định rằng sự kết hợp giữa CF, SA, học sâu và các thuật toán tối ưu có thể giúp nâng cao chất lượng và độ chính xác của hệ thống gợi ý, từ đó cải thiện trải nghiệm người dùng.

Tuy nhiên, các hệ thống gợi ý hiện đại vẫn gặp nhiều thách thức, đặc biệt là vấn đề cold start và khả năng mở rộng [9]. Giải pháp tiềm năng bao gồm việc áp dụng mô hình hybrid, kết hợp cả lọc cộng tác và lọc nội dung để tận dụng thông tin từ nhiều nguồn khác nhau [4]. Ngoài ra, các phương pháp học chuyển giao (Transfer Learning) và mô hình học sâu sử dụng metadata cũng đang được nghiên cứu nhằm cải thiện độ chính xác trong bối cảnh dữ liệu hạn chế [1], [12].

Các hướng nghiên cứu trong tương lai có thể tập trung vào việc mở rộng khả năng học của hệ thống gợi ý thông qua tích hợp học sâu với Transfer Learning, nhằm nâng cao khả năng khái quát hóa trong các môi trường dữ liệu đa miền. Đồng thời, việc ứng dụng các kiến trúc hiện đại như Transformer và GPT có tiềm năng lớn trong việc xử lý ngữ cảnh và cải thiện chất lượng gợi ý theo thời gian thực [18]. Bên cạnh đó, các nghiên cứu gần đây cũng nhấn mạnh vai trò của học tăng cường sâu (DRL) trong việc thích ứng với hành vi người dùng động, với các hướng tiếp cận triển vọng như tích hợp Multi-Agent RL và Hierarchical RL để xử lý các tác vụ phức tạp, áp dụng Inverse Reinforcement Learning (IRL) để tự động suy luận hàm thưởng, cũng như kết hợp với Graph Neural Networks (GNNs) nhằm tăng cường khả năng biểu diễn và giải thích của mô hình [19].

5. KẾT LUẬN

Nghiên cứu này tổng quan các phương pháp phổ biến trong hệ thống gợi ý, đồng thời đánh giá hiệu suất của từng phương pháp dựa trên các nghiên cứu thực nghiệm. Kết quả cho thấy mô hình lai kết hợp CF, NCF, CBF và RNN (HRS-IU-DL) vượt trội hơn so với các phương pháp riêng lẻ.

Các hướng nghiên cứu tương lai sẽ tập trung vào việc tích hợp học sâu với Transfer Learning, sử dụng Transformer/GPT cho gợi ý ngữ cảnh, và tối ưu hóa hiệu suất thông qua điện toán đám mây. Việc tiếp tục phát triển các mô hình gợi ý thông minh sẽ mở ra nhiều cơ hội ứng dụng trong thương mại điện tử, truyền thông số, và các lĩnh vực công nghệ tiên tiến khác.

TÀI LIỆU THAM KHẢO

1. Ghadami A. - Bridging Content and Behavior: A Deep Dive into Hybrid Recommendation Systems for Enhancing Personalized User Experiences. University of Ottawa (2024).
2. Ilyes H. - Product Recommendation System for E-Commerce. Univ-Biskra (2024).
3. Bhavani B. and Haritha D. - DLERSE: Deep Learning-Enhanced Recommendation Systems for E-Commerce User Interaction. *J Theor Appl Inf Technol* **103** (2025) 1230–1244.
4. Messaoudi F. and Loukili M. - E-commerce personalized recommendations: A deep neural collaborative filtering approach. *Oper Res Forum* **3** (2024) 66.
5. Pérez-Almaguer Y., Yera R., Alzahrani A. A. and Martínez L. - Content-based group recommender systems: A general taxonomy and further improvements. *Expert Syst Appl* **184** (2021) 115444.
6. Kalidindi R. R. and Bhaskar G. - Movie Recommendation System Using Sentiment Analysis from Micro Blogging Data. *Int J Inf Technol Comput Eng* **12** (2024) 567–580.
7. Karabila I., Khazzani M. E. and Ahaitouf A. - Enhancing collaborative filtering-based recommender system. *Future Internet* **15** (2023).
8. Karnam D. K. - Improving Recommender System Accuracy with Category-Specific Techniques. IIIT Hyderabad (2024).
9. Mishra K. N., Khan A. A. and Tripathi R. C. - Natural Language Processing and Machine Learning-Based Solution of Cold Start Problem. *Electronics* **13** (2024) 4331.
10. Jameii S. M. and Reza A. - Deep learning-based collaborative filtering. *Neural Comput Appl* (2023).
11. Kalideen M. R. and Yağlı C. - Machine Learning-based Recommendation Systems. SEU University (2025).
12. Saifudin I. and Widiyaningtyas T. - Systematic literature review on recommender system. *IEEE Access* (2024).
13. Sami S., El Adrousy W., Sarhan S. and Elmougy S. - A deep learning-based hybrid recommendation model for internet users. *Sci Rep* **14** (2024) 29390.
14. Neve J. - Content-Based Filtering. In: *Reciprocal Recommender Systems*. SpringerBriefs in Computer Science. Springer, Cham (2025).
15. Xu Z., Yamamoto S. and Joho H. - Research Paper Recommender System by Considering Users' Information Seeking Behaviors. *arXiv 2504.02377* (2025).

16. Khababa G. et al. - Collaborative Filtering Techniques for Predicting Web Service QoS Values in Static and Dynamic Environments: A Systematic and Thorough Analysis. *IEEE Access* **13** (2024) 45350–45376.
17. Phalguna K. E. S. et al. - Enhancing E-commerce recommendations with sentiment analysis using MLA-EDTCNet and collaborative filtering. *Sci Rep* **15** (2025) 6739.
18. Singhal S. and Pal K. - State of art and emerging trends on group recommender systems. *Int J Multimedia Inf* (2024).
19. Chen X., Yao L., McAuley J., Zhou G. and Wang X. - Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowl Based Syst* **264** (2023) 110335.

ABSTRACT

SOME METHODS IN BUILDING RECOMMENDATION SYSTEMS: A REVIEW

Dinh Thi Tam^{1,*}

Ho Chi Minh City University of Industry and Trade

Email: tamdt@huit.edu.vn

Recommendation systems play a crucial role in enhancing user experience by analyzing behavior and preferences to provide personalized recommendations. This paper reviews and analyzes advanced methodologies applied in recommender systems, including collaborative filtering, content-based filtering, and hybrid models, as well as modern techniques such as deep learning and reinforcement learning. Notably, the study emphasizes the significance of sentiment analysis, parameters optimization, and strategies for improving system performance in dynamic environments. Additionally, key challenges such as data sparsity, cold start problems, and temporal adaptability are presented, leading to potential research directions aimed at improving the effectiveness of recommendation systems in the future.

Keywords: Recommender systems; Collaborative filtering; Content-based filtering; Hybrid models.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Đinh Thị Tâm

Số điện thoại liên hệ: 0938224968

KHẢO SÁT TÁC ĐỘNG CỦA AI ĐẾN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN TẠI ĐẠI HỌC CÔNG THƯƠNG TP.HỒ CHÍ MINH

Nguyễn Quốc Sử*, Trần Thị Bích Vân¹

¹Trường Đại học Công thương Thành phố Hồ Chí Minh

*Email: sunq@huit.edu.vn

Ngày nhận bài: 28/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Công cụ Trí tuệ Nhân tạo (Artificial Intelligence – AI) đã và đang ngày càng dễ tiếp cận, điều này cũng thay đổi đáng kể cách sinh viên tiếp cận tri thức. Sự phổ biến của ChatGPT, Gemini, ... không chỉ hỗ trợ viết bài, lập trình, kiểm tra lỗi mà còn cá nhân hóa trải nghiệm học tập. Tuy nhiên, liệu AI có thực sự giúp cải thiện kết quả học tập hay chỉ đơn thuần thay đổi phương thức tiếp thu kiến thức? Nghiên cứu này dựa trên bộ dữ liệu khảo sát 523 sinh viên đại học tại trường Đại học Công thương TP.Hồ Chí Minh, các câu hỏi tập trung vào tần suất sử dụng AI, mục đích ứng dụng, mức độ phụ thuộc, nhận thức cá nhân và tác động đối với kết quả học tập. Đây là một trong những nghiên cứu có quy mô mẫu khảo sát lớn, giúp tăng cường tính đại diện và độ tin cậy của kết quả phân tích. Dữ liệu được thu thập thông qua Google Forms và xử lý bằng Microsoft Excel để phân tích và xác định mối tương quan giữa cường độ sử dụng AI và sự thay đổi kết quả học tập.

Kết quả nghiên cứu cho thấy nhóm sinh viên sử dụng AI có chọn lọc và kết hợp với phương pháp học truyền thống có xu hướng cải thiện điểm số, trong khi một số sinh viên dựa hoàn toàn vào AI lại có nhận thức trái chiều về hiệu quả của công cụ này. Đặc biệt, sinh viên thuộc các chuyên ngành liên quan đến công nghệ thông tin và lập trình có xu hướng sử dụng AI với tần suất cao hơn, nhưng không phải lúc nào cũng đạt được sự cải thiện tương ứng về điểm số.

Từ khóa: Trí tuệ nhân tạo trong giáo dục, dữ liệu khảo sát sinh viên, cải thiện điểm số, ứng dụng AI.

1. ĐẶT VẤN ĐỀ

Xu hướng sử dụng công cụ AI của sinh viên đại học đã gia tăng đáng kể trong những năm gần đây. Theo nghiên cứu của nhóm tác giả Đặng Văn Em, Nguyễn Đình Loan Phương, Nguyễn Thị Hảo [1] khoảng 98.1% sinh viên đại học đã sử dụng nền tảng AI là ChatGPT để hỗ trợ quá trình học tập, bao gồm viết bài, lập trình, giải bài tập và tra cứu tài liệu. AI mang lại lợi ích rõ ràng trong việc tăng tốc độ xử lý thông tin và cá nhân hóa nội dung học tập, đồng thời hỗ trợ sinh viên cải thiện kỹ năng giải quyết vấn đề của nhóm tác giả Lê Anh Vinh, Trần Mỹ Ngọc [2]. Theo Nghiên cứu của nhóm tác giả Alberto Grájeda, Johnny Burgos, Pamela

Córdova* and Alberto Sanjinés [3] đã khảo sát cho thấy AI có ảnh hưởng tích cực đến tốc độ xử lý thông tin, khả năng sáng tạo và năng suất học tập. Gần đây, các nhóm tác giả Leon Furze, Mike Perkins, Jasper Roe, Jason MacVaugh [4] đã chỉ ra việc sử dụng AI có thể cải thiện điểm số và nâng cao hiệu quả học tập nếu sinh viên chủ động kiểm chứng thông tin thay vì sao chép thụ động.

Phần lớn các nghiên cứu trước đây chỉ ở qui mô ứng dụng AI của sinh viên trong nghiên cứu [1], một số nghiên cứu khảo sát môi trường ngoài nước. Ngoài ra, trước sự phát triển như vũ bão của các mô hình và công cụ AI mới thì cấp thiết phải có những dữ liệu mới và phù hợp để nghiên cứu và đánh giá chính xác hơn.

Nghiên cứu này khắc phục khoảng trống trên bằng cách thực hiện khảo sát trên 523 sinh viên, được thực hiện tại Trường Đại học Công thương TP.Hồ Chí Minh (HUIT), nhằm có cái nhìn thực tế và đa chiều về tần suất và mục đích sử dụng AI trong học tập, sự thay đổi về điểm số trước và sau khi áp dụng AI, nhận thức của sinh viên về tác động của AI đến kỹ năng tư duy và tự học.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Nghiên cứu được thực hiện trên 523 sinh viên đang theo học tại trường Đại học Công Thương Tp.HCM (HUIT), đại diện cho hai nhóm ngành lần lượt là nhóm Kỹ thuật - Công nghệ với sinh viên Công nghệ thông tin, nhóm Kinh tế – Xã hội: sinh viên Quản trị kinh doanh, Marketing, Thương mại điện tử, một số sinh viên các ngành ngôn ngữ,

Mẫu khảo sát được thu thập đa dạng từ các sinh viên năm nhất đến các sinh viên năm cuối để đảm bảo góc nhìn đa dạng và đầy đủ về trải nghiệm và mức độ sử dụng AI.

2.2. Phương pháp nghiên cứu

2.2.1. Phương pháp thu thập dữ liệu

Dữ liệu được thu thập bằng khảo sát trực tuyến với việc sử dụng Google Forms giúp sinh viên dễ dàng khảo sát hiệu quả và nhanh chóng. Mẫu gồm tổng cộng 21 câu hỏi định tính và định lượng với các phần về thông tin để xác định ngành học, đối tượng niên khóa và đặc biệt 13 câu hỏi trọng tâm nhằm xác định trọng tâm nhằm xác định các yếu tố về thói quen, tần suất sử dụng AI, mục đích, nhận thức và so sánh kết quả khi ứng dụng AI, cũng như gợi mở về công nhận kết quả từ AI trong giáo dục.

2.2.2. Phương pháp xử lý và phân tích dữ liệu

Dữ liệu sau thu thập được nhóm tác giả tối ưu, làm sạch và tổng hợp bằng Microsoft Excel để đảm bảo chính xác. Sau đó tính toán và thống kê để so sánh và xác định các mối tương quan về AI với kết quả học tập cũng như góc nhìn của sinh viên về AI trong học tập.

2.3. Đóng góp của nghiên cứu này

Khác với các nghiên cứu trước đây chủ yếu tập trung vào nhận thức sinh viên về AI, nghiên cứu này đóng góp những điểm mới sau: Quy mô dữ liệu lớn (523 sinh viên) và đa dạng niên khóa (sinh viên năm nhất đến sinh viên năm cuối), đa dạng ngành học (Kỹ thuật - Công nghệ, Tự nhiên – Xã hội) giúp tăng độ tin cậy của phân tích thống kê. Đánh giá sự thay đổi điểm số thực tế trước và sau khi sử dụng AI, thay vì chỉ khảo sát ý kiến chủ quan. Phân loại

sinh viên theo tần suất sử dụng AI để xác định mức độ tối ưu trong việc cải thiện hiệu suất học tập. Đưa ra những cảnh báo về nguy cơ lệ thuộc AI trong học tập, thay vì chỉ đánh giá lợi ích. Nghiên cứu này không chỉ giúp giảng viên, nhà quản lý giáo dục và sinh viên tại HUIT có cái nhìn thực tế hơn về tác động của AI với kết quả học tập, mà còn cung cấp bằng chứng thực nghiệm để nghiên cứu và cải tiến, thậm chí điều chỉnh hành vi để sinh viên có thể có kết quả học tập tốt hơn thông qua việc khai thác sức mạnh ngày càng tăng nhanh chóng của AI.

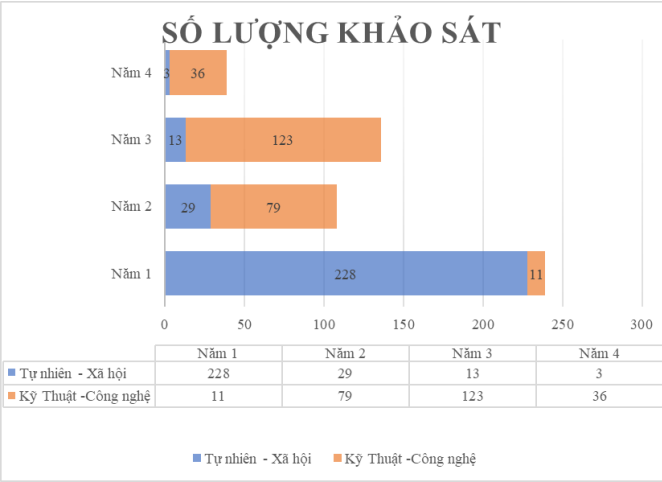
2.4. Hạn chế của nghiên cứu:

Dữ liệu tự báo cáo có thể bị ảnh hưởng bởi sai số chủ quan của sinh viên khi đánh giá điểm số của mình. Không thể kiểm soát tất cả các yếu tố bên ngoài (phương pháp giảng dạy, tài liệu tham khảo, động lực học tập cá nhân,...). Chưa đánh giá tác động dài hạn, vì nghiên cứu chủ yếu tập trung vào sự thay đổi điểm số trong một khoảng thời gian ngắn.

3. KẾT QUẢ VÀ THẢO LUẬN

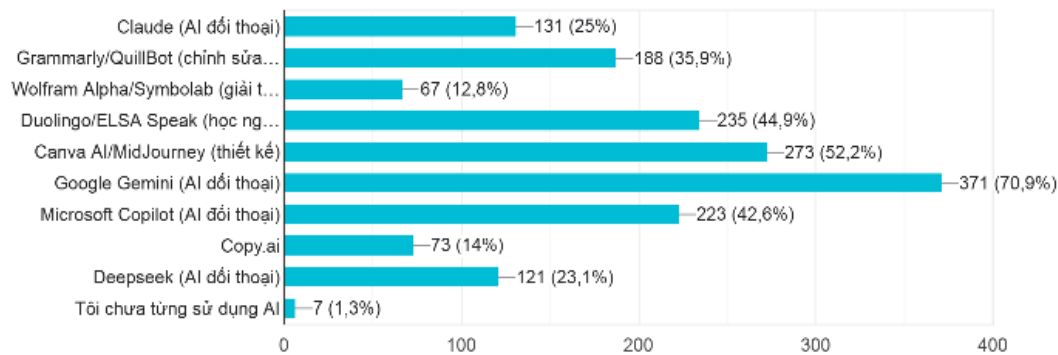
3.1. Đối tượng tham gia khảo sát

Nhóm tác giả thực hiện nghiên cứu bằng khảo sát hỏi đáp và thống kê mô tả trên dữ liệu 523 sinh viên (Hình 1), với 249 sinh viên nhóm A (Sinh viên Khoa Công nghệ Thông tin với các nhóm ngành đại diện cho Khoa học – Công nghệ) và 274 nhóm B (Tự nhiên – Xã hội với các ngành Ngôn ngữ, Quản trị kinh doanh, Kinh tế, truyền thông,...). Tỷ lệ giữa 2 nhóm này lần lượt là 48% và 52% để thấy được tính cân đối của khảo sát.



Hình 1. Dữ liệu khảo sát

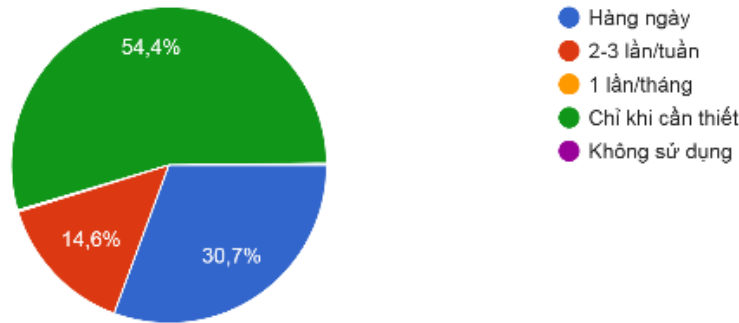
Tính đến thời điểm tháng 03/2025, các đối tượng tham gia phỏng vấn về trải nghiệm AI của bản thân và kết quả đạt được khi sử dụng các công cụ AI được chia làm ba nhóm: nhóm sử dụng AI đối thoại để giải quyết tình huống, nhóm sử dụng các AI hỗ trợ để viết bài luận, học ngoại ngữ và nhóm còn lại có xu hướng sử dụng các AI thiết kế (Hình 2)



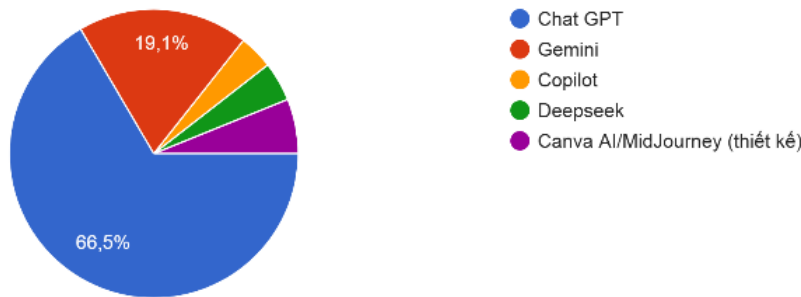
Hình 2: Tỷ lệ sử dụng các loại công cụ AI

3.2. Tần suất và công cụ

Trong số 523 sinh viên được khảo sát có 54.3% chỉ sử dụng AI khi cần thiết, cho thấy sinh viên xem AI với vai trò hỗ trợ hơn là thay thế phương pháp học tập truyền thống. Tuy nhiên, có đến 30.7% sinh viên sử dụng AI hàng ngày cho việc học tập, cho thấy một bộ phận sinh viên đang dựa vào công nghệ này trong quá trình học tập (Hình 3). Trong số 85.6% công cụ AI sinh viên ưu tiên sử dụng là các AI đối thoại thông dụng như Chat GPT, Gemini, ngoài ra các AI thiết kế cũng được khai thác, điều này cho thấy đã có sự đa dạng hóa trong việc lựa chọn các công cụ AI hỗ trợ việc học của sinh viên hiện nay (Hình 4).



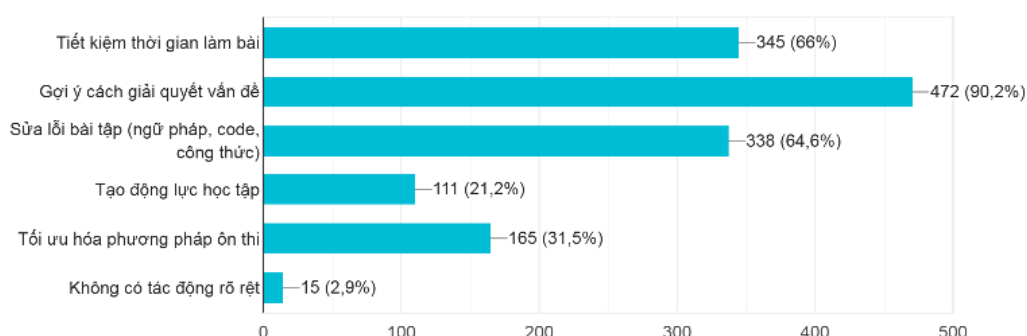
Hình 3: Tần suất sử dụng AI



Hình 4. Tỷ lệ sử dụng các công cụ AI

3.3 Tác động của AI đến kết quả học tập

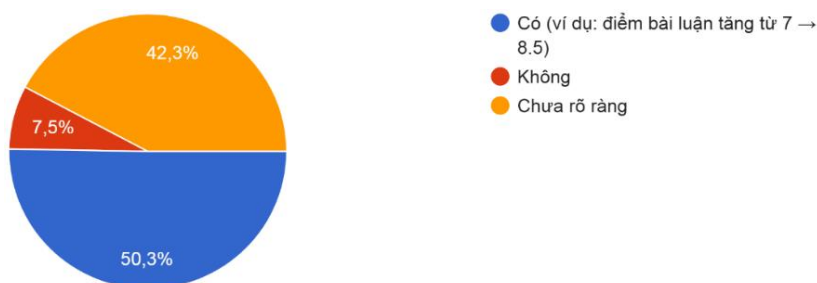
Kết quả khảo sát cho thấy gần như 90.2%, tương đương 472 sinh viên sử dụng AI để gợi ý cách giải quyết vấn đề (Hình 5). Điều này cho thấy sự liên quan của AI đến các môn học cần tư duy như toán hoặc lập trình, hay các lĩnh vực khoa học kỹ thuật. Theo sau đó cũng là thói quen phụ thuộc vào AI với 66% tức tương đương 345 sinh viên chọn phương án “*Tiết kiệm thời gian làm bài*”, tiếp đó là phương án “*Sửa lỗi bài tập*” với 338 sinh viên chọn. Mặc dù trong khảo sát có thể thấy có một số ít khoảng 2.9% nhận thấy “*Không có tác dụng rõ rệt*”, tỷ lệ này cũng có thể bao gồm nhóm sinh viên có kỹ năng sử dụng AI thấp. Cần phải lưu ý rằng điều này cũng phản ánh việc các gợi ý hoặc hướng dẫn tiếp cận từ AI là yếu tố then chốt quan trọng ảnh hưởng trực tiếp đến việc cải thiện điểm số. Xét về mặt tích cực nó có thể hỗ trợ sinh viên với các gợi ý nhanh, bài mẫu giải sẵn,.. Mặt khác, đây là thói quen nguy hiểm khi kết quả học tập có sự trợ giúp quá nhiều từ AI thay vì nỗ lực học tập theo phương pháp truyền thống.



Hình 5. Tác động của AI đến kết quả học tập

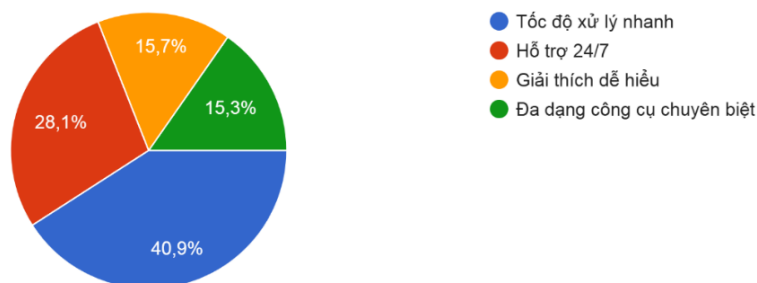
Nghiên cứu cho thấy khoảng 50.3% sinh viên nhìn ra được có sự cải thiện điểm số, chứng tỏ AI có tác động tích cực trong hỗ trợ học tập. Tuy nhiên, có đến 42.3% chưa thấy rõ sự thay đổi, điều này có thể do sinh viên chưa sử dụng AI đúng cách hoặc chỉ dùng AI để hỗ trợ thay vì nâng cao tư duy cá nhân. Từ kết quả khảo sát và phân tích, mục tiêu chính của nghiên cứu này là việc cải thiện điểm số, có thể thấy khi sử dụng AI, có sự cải thiện điểm số rõ rệt với thang tăng 1.5 điểm tương đương 15%. Ngoài ra, với tỷ lệ “chưa thấy rõ sự thay đổi” hoặc “Không thay đổi” nguyên nhân chính là do kỹ năng khai thác AI còn chưa tốt hoặc chỉ dùng AI để hỗ trợ chứ không để nâng cao tư duy và kết quả học tập cá nhân.

3.4. Góc nhìn về ưu điểm và hạn chế của AI với sinh viên



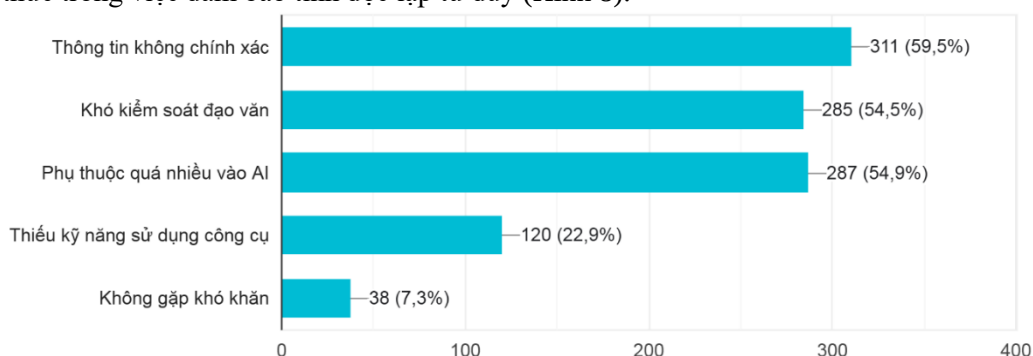
Hình 6. Tỷ lệ cải thiện điểm số với AI

Kết quả đánh giá khi sử dụng AI của sinh viên, nghiên cứu thu được nhiều góc nhìn đa dạng, phần lớn (40.9%) cho rằng tốc độ xử lý của AI nhanh, có thể phản hồi về các gợi ý, giải bài tập, sửa lỗi nhanh, với tỷ lệ 28.1% sự ưa thích dành cho tiêu chí “Hỗ trợ liên tục 24/7” của AI, điều này chứng tỏ ưu điểm luôn sẵn sàng của các công cụ AI, là một trợ thủ đắc lực giúp cho sinh viên tiết kiệm được nhiều thời gian và có thể linh động với giờ giấc học tập của mình. Đây cũng là một yếu tố giúp cải thiện kết quả học tập cho sinh viên.



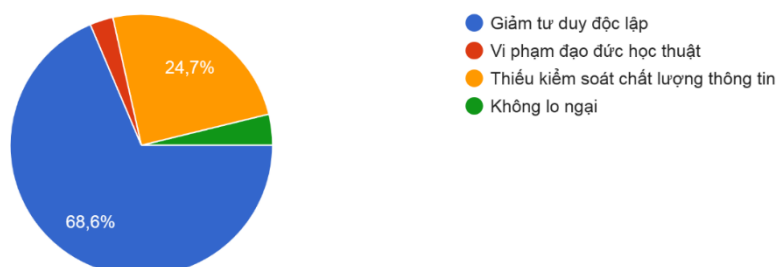
Hình 7. Đánh giá ưu điểm của AI

Bên cạnh đó, một số hạn chế của AI cũng đã được sinh viên ghi nhận, có thể ảnh hưởng đến kết quả học tập, thậm chí thay vì là công cụ hữu ích lại là rào cản. Sinh viên chủ yếu lo ngại về chất lượng thông tin AI cung cấp (59.6%) và vấn đề đạo văn (54.6%). Đáng chú ý, 55% sinh viên nhận thấy mình có xu hướng phụ thuộc quá nhiều vào AI, điều này đặt ra thách thức trong việc đảm bảo tính độc lập tư duy (Hình 8).



Hình 8. Một số hạn chế của AI

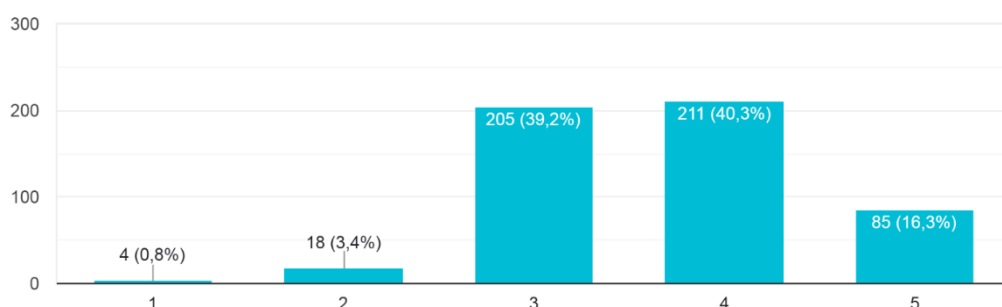
Mặc dù hiện nay có khá nhiều ý kiến tranh cãi về vấn đề AI có thể gây ảnh hưởng đến kết quả học tập từ kiến thức đến kỹ năng của sinh viên, thậm chí có thể tạo ra **“vùng thoải mái kỹ thuật số”**, có thể gây ra tình trạng ỷ lại và giả học từ AI của sinh viên do lạm dụng AI quá mức. Tuy nhiên, với kết quả khảo sát thu được từ nghiên cứu, sinh viên HUIT luôn có ý thức rất cao trong việc sử dụng AI trong học tập, với tỷ lệ lần lượt 68.6%, 24.7% và tỷ lệ nhỏ tương ứng với giảm tư duy độc lập, thiếu kiểm soát chất lượng và các vấn đề vi phạm đạo đức học thuật. Điều này có thể giúp cho những người đang trăn trở hoặc các nhà giáo dục có thể yên tâm phần nào đó và tin tưởng về ý thức học tập của sinh viên trước kỷ nguyên AI.



Hình 9. Một số điểm e ngại khi sử dụng AI

3.5. Mức độ hiệu quả của AI

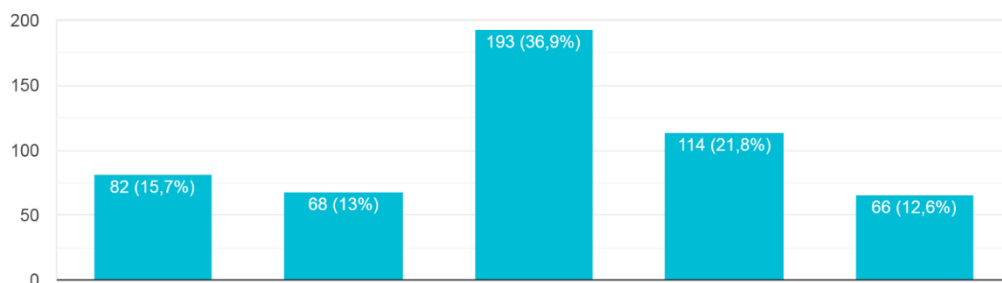
Các câu hỏi định lượng nhằm tìm ra sự công nhận của sinh viên về điểm số khi sử dụng AI, khi yêu cầu đánh giá mức độ cải thiện kết quả học tập từ mức 1 - Không hiệu quả đến mức 5 - Rất hiệu quả, kết quả hầu hết nằm ở mức hiệu quả trung bình đến cao (3-4 điểm) cho thấy phần lớn sinh viên công nhận vai trò hỗ trợ của AI trong học tập.



Hình 10. Hiệu quả của việc sử dụng AI trong việc cải thiện kết quả học tập

3.6. Thái độ của sinh viên về giảng viên sử dụng AI để giảng dạy

Kết quả khi được gọi mở về đề nghị lựa chọn học tập với hình thức giảng dạy truyền thống hay một giảng viên sử dụng AI cho giảng dạy, với mức 1 -Không muốn đến 5 – Hoàn toàn mong muốn, kết quả cho thấy sinh viên có thái độ trung lập đến tích cực với giảng viên sử dụng AI trong giảng dạy. 34.2% mong muốn học với giảng viên dùng AI (mức 4-5), trong khi 15.7% không mong muốn (mức 1-2) (Hình 11). Điều này phản ánh sự quan tâm của sinh viên đối với công nghệ nhưng cũng cho thấy sự cần thiết của một phương pháp giảng dạy kết hợp AI một cách hợp lý.



Hình 11. Mức độ mong muốn GV áp dụng AI trong giảng dạy

4. KẾT LUẬN

Dựa trên dữ liệu thu thập được về mức độ sử dụng AI trong học tập của sinh viên tại HUIT cho thấy AI đã trở thành một công cụ phổ biến, nhưng mức độ và cách thức sử dụng có sự khác biệt đáng kể giữa các nhóm sinh viên. Đầu tiên, nói về cải thiện kết quả học tập, khoảng 50.1% sinh viên nhận thấy sự cải thiện rõ rệt về điểm số sau khi sử dụng AI, chủ yếu nhờ tiết kiệm thời gian, hỗ trợ giải quyết vấn đề và tối ưu hóa phương pháp ôn tập. Tuy nhiên, 42.3% vẫn chưa thấy thay đổi đáng kể, cho thấy việc sử dụng AI chưa thực sự tối ưu hoặc có sự phụ thuộc quá mức mà không phát triển kỹ năng cá nhân. Một số kết luận quan trọng khác có thể được rút ra từ thực tế này đó là mức độ sử dụng AI trong học tập với phần lớn sinh viên (54.2%) chỉ sử dụng AI khi cần thiết, trong khi 30.6% sử dụng hàng ngày. Điều này cho thấy AI đóng vai trò hỗ trợ nhưng chưa thực sự thay thế hoàn toàn phương pháp học truyền thống. Ngoài ra sinh viên cũng đánh giá cao về AI với tốc độ xử lý nhanh và giải thích dễ hiểu cũng như có gần 70% ủng hộ giảng viên sử dụng AI trong giảng dạy, điều này có thể cũng cổ thêm về tiềm năng của AI cải thiện chất lượng giảng dạy.

Tóm lại, sự dần phổ biến và gia tăng nhanh về sức mạnh của các công cụ AI không chỉ là một bước tiến công nghệ quan trọng, đóng góp đáng kể đến nhiều lĩnh vực mà nó còn tác động lớn đến giáo dục, đặc biệt là giáo dục đại học. Sử dụng AI có thể nâng cao hiệu quả và kết quả học tập của đa số người học tuy nhiên cần thiết hướng dẫn sinh viên sử dụng AI một cách khoa học và có trách nhiệm, thêm vào đó là tập trung và việc kiểm chứng thông tin cũng như phát triển tư duy phản biện và giải quyết vấn đề.

TÀI LIỆU THAM KHẢO

- [1] Đặng Văn Em, Nguyễn Đình Loan Phương, Nguyễn Thị Hào, “*Thực trạng ứng dụng ChatGPT trong việc học tập, nghiên cứu của sinh viên Đại học Quốc gia Thành phố Hồ Chí Minh*”, Tạp chí Giáo dục, vol. 24, no. 1, pp. 36-41, 2024.
- [2] Lê Anh Vinh, Trần Mỹ Ngọc, “*Tác động của Trí tuệ nhân tạo (AI) đối với hệ thống*”, Tạp chí giáo dục, vol. Tập 20, no. Số 05, pp. 1-15, 2024.
- [3] Alberto Grájeda, Johnny Burgos, Pamela Córdova* and Alberto Sanjinés, "Assessing the student-perceived impact of using artificial intelligence tools: Construction of a synthetic index of application in higher education," Cogent Education, vol. 11, pp. 1-24, 2024.
- [4] Leon Furze, Mike Perkins, Jasper Roe, Jason MacVaugh, “*The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI supported assessment*”, Australasian Journal of Educational, vol. 40, pp. 38-55, (2024).

ABSTRACT

A STUDY ON THE EFFECTS OF AI ON STUDENT LEARNING AT HO CHI MINH CITY UNIVERSITY OF INDUSTRY AND TRADE

Quoc-Su NGUYEN^{1*}, Bich-Van Thi TRAN¹

Ho Chi Minh City University of Industry and Trade

Artificial Intelligence (AI) tools have become increasingly accessible, significantly transforming how students acquire knowledge. The widespread adoption of ChatGPT, Gemini, and similar AI technologies not only supports writing, programming, and error detection but also personalizes the learning experience. However, does AI truly enhance academic performance, or does it merely alter how knowledge is acquired?

This study is based on survey data from 523 university students at Ho Chi Minh City University of Industry and Trade (HUIT). The questionnaire focused on AI usage frequency, application purposes, dependency levels, personal perceptions, and its impact on academic performance. This research features a large sample size, enhancing the representativeness and reliability of the analysis. Data was collected via Google Forms and processed using Microsoft Excel to analyze correlations between AI usage intensity and changes in academic performance.

The findings indicate that students who selectively use AI with traditional learning methods tend to improve their grades. In contrast, some students who rely entirely on AI have mixed perceptions regarding its effectiveness. Notably, students in technology-related and programming disciplines tend to use AI more frequently, but this does not always correlate with improved academic performance.

Keywords: Artificial Intelligence in Education, Student Survey Data, Academic Performance Improvement, AI Applications.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Quốc Sĩ

Số điện thoại liên hệ: 0938269603

PHỤ LỤC
BẢNG CÂU HỎI KHẢO SÁT

STT	Câu hỏi	Loại câu hỏi
1	(1) Bạn đã từng sử dụng hệ thống AI nào bên dưới (có thể chọn nhiều câu trả lời)? Claude (AI đối thoại) Grammarly/QuillBot (chỉnh sửa văn bản) Wolfram Alpha/Symbolab (giải toán) Duolingo/ELSA Speak (học ngoại ngữ) Canva AI/MidJourney (thiết kế) Google Gemini (AI đối thoại) Microsoft Copilot (AI đối thoại) Copy.ai Deepseek (AI đối thoại) Tôi chưa từng sử dụng AI	Lựa chọn nhiều
2	(2) Tần suất sử dụng AI trong học tập của bạn: Hàng ngày 2-3 lần/tuần 1 lần/tháng Chỉ khi cần thiết Không sử dụng	Lựa chọn đơn
3	(3) Nếu chỉ được phép sử dụng một công cụ AI để sử dụng trong một ngày học tập , bạn sẽ sử dụng công cụ nào? Chat GPT Gemini Copilot Deepseek Canva AI/MidJourney (thiết kế)	Lựa chọn đơn
4	(4) Theo bạn, AI giúp cải thiện điểm số thông qua những yếu tố nào? (Chọn nhiều đáp án) Tiết kiệm thời gian làm bài Gợi ý cách giải quyết vấn đề Sửa lỗi bài tập (ngữ pháp, code, công thức) Tạo động lực học tập Tối ưu hóa phương pháp ôn thi Không có tác động rõ rệt	Lựa chọn nhiều
5	(5) Bạn đã từng trải nghiệm cải thiện điểm số sau khi sử dụng AI chưa? Có (ví dụ: điểm bài luận tăng từ 7 → 8.5) Không	Câu hỏi định tính

STT	Câu hỏi	Loại câu hỏi
	Chưa rõ ràng	
6	(6) Điều bạn thích nhất khi dùng AI để học tập là gì? Tốc độ xử lý nhanh Hỗ trợ 24/7 Giải thích dễ hiểu Đa dạng công cụ chuyên biệt	Lựa chọn đơn
7	(7) Bạn gặp khó khăn gì khi sử dụng AI cho học tập? (Chọn nhiều đáp án) Thông tin không chính xác Khó kiểm soát đạo văn Phụ thuộc quá nhiều vào AI Thiếu kỹ năng sử dụng công cụ Không gặp khó khăn	Lựa chọn nhiều
8	(8) Bạn có đồng ý rằng AI nên được tích hợp vào giáo dục chính thức (có nghĩa là sẽ tính điểm cho kết quả thực hiện bằng AI) ? Hoàn toàn đồng ý Đồng ý Trung lập Không đồng ý Phản đối	Lựa chọn đơn
9	(9) Điều bạn lo ngại nhất về việc sử dụng AI trong học tập là gì? Giảm tư duy độc lập Vi phạm đạo đức học thuật Thiếu kiểm soát chất lượng thông tin Không lo ngại	Lựa chọn đơn
10	(10) Đánh giá mức độ AI giúp bạn cải thiện kết quả học tập 1 (Không hiệu quả) → 5 (Rất hiệu quả)	Câu hỏi định lượng
11	(11) Bạn có sẵn sàng giới thiệu AI như một công cụ học tập/Hỗ trợ thi, kiểm tra cho người khác?	Lựa chọn đơn
12	(12) Nếu một sinh viên học tập chăm chỉ không sử dụng AI và kết quả học tập đạt thang điểm 10, theo bạn một sinh viên sử dụng AI toàn bộ thời gian học tập thì kết quả nên ở thang điểm mấy?	Câu hỏi định lượng
13	(13) Nếu được chọn học với một giảng viên truyền thống giảng dạy dưới hình thức Thuyết giảng/Trình chiếu và một giảng viên sử dụng AI để giảng dạy , mức độ mong muốn học tập với giảng viên sử dụng AI để giảng dạy của bạn theo thang 1(Không muốn) → 5 (Hoàn toàn mong muốn)	Câu hỏi định lượng

TÍCH HỢP CÔNG NGHỆ THỬ KÍNH ẢO THỜI GIAN THỰC TRÊN WEBSITE BÁN MẮT KÍNH THỜI TRANG DỰA TRÊN FACEMESH VÀ DLIB

Nguyễn Trọng Hoàng Phúc¹, Ngô Dương Hà¹, Trần Như Ý^{1*}

Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: ytn@huit.edu.vn

Ngày nhận bài: 30/3/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Việc tích hợp công nghệ thử kính ảo (Virtual Try-On – VTO) vào các nền tảng thương mại điện tử đang trở thành một xu hướng nổi bật, đặc biệt trong lĩnh vực kinh doanh mắt kính thời trang. Công nghệ này cho phép người dùng trải nghiệm thử kính trực tuyến thông qua mô hình 3D, từ đó nâng cao chất lượng trải nghiệm mua sắm, tăng khả năng ra quyết định và giảm tỷ lệ hoàn trả sản phẩm. Trong bối cảnh mua sắm trực tuyến phát triển mạnh mẽ sau đại dịch Covid-19, VTO ngày càng đóng vai trò thiết yếu trong hoạt động kinh doanh trực tuyến. Tuy nhiên, một số thách thức kỹ thuật vẫn tồn tại, đáng chú ý là hạn chế về độ chính xác và khả năng tương tác thời gian thực khi hiển thị các sản phẩm thời trang trên khuôn mặt người dùng. Bài báo này đề xuất và đánh giá hiệu quả của hai mô hình nhận diện khuôn mặt, bao gồm FaceMesh và Dlib, trong việc hỗ trợ chức năng thử kính ảo trực tuyến. Các mô hình được tích hợp vào hệ thống thông qua TensorFlow.js và Three.js nhằm hiển thị kính 3D theo thời gian thực. Kết quả thực nghiệm cho thấy cả hai mô hình đều hoạt động hiệu quả, tuy nhiên FaceMesh cho kết quả vượt trội hơn về độ chính xác và khả năng tương tác, đặc biệt trong việc theo dõi khuôn mặt từ nhiều góc nhìn khác nhau. Hệ thống thử kính ảo được đề xuất góp phần nâng cao trải nghiệm người dùng và mở ra hướng phát triển tiềm năng cho các ứng dụng thương mại điện tử trong tương lai.

Từ khóa: AR, AI, Virtual Try-on, FaceMesh.

1. MỞ ĐẦU

Sự gia tăng mạnh mẽ của nhu cầu mua sắm trực tuyến trong những năm gần đây, đặc biệt là trong giai đoạn giãn cách xã hội do đại dịch Covid-19, đã làm thay đổi đáng kể hành vi tiêu dùng trên toàn cầu. Một trong những khó khăn lớn mà người tiêu dùng gặp phải là không thể trực tiếp đến cửa hàng để thử sản phẩm, đặc biệt là với các phụ kiện thời trang như mắt kính – vốn yêu cầu cao về tính vừa vặn và thẩm mỹ. Điều này đặt ra thách thức trong việc duy trì trải nghiệm mua sắm tích cực và hạn chế tỷ lệ trả hàng – một trong những vấn đề phổ biến trong thương mại điện tử.

Trong bối cảnh đó, công nghệ thử ảo (Virtual Try-On – VTO) nổi lên như một giải pháp hiệu quả, cho phép người dùng thử sản phẩm trực tuyến thông qua hình ảnh ảo mà không cần đến cửa hàng. VTO không chỉ nâng cao tính tương tác và cá nhân hóa trong trải nghiệm mua sắm, mà còn góp phần giảm thiểu rủi ro hoàn trả sản phẩm, từ đó cải thiện hiệu quả hoạt động kinh doanh cho các doanh nghiệp thương mại điện tử [1,3].

Mặc dù VTO đã được nghiên cứu và ứng dụng trong nhiều lĩnh vực, điểm mới của nghiên cứu này là đề xuất và triển khai giải pháp thử kính ảo thời gian thực trực tiếp trên nền tảng web bằng cách tích hợp hai công nghệ hiện đại là TensorFlow.js và Three.js. Cách tiếp cận này không chỉ giúp tiết kiệm chi phí phát triển mà còn tăng tính tiện lợi khi người dùng không cần cài đặt phần mềm bổ sung. Tuy nhiên, việc tối ưu hóa hiệu năng và độ chính xác của VTO trên trình duyệt web vẫn đang là một thách thức. Các nghiên cứu hiện tại chủ yếu tập trung vào việc ứng dụng trí tuệ nhân tạo và học sâu để nâng cao khả năng nhận diện và mô phỏng sản phẩm, song khả năng tái hiện hình ảnh sản phẩm 3D một cách chính xác và mượt mà trên khuôn mặt người dùng vẫn chưa đạt mức tối ưu, đặc biệt là khi xử lý trong thời gian thực [3].

Nghiên cứu này nhằm xây dựng hệ thống thử kính ảo trực tuyến thông qua việc tích hợp hai mô hình nhận diện khuôn mặt là FaceMesh và Dlib, từ đó đánh giá hiệu quả của từng mô hình trong việc hiển thị kính 3D trên khuôn mặt theo thời gian thực. Mục tiêu cuối cùng là lựa chọn mô hình tối ưu, góp phần cải thiện trải nghiệm người dùng và tăng tính ứng dụng thực tiễn của công nghệ VTO trong lĩnh vực thương mại điện tử.

2. CƠ SỞ LÝ THUYẾT

2.1. Mô hình FaceMesh

FaceMesh là một mô hình nhận diện khuôn mặt tiên tiến dựa trên học sâu (deep learning), có khả năng phát hiện chính xác 468 điểm đặc trưng (landmark) trên khuôn mặt người dùng. Các điểm này bao phủ các khu vực quan trọng như mắt, mũi, miệng và cằm, từ đó cho phép mô hình tái hiện chính xác hình dạng và vị trí của khuôn mặt [4]. Nhờ vậy, các ứng dụng như thử kính ảo có thể đặt và điều chỉnh mô hình kính 3D theo thời gian thực, bám sát chuyển động và vị trí khuôn mặt được ghi lại từ hình ảnh webcam.

FaceMesh kết hợp các kỹ thuật xử lý hình ảnh hiện đại và mạng học sâu để phân tích và nhận diện đặc điểm khuôn mặt một cách hiệu quả. Mô hình này hoạt động trực tiếp trong trình duyệt nhờ vào nền tảng TensorFlow.js, cho phép xử lý thời gian thực mà không cần cài đặt phần mềm bổ sung, từ đó nâng cao tính linh hoạt và tối ưu trải nghiệm người dùng trên các nền tảng thương mại điện tử [5].

2.1.1. Nhận diện điểm landmark

Trong mô hình FaceMesh, các điểm landmark trên khuôn mặt được xác định dựa trên việc phân tích hình ảnh từ camera. Các điểm landmark được xác định bằng tọa độ 2D (x, y) trong không gian ảnh và có thể được biểu diễn dưới dạng vector:

$$p_i = (x_i, y_i), \text{ với } i = 1, 2, \dots, N$$

Trong đó:

p_i là tọa độ của điểm landmark thứ i ,

x_i, y_i là tọa độ của điểm đó trong không gian ảnh 2D,

N là tổng số điểm landmark (468 điểm trong mô hình FaceMesh).

2.1.2. Trung điểm giữa hai mắt

Để xác định trung điểm giữa hai mắt trong ứng dụng thử kính 3D, trước tiên cần trích xuất tọa độ của mắt trái và mắt phải từ các điểm landmark do mô hình FaceMesh của TensorFlow.js cung cấp. Các landmark này biểu thị vị trí của các đặc điểm khuôn mặt, bao gồm các điểm tại góc trong và ngoài của mỗi mắt. Khi đã thu được tọa độ của hai điểm đại diện cho mắt trái và mắt phải, trung điểm giữa hai mắt có thể được tính bằng cách lấy trung bình cộng của tọa độ tương ứng theo công thức:

$$MidEye = \left(\frac{x_{leftEye} + x_{rightEye}}{2}, \frac{y_{leftEye} + y_{rightEye}}{2} \right)$$

Trong đó:

$(x_{leftEye}, y_{leftEye})$ là tọa độ của mắt trái,

$(x_{rightEye}, y_{rightEye})$ là tọa độ của mắt phải.

2.1.3. Khoảng cách giữa hai mắt

Khoảng cách giữa hai mắt là một yếu tố quan trọng trong việc hiệu chỉnh kích thước của mô hình kính 3D sao cho phù hợp với khuôn mặt người dùng. Khoảng cách này được tính toán dựa trên công thức khoảng cách Euclid, áp dụng cho hai điểm có tọa độ lần lượt là (x_1, y_1) và (x_2, y_2) , tương ứng với vị trí của mắt trái và mắt phải trên khuôn mặt.

$$\theta = \arctan\left(\frac{y_{rightEye} - y_{leftEye}}{x_{rightEye} - x_{leftEye}}\right)$$

Trong đó:

θ là góc xoay của kính (góc roll),

$x_{leftEye}, x_{rightEye}$ là tọa độ ngang của mắt trái và mắt phải,

$y_{leftEye}, y_{rightEye}$ là tọa độ dọc của mắt trái và mắt phải.

Khoảng cách này phản ánh độ rộng thực tế giữa hai mắt của người dùng, đóng vai trò then chốt trong việc điều chỉnh kích thước mô hình kính 3D để đảm bảo tỷ lệ hiển thị phù hợp với khuôn mặt. Việc tính toán chính xác khoảng cách giữa hai mắt giúp mô hình kính 3D được hiển thị với tỷ lệ đúng đắn, tránh tình trạng quá lớn hoặc quá nhỏ, từ đó mang lại trải nghiệm thử kính trực quan và chân thực hơn cho người dùng.

2.2. Dlib

Dlib là một thư viện mã nguồn mở được thiết kế nhằm cung cấp các công cụ hỗ trợ cho các bài toán trong lĩnh vực học máy và thị giác máy tính. Trong nhận diện khuôn mặt, Dlib áp dụng các phương pháp học máy truyền thống, nổi bật là kỹ thuật Histogram of Oriented Gradients (HOG) kết hợp với mô hình Support Vector Machine (SVM) để phát hiện khuôn mặt trong ảnh. Sau khi khuôn mặt được phát hiện, Dlib tiếp tục sử dụng một mô hình hồi quy để xác định các điểm đặc trưng (landmark) trên khuôn mặt, bao gồm các vị trí quan trọng

như mắt, mũi, miệng và cằm [6]. Các điểm landmark này đóng vai trò quan trọng trong việc phân tích cấu trúc khuôn mặt và hỗ trợ các ứng dụng như căn chỉnh, theo dõi và thử kính ảo.

2.2.1. Phát hiện khuôn mặt với HOG và SVM

Dlib sử dụng kỹ thuật Histogram of Oriented Gradients (HOG) để phát hiện khuôn mặt trong ảnh. Phương pháp này hoạt động bằng cách phân tích gradient – bao gồm hướng và độ lớn – của các pixel, sau đó xây dựng các histogram thể hiện phân bố hướng gradient trong từng vùng nhỏ của ảnh. Các histogram này được kết hợp lại để tạo thành một đặc trưng tổng quát, giúp mô hình phân biệt khuôn mặt với các đối tượng khác trong ảnh một cách hiệu quả.

Công thức cơ bản cho gradient của một điểm ảnh (x, y) trong một ảnh số hóa được tính như sau:

$$G_x(x, y) = I(x + 1, y) - I(x - 1, y)$$

$$G_y(x, y) = I(x, y + 1) - I(x, y - 1)$$

Trong đó:

$G_x(x, y)$ và $G_y(x, y)$ là gradient theo trục ngang và trục dọc

$I(x, y)$ là giá trị độ sáng tại vị trí x, y của ảnh

Histogram của các hướng gradient sau đó được tính toán cho từng vùng trong ảnh. Kết quả từ HOG được đưa vào mô hình SVM để phân loại liệu một vùng trong ảnh có chứa khuôn mặt hay không.

2.2.2. Dự đoán điểm landmark với hàm hồi quy

Sau khi phát hiện khuôn mặt, Dlib áp dụng phương pháp Ensemble of Regression Trees để dự đoán các điểm đặc trưng (landmark) trên khuôn mặt. Đây là một kỹ thuật hồi quy sử dụng tập hợp các cây quyết định (regression trees), trong đó mỗi cây quyết định đóng vai trò như một bộ dự đoán riêng biệt. Các kết quả dự đoán từ từng cây quyết định được tổng hợp lại để đưa ra vị trí cuối cùng của từng điểm landmark, giúp tăng độ chính xác và khả năng khái quát của mô hình.

Về mặt toán học, quá trình dự đoán vị trí của các điểm landmark $\hat{y}_i$ được mô hình hóa như một hàm hồi quy dựa trên đặc trưng hình ảnh x , cụ thể như sau:

$$\hat{y}_i = f(x) = \sum_{j=1}^N w_j h_j(x)$$

Trong đó:

$\hat{y}_i$ là vị trí của điểm landmark thứ i .

$h_j(x)$ là các hàm hồi quy trong mô hình.

w_j là trọng số của hàm hồi quy h_j .

N là tổng số hàm hồi quy trong mô hình.

Phương pháp này cho phép mô hình học được các quan hệ phi tuyến giữa đặc trưng hình ảnh và vị trí landmark, đồng thời duy trì tốc độ xử lý nhanh, phù hợp với các ứng dụng thời gian thực như thử kính ảo.

2.3. Three.js

Three.js là một thư viện đồ họa 3D mạnh mẽ được xây dựng trên nền tảng WebGL, cho phép hiển thị và tương tác với các đối tượng 3D trực tiếp trong trình duyệt web mà không cần cài đặt thêm phần mềm bổ trợ. Nhờ khả năng kết xuất mô hình 3D theo thời gian thực, Three.js đóng vai trò quan trọng trong việc phát triển các ứng dụng Virtual Try-On (VTO), đặc biệt trong lĩnh vực thương mại điện tử [1]. Việc tích hợp Three.js vào các nền tảng mua sắm trực tuyến mang đến trải nghiệm trực quan, sinh động cho người dùng, từ đó nâng cao mức độ tương tác và hỗ trợ quyết định mua hàng.

Kiến trúc hoạt động của Three.js bao gồm bốn thành phần chính: Scene, Camera, Canvas và Renderer, như minh họa trong Hình 1 [7, 8].

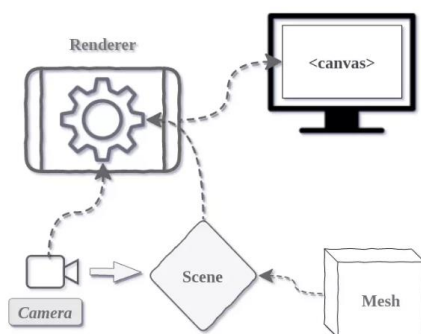
Scene là không gian trung tâm chứa tất cả các đối tượng 3D như mô hình kính, ánh sáng và các yếu tố môi trường. Các mô hình 3D, chẳng hạn kính mắt trong ứng dụng VTO, được tải và thêm vào Scene để phục vụ hiển thị.

Camera đóng vai trò là điểm quan sát trong không gian 3D, xác định góc nhìn và phối cảnh của người dùng. Trong hầu hết các ứng dụng VTO, PerspectiveCamera được ưu tiên sử dụng nhằm mô phỏng cách con người quan sát đối tượng trong không gian thực, tạo ra hiệu ứng phối cảnh tự nhiên. Camera có thể được điều chỉnh để xoay, thu phóng hoặc di chuyển, giúp người dùng quan sát mô hình từ nhiều góc độ khác nhau.

Canvas là phần tử HTML nơi hình ảnh 3D được kết xuất. Đây là khu vực hiển thị đầu ra, nơi mà công cụ kết xuất (Renderer) thực hiện các phép tính đồ họa và trình bày hình ảnh lên màn hình.

Renderer, cụ thể là WebGLRenderer, chịu trách nhiệm chuyển đổi dữ liệu cảnh 3D thành hình ảnh mà người dùng có thể nhìn thấy. WebGLRenderer liên tục cập nhật và hiển thị các thay đổi trong mô hình hoặc góc nhìn camera, đảm bảo khả năng tương tác và hiển thị mượt mà với hiệu suất cao.

Trong Three.js, mỗi mô hình 3D được cấu thành từ các thành phần hình học (geometry) và vật liệu (material). Khi sử dụng GLTFLoader để tải mô hình kính 3D, các yếu tố như hình dạng, chất liệu bề mặt và ánh sáng được xử lý nhằm tạo ra hình ảnh chân thực. Hệ thống ánh sáng đóng vai trò thiết yếu trong việc tạo chiều sâu và hiệu ứng đổ bóng, bao gồm các loại như Directional Light (ánh sáng định hướng) và Ambient Light (ánh sáng môi trường), giúp mô hình được chiếu sáng từ nhiều hướng và cải thiện tính thẩm mỹ thị giác.



Hình 1. Cấu trúc để xây dựng ứng dụng 3D Three.js.

2.3.1. Mô hình kính 3D

Trong ứng dụng thử kính 3D sử dụng kết hợp Three.js và TensorFlow.js, thành phần GLTFLoader đóng vai trò then chốt trong việc tải mô hình kính từ các định dạng phổ biến như GLB hoặc GLTF. Sau khi mô hình kính được tải vào, hệ thống tiến hành căn chỉnh vị trí của kính dựa trên dữ liệu landmark do mô hình FaceMesh cung cấp, bao gồm các điểm đặc trưng như mắt và mũi [4,9]. Để đảm bảo kính hiển thị chính xác và vừa vặn với khuôn mặt người dùng, các phép biến đổi hình học như dịch chuyển (translation), xoay (rotation) và thay đổi tỷ lệ (scaling) được áp dụng lên mô hình kính.

Ngoài việc căn chỉnh về mặt vị trí, yếu tố ánh sáng và hiệu ứng thị giác cũng đóng vai trò quan trọng trong việc tạo ra trải nghiệm chân thực. Three.js hỗ trợ nhiều loại ánh sáng khác nhau, trong đó ánh sáng định hướng (Directional Light) giúp tạo ra bóng đổ rõ nét theo hướng chiếu sáng, còn ánh sáng môi trường (Ambient Light) có tác dụng chiếu sáng toàn cảnh một cách đồng đều mà không tạo bóng, từ đó làm nổi bật mô hình kính một cách tự nhiên và sinh động [1,5].

Toàn bộ hệ thống hoạt động trong một môi trường không gian ba chiều do Three.js cung cấp, bao gồm các thành phần chính như Scene (nơi chứa mô hình kính và nguồn sáng), PerspectiveCamera (mô phỏng góc nhìn tự nhiên của mắt người) và WebGLRenderer (chịu trách nhiệm kết xuất hình ảnh 3D lên trình duyệt với hiệu suất cao). Sự phối hợp chặt chẽ giữa các thành phần này cho phép tạo ra một trải nghiệm thử kính ảo mượt mà, trực quan và có tính tương tác cao đối với người dùng.

2.3.2. Phép tịnh tiến và xoay trong Three.js

Trong Three.js, các phép biến đổi hình học thuộc nhóm phép biến đổi affine như tịnh tiến (translation), xoay (rotation) và thay đổi tỷ lệ (scaling) được sử dụng để điều chỉnh vị trí, hướng và kích thước của mô hình kính 3D sao cho phù hợp với khuôn mặt người dùng. Các phép biến đổi này được áp dụng dựa trên dữ liệu landmark thu được từ mô hình FaceMesh, nhằm đảm bảo mô hình kính được hiển thị một cách chính xác và trực quan trong không gian 3D.

Phép tịnh tiến (Translation): Mô hình kính ban đầu được tải ở một vị trí mặc định trong không gian 3D. Để đưa kính đến đúng vị trí trên khuôn mặt, phép tịnh tiến được áp dụng dựa trên các điểm landmark, đặc biệt là vị trí mắt và mũi. Trung điểm giữa hai mắt thường được sử dụng làm tham chiếu để xác định vị trí cần dịch chuyển mô hình. Trung điểm này được tính theo công thức:

$$MidEye = \left(\frac{x_{leftEye} + x_{rightEye}}{2}, \frac{y_{leftEye} + y_{rightEye}}{2} \right)$$

Sau khi tính được trung điểm, mô hình kính được tịnh tiến sao cho trùng với vị trí này trong không gian hiển thị.

Phép xoay (Rotation): Để mô hình kính khớp với góc nghiêng của khuôn mặt, cần xác định góc nghiêng giữa hai mắt dựa trên tọa độ landmark. Góc quay (θ) được tính theo công thức:

$$\theta = \arctan\left(\frac{y_{rightEye} - y_{leftEye}}{x_{rightEye} - x_{leftEye}}\right)$$

Góc này sau đó được áp dụng để xoay mô hình kính quanh trục Z, giúp mô phỏng góc nhìn của người dùng trong không gian 3D một cách chính xác.

Thay đổi tỷ lệ (Scaling): Để đảm bảo mô hình kính có kích thước phù hợp với khuôn mặt, một phép biến đổi tỷ lệ được áp dụng dựa trên khoảng cách Euclid giữa hai mắt. Khoảng cách này phản ánh độ rộng thực tế của khuôn mặt người dùng và được dùng để điều chỉnh tỉ lệ mô hình kính sao cho không quá lớn hoặc quá nhỏ. Nhờ đó, mô hình kính được hiển thị với tỷ lệ chính xác, góp phần tạo nên trải nghiệm thử kính trực quan và chân thực.

2.3.3. Ánh sáng và đổ bóng (Shading)

Trong Three.js, hệ thống ánh sáng và đổ bóng (shading) đóng vai trò quan trọng trong việc nâng cao tính chân thực và thẩm mỹ của mô hình kính 3D. Bằng cách mô phỏng cách ánh sáng tương tác với bề mặt vật thể, hệ thống này giúp mô hình kính phản ánh chính xác các đặc điểm về chất liệu, độ cong và độ phản xạ, từ đó mang lại trải nghiệm thị giác sống động cho người dùng [5].

Một trong những mô hình chiếu sáng cơ bản được sử dụng là mô hình chiếu sáng Lambert, được áp dụng để tính toán ánh sáng khuếch tán (diffuse lighting). Mô hình này giả định rằng ánh sáng phản xạ đồng đều theo mọi hướng từ bề mặt vật thể và cường độ ánh sáng quan sát được phụ thuộc vào góc giữa ánh sáng chiếu tới và pháp tuyến bề mặt.

Phương trình ánh sáng khuếch tán theo mô hình Lambert được biểu diễn như sau:

$$I_{diffuse} = I_L \cdot \max(0, N \cdot L)$$

Trong đó:

$I_{diffuse}$ là cường độ ánh sáng khuếch tán,

I_L là cường độ ánh sáng từ nguồn sáng,

N là vector pháp tuyến tại điểm trên mô hình kính,

L là vector hướng tới nguồn sáng.

2.4. HelenDataset

Tập dữ liệu hình ảnh khuôn mặt được minh họa trong Hình 2 là một nguồn dữ liệu phổ biến, được sử dụng rộng rãi trong các nghiên cứu về nhận dạng và phân tích khuôn mặt. Bộ dữ liệu này bao gồm 2.330 hình ảnh khuôn mặt với độ phân giải 400×400 pixel, trong đó có 2.000 ảnh được sử dụng cho huấn luyện và 330 ảnh cho kiểm tra [10].



Hình 2. Các hình ảnh khuôn mặt trong HelenDataset.

Mỗi hình ảnh trong tập dữ liệu được gán nhãn chi tiết các thành phần chính của khuôn mặt như mắt, lông mày, mũi, miệng và đường viền hàm, thông qua việc đánh dấu và vẽ các đường viền đặc trưng một cách chính xác và nhất quán. Các nhãn này phục vụ cho việc huấn luyện và đánh giá hiệu quả của các mô hình nhận diện khuôn mặt, đặc biệt là trong các bài toán liên quan đến phát hiện landmark, căn chỉnh khuôn mặt và các ứng dụng thị giác máy tính khác.

3. XÂY DỰNG ỨNG DỤNG

3.1. Quy trình xử lý bài toán thử kính 3D

Trong nghiên cứu này, chúng tôi xây dựng một hệ thống thử kính 3D trực tuyến dựa trên sự kết hợp giữa công nghệ nhận diện khuôn mặt và hiển thị đồ họa 3D thời gian thực. Quy trình xử lý tổng thể của hệ thống được chia thành bốn bước chính như sau:

Bước 1: Nhận diện khuôn mặt và trích xuất điểm landmark với TensorFlow.js (FaceMesh)

Ứng dụng sử dụng camera (webcam) của người dùng để thu nhận hình ảnh khuôn mặt theo thời gian thực. Sau đó, mô hình FaceMesh được triển khai bằng TensorFlow.js sẽ thực hiện nhận diện khuôn mặt và xác định các điểm landmark quan trọng như mắt, mũi và miệng. Các điểm này đóng vai trò then chốt trong việc định vị và căn chỉnh mô hình kính 3D trong các bước tiếp theo.

Bước 2: Tải mô hình kính 3D bằng Three.js

Mô hình kính 3D được tải vào không gian hiển thị thông qua GLTFLoader của thư viện Three.js. Ban đầu, mô hình kính được thiết lập ở kích thước tiêu chuẩn và chưa căn chỉnh vị trí trên khuôn mặt. Các dữ liệu hình học và vật liệu của mô hình sẽ được xử lý để sẵn sàng cho hiển thị.

Bước 3: Căn chỉnh vị trí và kích thước kính 3D

Dựa trên thông tin landmark thu được từ bước 1, hệ thống tiến hành điều chỉnh mô hình kính để khớp với vị trí và tỷ lệ khuôn mặt của người dùng. Các phép biến đổi hình học được áp dụng bao gồm:

Tịnh tiến (translation): đưa kính đến đúng vị trí giữa hai mắt.

Xoay (rotation): điều chỉnh góc nghiêng của kính sao cho phù hợp với hướng khuôn mặt.

Thay đổi tỷ lệ (scaling): điều chỉnh kích thước kính dựa trên khoảng cách giữa hai mắt, đảm bảo kính vừa vặn và tự nhiên khi hiển thị.

Bước 4: Hiển thị và tương tác thời gian thực

Sau khi mô hình kính được căn chỉnh, hệ thống sẽ hiển thị kính trên khuôn mặt người dùng trong môi trường không gian 3D. Thư viện Three.js hỗ trợ các hiệu ứng ánh sáng như Directional Light và Ambient Light, giúp mô phỏng hiệu ứng chiếu sáng và đổ bóng một cách chân thực. Ngoài ra, phương trình Lambert được sử dụng để tính toán ánh sáng khuếch tán (diffuse lighting), tạo ra các hiệu ứng đổ bóng mượt mà trên bề mặt kính, từ đó nâng cao tính chân thực và chất lượng thị giác của mô hình khi người dùng tương tác trực tiếp.

3.2. Các chức năng của ứng dụng thử kính

Ứng dụng thử kính trực tuyến được phát triển với mục tiêu mang lại trải nghiệm tương tác tối ưu cho người dùng. Các chức năng chính của hệ thống bao gồm:

Chức năng thử kính 3D: Ứng dụng sử dụng kết hợp công nghệ nhận diện khuôn mặt và đồ họa 3D để hiển thị mô hình kính trực tiếp lên khuôn mặt người dùng thông qua hình ảnh từ webcam. Mô hình kính 3D được tải lên và hiển thị trong không gian ảo, đồng thời được tự động căn chỉnh về vị trí và tỷ lệ dựa trên các điểm landmark khuôn mặt thu được từ mô hình FaceMesh. Nhờ khả năng xử lý thời gian thực, người dùng có thể quan sát kính từ nhiều góc độ và đánh giá mức độ phù hợp một cách trực quan trước khi đưa ra quyết định lựa chọn sản phẩm.

Chức năng bật/tắt kính: Hệ thống cung cấp tùy chọn cho phép người dùng linh hoạt bật hoặc tắt mô hình kính trong giao diện thử nghiệm. Điều này tạo điều kiện thuận lợi cho việc so sánh nhiều mẫu kính khác nhau, giúp người dùng dễ dàng lựa chọn kiểu dáng phù hợp với khuôn mặt và sở thích cá nhân.

Giao diện người dùng thân thiện: Giao diện ứng dụng được thiết kế tối giản, trực quan và dễ thao tác, nhằm đảm bảo trải nghiệm người dùng liền mạch, kể cả với những người không có kỹ năng công nghệ cao. Ứng dụng hỗ trợ hoạt động ổn định trên nhiều trình duyệt phổ biến như Google Chrome, Mozilla Firefox, Safari và Microsoft Edge, đồng thời không yêu cầu cài đặt thêm phần mềm hay plugin. Điều này góp phần nâng cao tính tiện lợi, khả năng truy cập đa nền tảng và tăng độ phổ cập của ứng dụng đến đông đảo người dùng.

3.3. Thuật toán xác định vị trí và hiển thị mô hình kính 3D

Thuật toán được đề xuất trong nghiên cứu này có nhiệm vụ xác định chính xác vị trí mắt người dùng từ luồng video đầu vào và hiển thị mô hình kính 3D được căn chỉnh theo thời gian thực trên khuôn mặt. Dữ liệu đầu vào bao gồm các khung hình (video frames) từ webcam và các điểm landmark khuôn mặt được trích xuất từ mô hình FaceMesh. Kết quả đầu ra là tọa độ mắt đã được xác định và mô hình kính 3D được render trực tiếp lên khuôn mặt người dùng.

Quy trình xử lý được chia thành các bước chính như sau:

Bước 1: Khởi tạo webcam và mô hình nhận diện khuôn mặt

Bước 1.1: Yêu cầu quyền truy cập webcam từ trình duyệt.

Bước 1.2: Thiết lập luồng video trực tiếp từ webcam và hiển thị trên giao diện web.

Bước 1.3: Chờ cho đến khi video sẵn sàng để đảm bảo dữ liệu ổn định cho các bước xử lý tiếp theo.

Bước 2: Khởi tạo môi trường 3D với Three.js

Bước 2.1: Khởi tạo thư viện Three.js để phục vụ việc render mô hình kính.

Bước 2.2: Tạo các thành phần cơ bản bao gồm scene (khung cảnh), camera (góc nhìn) và renderer (bộ kết xuất đồ họa).

Bước 2.3: Gắn camera vào scene và thiết lập các thông số vị trí phù hợp.

Bước 2.4: Thêm hệ thống ánh sáng (directional light, ambient light) để tăng tính chân thực khi hiển thị kính.

Bước 3: Tải và hiển thị mô hình kính 3D

Bước 3.1: Sử dụng GLTFLoader để tải mô hình kính 3D từ định dạng GLB hoặc GLTF.

Bước 3.2: Thêm mô hình kính vào scene.

Bước 3.3: Áp dụng phép biến đổi để điều chỉnh tỷ lệ (scaling).

Bước 3.4: Thiết lập kích thước ban đầu cho mô hình kính sao cho phù hợp với tỷ lệ khuôn mặt trung bình.

Bước 4: Tải và triển khai mô hình FaceMesh từ TensorFlow.js

Bước 4.1: Tải mô hình nhận diện khuôn mặt FaceMesh từ thư viện TensorFlow.js.

Bước 4.2: Kích hoạt quá trình nhận diện điểm landmark trên khuôn mặt.

Bước 4.3: Với mỗi khung hình từ webcam, mô hình sẽ dự đoán và trả về 468 điểm landmark, bao gồm các vị trí mắt, mũi, miệng và các đặc trưng khác.

Bước 5: Tính toán vị trí và điều chỉnh kính 3D

Bước 5.1: Tính toán trung điểm giữa hai mắt để xác định vị trí đặt kính:

$$\text{MidEye} = \left(\frac{x_{\text{leftEye}} + x_{\text{rightEye}}}{2}, \frac{y_{\text{leftEye}} + y_{\text{rightEye}}}{2} \right)$$

Bước 5.2: Chuyển đổi tọa độ từ hệ trục video sang hệ tọa độ của Three.js.

Bước 5.3: Tính khoảng cách giữa hai mắt để điều chỉnh tỷ lệ kính và tính góc xoay (rotation) dựa trên công thức:

$$\theta = \arctan \left(\frac{y_{\text{rightEye}} - y_{\text{leftEye}}}{x_{\text{rightEye}} - x_{\text{leftEye}}} \right)$$

Bước 5.4: Cập nhật vị trí, tỷ lệ và góc xoay của mô hình kính tương ứng với khuôn mặt người dùng.

Bước 6: Hiển thị kính 3D theo thời gian thực

Bước 6.1: Khởi chạy vòng lặp render để cập nhật liên tục giao diện 3D.

Bước 6.2: Trong mỗi khung hình, cập nhật các điểm landmark mới từ mô hình FaceMesh.

Bước 6.3: Điều chỉnh vị trí, kích thước và góc xoay của kính dựa trên chuyển động khuôn mặt, đảm bảo độ khớp và tính chân thực trong tương tác thời gian thực.

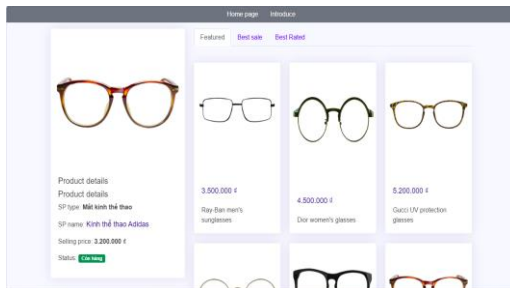
3.4. Ứng dụng

Quy trình người dùng trải nghiệm chức năng thử kính mắt ảo trên nền tảng web được thiết kế để đảm bảo sự tiện lợi, tương tác trực quan và hoạt động hoàn toàn theo thời gian thực. Người dùng bắt đầu bằng cách truy cập vào trang web có tích hợp hệ thống thử kính ảo. Sau đó, người dùng được yêu cầu bật webcam, từ đó hệ thống sẽ tự động thu nhận hình ảnh khuôn mặt và tiến hành nhận diện các đặc điểm hình học quan trọng như vị trí của mắt, mũi và miệng.

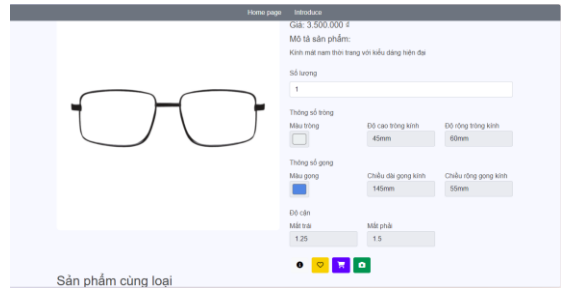
Dựa trên các điểm landmark được xác định thông qua mô hình nhận diện khuôn mặt, hệ thống tiến hành định vị và hiển thị mô hình kính 3D lên khuôn mặt người dùng. Quá trình này bao gồm các bước xử lý hình học như tịnh tiến, xoay và thay đổi tỷ lệ mô hình kính để đảm bảo sự khớp chính xác với khuôn mặt trong từng khung hình video.



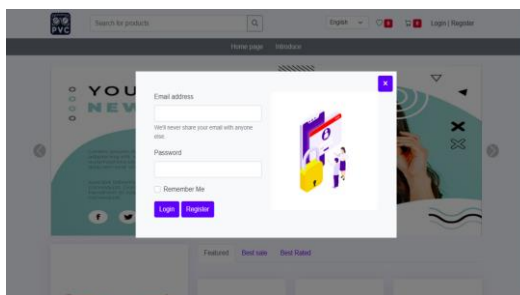
Hình 3. Quy trình người dùng thử kính ảo trên website.



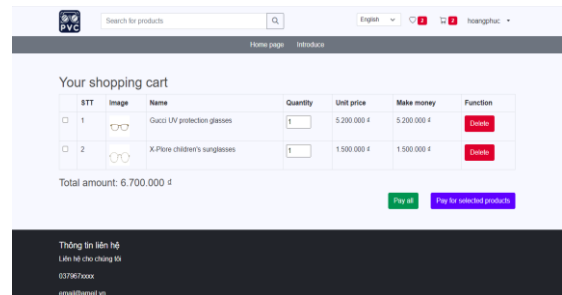
(a)



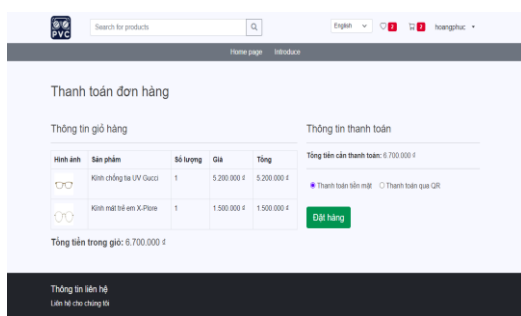
(b)



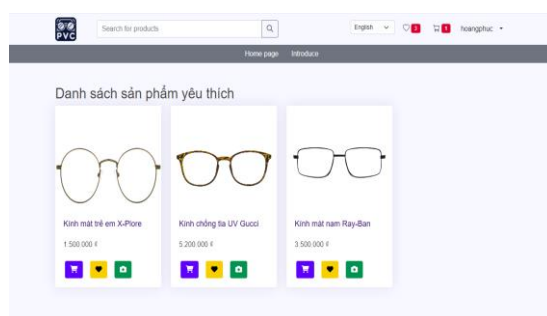
(c)



(d)



(e)



(f)



(g)

Hình 4. Trang chủ (a), trang thông tin sản phẩm (b), trang đăng nhập/đăng ký (c), trang giỏ hàng (d), trang đơn hàng (e), trang sản phẩm được lưu (f), trang thử mắt kính ảo (g).

Toàn bộ hệ thống hoạt động trực tiếp trên trình duyệt, cho phép người dùng sử dụng trên nhiều thiết bị khác nhau như máy tính cá nhân, máy tính bảng hoặc điện thoại di động mà không cần cài đặt thêm phần mềm. Mô hình kính được hiển thị theo thời gian thực, mang lại trải nghiệm tương tác sống động và chân thực. Người dùng có thể dễ dàng lựa chọn, thử nghiệm và so sánh các kiểu dáng, màu sắc và tùy chọn sản phẩm khác nhau trước khi đưa ra quyết định mua hàng, như minh họa trong Hình 3.

Chúng tôi đã xây dựng một ứng dụng thử kính 3D trực tuyến trên nền tảng web, trong đó chức năng thử kính ảo được tích hợp trực tiếp vào giao diện người dùng nhằm mang lại trải nghiệm tương tác mượt mà và thuận tiện. Hệ thống ứng dụng kết hợp mô hình học sâu FaceMesh để nhận diện và trích xuất các điểm landmark trên khuôn mặt, cùng với thư viện đồ họa Three.js để hiển thị mô hình kính 3D theo thời gian thực một cách chính xác và ổn định.

Thông qua ứng dụng này, người dùng có thể dễ dàng tìm kiếm, thử nghiệm và lựa chọn các loại kính mắt khác nhau ngay trong trình duyệt mà không cần đến cửa hàng thực tế. Quá trình thử kính được thực hiện một cách trực quan, giúp người dùng đưa ra quyết định mua hàng chính xác hơn, đồng thời nâng cao chất lượng trải nghiệm mua sắm trực tuyến.

Hệ thống bao gồm nhiều chức năng hỗ trợ, trong đó các chức năng chính và đặc trưng của ứng dụng được minh họa trong Hình 4.

4. KẾT QUẢ VÀ THẢO LUẬN

Trong quá trình nghiên cứu và thực nghiệm hệ thống thử kính ảo, chúng tôi đã tiến hành đánh giá hiệu suất nhận diện khuôn mặt của mô hình thông qua khoảng cách giữa hai điểm landmark đại diện cho mắt trái và mắt phải, được xác định bởi mô hình FaceMesh. Việc đánh giá được thực hiện trên tập dữ liệu HelenDataset với hơn 2.000 ảnh khuôn mặt đã được gán nhãn chính xác [10].

Kết quả cho thấy mô hình FaceMesh có khả năng nhận diện các điểm landmark với độ chính xác tương đối cao, đặc biệt trong các điều kiện ánh sáng ổn định và góc nhìn chính diện. Trung bình, sai số khoảng cách giữa hai mắt được dự đoán so với nhãn thực tế là 23.11 đơn vị, phản ánh hiệu quả của mô hình trong việc tái tạo cấu trúc khuôn mặt phục vụ cho việc căn chỉnh mô hình kính 3D.

Để đánh giá hiệu quả của mô hình một cách toàn diện, chúng tôi áp dụng chiến lược kiểm tra hai giai đoạn đối với cả hai mô hình FaceMesh và Dlib:

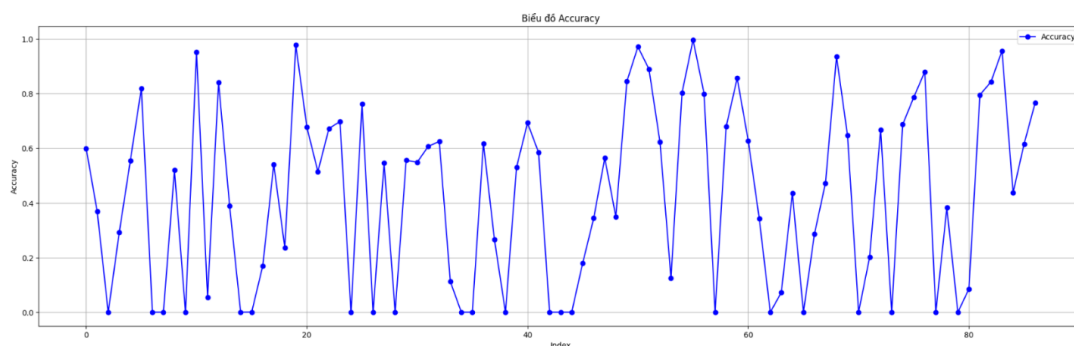
Giai đoạn 1: Tiến hành kiểm tra ban đầu trên 100 ảnh đầu tiên trong tập dữ liệu nhằm đảm bảo các biểu đồ kết quả có thể hiện thị rõ ràng các đường xu hướng và sai số, từ đó hỗ trợ quá trình phân tích trực quan và điều chỉnh thuật toán nếu cần thiết. Kết quả từ giai đoạn này được trình bày trong Hình 5.

Giai đoạn 2: Sau khi xác nhận tính ổn định và xu hướng của mô hình từ giai đoạn đầu, chúng tôi tiếp tục mở rộng đánh giá trên toàn bộ hơn 2.000 ảnh trong HelenDataset nhằm thu thập dữ liệu đầy đủ và xây dựng các biểu đồ phản ánh chính xác hiệu suất tổng thể của hệ thống. Biểu đồ kết quả toàn diện được trình bày trong Hình 6.

Chiến lược hai giai đoạn này vừa giúp tối ưu hóa khả năng phân tích trong giai đoạn đầu, vừa đảm bảo tính đại diện và toàn diện của kết quả đánh giá cuối cùng. Qua đó, chúng tôi nhận thấy rằng mô hình FaceMesh hoạt động ổn định trong môi trường triển khai thực tế, đồng thời có tiềm năng ứng dụng cao trong các hệ thống thử kính ảo trực tuyến.

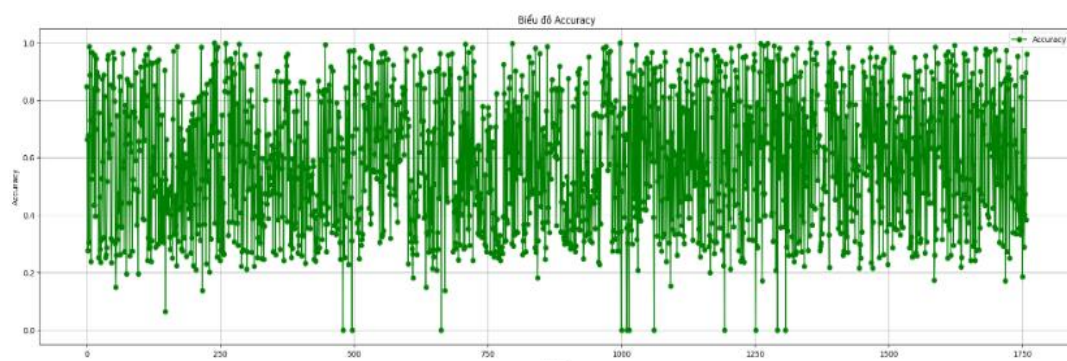


(a)

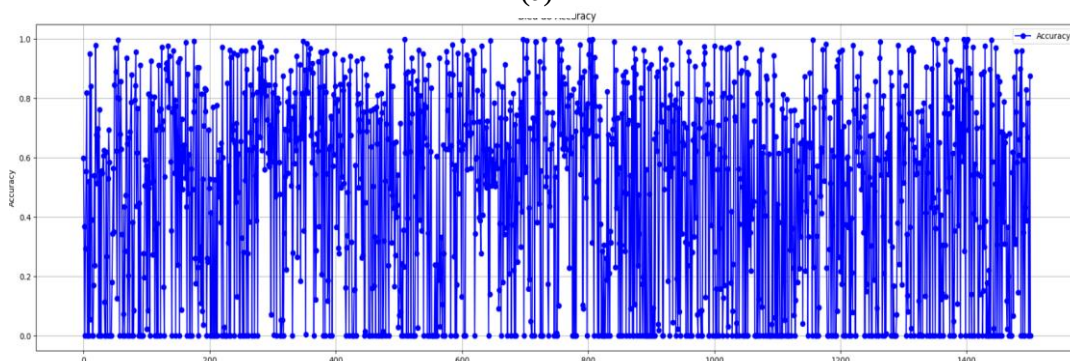


(b)

Hình 5. Biểu đồ biểu diễn độ chính xác của mô hình FaceMesh khi kiểm tra 100 ảnh đầu tiên (a), biểu đồ biểu diễn độ chính xác của mô hình Dlib khi kiểm tra 100 ảnh đầu tiên (b).



(a)



(b)

Hình 6. Biểu đồ biểu diễn độ chính xác của FaceMesh khi kiểm tra trên 2000 ảnh (a), biểu đồ biểu diễn độ chính xác của Dlib khi kiểm tra trên 2000 ảnh (b)

Dựa trên biểu đồ chính xác được trình bày trong Hình 6, cả hai mô hình FaceMesh và Dlib đều cho thấy độ biến thiên nhất định trong quá trình nhận diện các điểm landmark, với độ chính xác cao nhất gần tiệm cận giá trị 1.0. Tuy nhiên, trong một số trường hợp nhất định, đặc biệt khi ảnh đầu vào có chất lượng thấp, bị mờ hoặc góc chụp không thuận lợi, độ chính xác của cả hai mô hình đều sụt giảm nghiêm trọng, thậm chí xuống dưới mức 0, cho thấy việc nhận diện hoàn toàn thất bại.

Dù vậy, khi so sánh trực tiếp, mô hình FaceMesh thể hiện độ ổn định cao hơn, thể hiện qua việc số lượng ảnh có độ chính xác bằng 0 ít hơn đáng kể so với Dlib. Trung bình, độ

chính xác của FaceMesh đạt 59.11%, trong khi Dlib chỉ đạt 47.11%, cho thấy mô hình FaceMesh phù hợp hơn để tích hợp vào các ứng dụng thử kính ảo trên nền tảng web.

Kết quả thực nghiệm cũng chỉ ra rằng FaceMesh có khả năng nhận diện các điểm landmark nhanh và chính xác hơn trong điều kiện ánh sáng tốt và góc nhìn thuận lợi. Đây là một lợi thế lớn trong việc đảm bảo trải nghiệm người dùng liền mạch khi thử kính trực tuyến, giúp mô hình kính 3D được hiển thị đúng vị trí trên khuôn mặt và phản ánh chính xác kiểu dáng cũng như kích thước sản phẩm.

Tuy nhiên, FaceMesh vẫn gặp hạn chế khi xử lý hình ảnh trong điều kiện không lý tưởng, chẳng hạn như ảnh bị mờ, thiếu sáng hoặc chụp từ góc nghiêng. Trong các trường hợp này, sai số trong việc dự đoán vị trí landmark tăng lên, ảnh hưởng đến độ chính xác khi hiển thị kính 3D, đồng thời làm giảm tính chân thực của trải nghiệm thử kính.

Để khắc phục hạn chế trên, cần xem xét tích hợp các kỹ thuật tiền xử lý ảnh như nâng cao độ phân giải, cân bằng độ sáng, hoặc áp dụng bộ lọc khử nhiễu trước khi đưa vào mô hình nhận diện. Ngoài ra, việc kết hợp với các mô hình nhận diện khuôn mặt tiên tiến hơn, hoặc huấn luyện lại mô hình FaceMesh trên tập dữ liệu đa dạng hơn (bao gồm ảnh nghiêng, ảnh thiếu sáng, ảnh ngoài trời...) sẽ giúp cải thiện hiệu suất của hệ thống trong các tình huống thực tế phức tạp.

Cuối cùng, hệ thống có thể được mở rộng triển khai trên các nền tảng di động và máy tính xách tay tích hợp công nghệ thực tế tăng cường (AR), từ đó mang lại trải nghiệm thử kính trực tiếp sống động và linh hoạt hơn, góp phần nâng cao trải nghiệm mua sắm trực tuyến và gia tăng giá trị cho các nền tảng thương mại điện tử.

TÀI LIỆU THAM KHẢO

1. Nguyễn Hồng Quân, Trương Thị Hạnh, Đặng Ngọc Anh, Nghiêm Lê Mỹ Hạnh, Hà Thủy Tinh, Nguyễn Dương Lan Nhi - Ảnh hưởng của công nghệ thực tế ảo tăng cường đến ý định mua hàng trực tuyến của người tiêu dùng, Tạp chí Khoa học Đại học Huế 132 (5C) (2023) 19–38.
2. Nguyễn Thị Quỳnh Như - Ứng dụng của thực tế ảo (VR) và thực tế ảo tăng cường (AR) trong thương mại điện tử, 2023.
3. Marelli D., Bianco S., Ciocca G. - Designing an AI-Based Virtual Try-On Web Application, 2022.
4. Smilkov D., Thorat N., Assogba Y., Yuan A., Kreeger N., Yu P., Zhang K., Cai S., Nielsen E., Soergel D., Bileschi S., Terry M., Nicholson C., Gupta S. N., Sirajuddin S., Sculley D., Monga R., Corrado G., Viegas F. B., Wattenberg M. - TensorFlow.js: Machine Learning for the web and beyond, 2019.
5. Moghaddam M. H., Khazaiepoor M. - Facial recognition of people with masks using MediaPipe Facemesh and deep learning algorithms, Islamic Azad University, Birjand Branch, Iran.
6. Boyko N., Basystiuk O., Shakhovska N. - Performance Evaluation and Comparison of Software for Face Recognition, based on Dlib and OpenCV Library, 2018.
7. Link online: <https://viblo.asia/p/threejs-bai-1-lam-quen-voi-moi-truong-3d-cua-threejs-vyDZO7ROZwj>
8. Link online: <https://lptech.asia/kien-thuc/threejs-la-gi-tong-quan-thu-vien-threejs-cho-nguoi-moi>

9. Wu B., Zhang S., Tian W., Wang H. - Application and Development Prospect of Monitoring Screen based on Three.js Unit Equipment Control System, 2022.
10. Kazemi V., Sullivan J. - One Millisecond Face Alignment with an Ensemble of Regression Trees, Computer Vision and Active Perception Lab, KTH Royal Institute of Technology, Sweden, 2014.

ABSTRACT

REAL-TIME VIRTUAL EYEGLASS TRY-ON INTEGRATION FOR FASHION E-COMMERCE WEBSITES USING FACEMESH AND DLIB

Nguyen Trong Hoang Phuc¹, Ngo Duong Ha¹, Tran Nhu Y^{1*}

Ho Chi Minh City University of Industry and Trade

Email: ytn@huit.edu.vn

The integration of Virtual Try-On (VTO) technology into e-commerce platforms has emerged as a significant trend, particularly in the fashion eyewear industry. This technology enables users to virtually try on glasses through 3D models, thereby enhancing the shopping experience, increasing purchasing confidence, and reducing product return rates. In the context of growing online shopping demand—especially after the Covid-19 pandemic—VTO has become increasingly essential for online retail. However, several technical challenges remain, notably in terms of accuracy and real-time interaction when displaying fashion items on users' faces. This paper proposes and evaluates the effectiveness of two facial recognition models, FaceMesh and Dlib, in supporting online virtual try-on functionality. Both models are integrated into the system using TensorFlow.js and Three.js to render 3D glasses in real-time. Experimental results demonstrate that while both models perform effectively, FaceMesh outperforms Dlib in terms of accuracy and user interaction, particularly in tracking facial movements from various angles. The proposed virtual try-on system contributes to improving user experience and offers promising potential for future e-commerce applications.

Keywords: AR, AI, Virtual Try-on, FaceMesh.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Trọng Hoàng Phúc

Số điện thoại liên hệ: 0889271576

MỘT KHẢO SÁT VỀ CÁC PHƯƠNG PHÁP XÁC ĐỊNH NGUỒN CHO BÀI TOÁN KHAI THÁC TẬP HỮU ÍCH CAO

Nguyen Van Lẽ^{1,*}

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: lenv@huit.edu.vn

Ngày nhận bài 30/03/2025 ; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Khai thác tập hữu ích cao (High Utility Itemset Mining - HUIM) là một bài toán quan trọng trong khai phá dữ liệu, nhằm tìm ra các tập mục có giá trị hữu ích cao từ các cơ sở dữ liệu giao dịch lớn. Một thách thức quan trọng trong bài toán khai thác HUIM là xác định giá trị ngưỡng độ hữu ích tối thiểu ($\minUtil$) phù hợp, vì nó ảnh hưởng trực tiếp đến số lượng và chất lượng các tập hữu ích được khai thác. Bài báo này khảo sát một cách tổng quan về các phương pháp xác định ngưỡng $\minUtil$ trong các nghiên cứu liên quan, bao gồm lựa chọn ngưỡng thủ công, phương pháp elbow và phương pháp dựa trên khai thác top-k. Chúng tôi phân tích ưu điểm, hạn chế và khả năng ứng dụng của từng phương pháp, đồng thời thảo luận về các xu hướng và thách thức trong việc lựa chọn ngưỡng $\minUtil$ một cách tự động và thích ứng, mở ra hướng phát triển cho các kỹ thuật HUIM linh hoạt hơn trong tương lai.

Từ khóa: Tập hữu ích cao, ngưỡng độ hữu ích, cơ sở dữ liệu giao dịch, khai thác dữ liệu, xác định ngưỡng.

1. GIỚI THIỆU

Khai thác tập hữu ích cao (High Utility Itemset Mining - HUIM) là một bài toán quan trọng trong khai thác dữ liệu, giúp tìm ra các tập hợp mục (itemset) có độ hữu ích cao từ cơ sở dữ liệu giao dịch. Không giống như khai thác tập phổ biến (Frequent Itemset Mining - FIM), HUIM không chỉ dựa vào tần suất xuất hiện của các mục mà còn xem xét giá trị lợi nhuận hoặc trọng số của chúng, giúp phản ánh thực tế tốt hơn trong các lĩnh vực như thương mại, y tế và tài chính.

Một trong những thách thức quan trọng trong HUIM là xác định ngưỡng độ hữu ích tối thiểu ($\minUtil$) sao cho hợp lý, vì kết quả khai thác chịu ảnh hưởng trực tiếp bởi giá trị ngưỡng này. Ngưỡng quá cao có thể làm mất đi nhiều tập hữu ích quan trọng, trong khi ngưỡng quá thấp có thể dẫn đến số lượng tập kết quả quá lớn, gây lãng phí tài nguyên và làm phức tạp quá trình phân tích. Do đó, việc xác định một giá trị $\minUtil$ hợp lý là vấn đề cần được nghiên cứu để tối ưu hóa hiệu suất khai thác.

Trong bài báo này, chúng tôi tập trung phân tích ba phương pháp phổ biến để xác định ngưỡng $\minUtil$, bao gồm:

Phương pháp chọn ngưỡng thủ công: Người dùng xác định giá trị minUtil dựa trên kinh nghiệm hoặc thử nghiệm thực nghiệm. Một số thuật toán điển hình sử dụng phương pháp này bao gồm HUI-Miner [1], HUPMiner [2], và EFIM [3], FHM [4] và HMiner [5].

Phương pháp Elbow: Sử dụng phân tích đồ thị để tìm điểm gấp khúc (elbow point), từ đó xác định ngưỡng tối ưu. Phương pháp này ban đầu được áp dụng trong phân cụm [6, 7], và sau đó được mở rộng để xác định ngưỡng minSub và minUtil trong bài toán khai thác tập phổ biến và tập hữu ích cao [8, 9].

Phương pháp Top-K: Thay vì đặt một giá trị minUtil cố định, phương pháp này khai thác đúng k tập hữu ích cao nhất, giúp tự động điều chỉnh ngưỡng [10, 11, 12, 13].

Việc phân tích và so sánh các phương pháp này sẽ giúp tìm ra cách tiếp cận phù hợp hơn trong việc lựa chọn ngưỡng minUtil, nhằm tối ưu hóa hiệu suất khai thác tập hữu ích cao. Phần còn lại của bài báo sẽ trình bày chi tiết các phương pháp trên, phân tích ưu nhược điểm của từng phương pháp, và thảo luận về tiềm năng ứng dụng trong thực tế.

2. CÁC PHƯƠNG PHÁP XÁC ĐỊNH NGƯỠNG ĐỘ HỮU ÍCH CAO

2.1. Phương pháp chọn ngưỡng thủ công

Phương pháp này là cách tiếp cận đơn giản nhất để xác định ngưỡng minUtil trong bài toán khai thác tập hữu ích cao (HUIM). Người dùng tự đặt một giá trị minUtil dựa trên kinh nghiệm hoặc thông qua thử nghiệm. Giá trị này thường được xác định trước khi thuật toán khai thác được thực thi và có thể ảnh hưởng đáng kể đến số lượng tập hữu ích thu được. Tuy nhiên, do mang tính chủ quan, phương pháp này có một số hạn chế như sau:

- Nếu chọn ngưỡng quá cao, thuật toán có thể loại bỏ nhiều tập hữu ích quan trọng, dẫn đến kết quả không đầy đủ.
- Nếu *chọn* ngưỡng quá thấp, số lượng tập kết quả có thể quá lớn, làm tăng chi phí tính toán và gây khó khăn trong việc phân tích.
- Không *phù hợp* với mọi tập dữ liệu, vì giá trị ngưỡng tối ưu có thể thay đổi tùy vào đặc điểm của dữ liệu đầu vào.

Các nghiên cứu áp dụng chọn ngưỡng thủ công thường được chia thành hai nhóm chính:

Nhóm 1: Xác định ngưỡng minUtil dựa trên tỷ lệ phần trăm (%) của tổng độ hữu ích của toàn bộ cơ sở dữ liệu. Trong cách tiếp cận này, minUtil được đặt dựa trên một tỷ lệ phần trăm của tổng độ hữu ích của tất cả các giao dịch trong tập dữ liệu. Điều này giúp đảm bảo rằng ngưỡng được lựa chọn có tính tương đối với quy mô dữ liệu, thay vì là một giá trị tuyệt đối. Một số thuật toán điển hình sử dụng phương pháp này bao gồm: HUI-Miner (Liu và cộng sự, 2012) [1], HUPMiner (Krishnamoorthy, 2015) [2] và EFIM (Zida và cộng sự, 2017) [3].

Bảng 1. So sánh ngưỡng minUtil trong thuật toán HUI-Miner, HUPMiner và EFIM

Cơ sở dữ liệu	Ngưỡng ban đầu (%)			Bước tăng (%)		
	HUI-Miner	HUPMiner	EFIM	HUI-Miner	HUPMiner	EFIM
Chess	18	15	25	2	5	25
Mushroom	2	1	25	0.5	1	25
Accident	15	-	25	5	-	25
Kosarak	1	-	25	0.5	-	25
Chainstore	0.004	0.02	25	0.001	0.02	25
Retail	0.02	0.01	-	0.005	0.02	-

Nhóm 2: Xác định ngưỡng minUtil bằng một giá trị cố định. Thay vì tính toán dựa trên tỷ lệ tổng độ hữu ích, một số nghiên cứu đề xuất sử dụng một giá trị ngưỡng cụ thể, được xác

định theo từng bài toán cụ thể. Cách tiếp cận này giúp đơn giản hóa quá trình khai thác nhưng có thể không linh hoạt với các tập dữ liệu có quy mô khác nhau. Một số thuật toán tiêu biểu sử dụng cách tiếp cận này bao gồm: FHM (Fournier-Viger và cộng sự, 2014) [4], FHN (Lin và cộng sự, 2016) [14], HMiner (Krishnamoorthy, 2017) [5].

Bảng 2. So sánh ngưỡng minUtil trong thuật toán FHM, FHN và HMiner

Cơ sở dữ liệu	Ngưỡng ban đầu (x1000)			Bước tăng (x1000)		
	FHM	FHN	HMiner	FHM	FHN	HMiner
Chess	-	110	80	-	5	20
Mushroom	-	200	1	-	50	1
Accident	-	12000	5000	-	1000	5000
Kosarak	1200	-	1000	200	-	200
Chainstore	1200	-	1	700	-	0.2
Retail	5	1	1.6	5	1	0.2

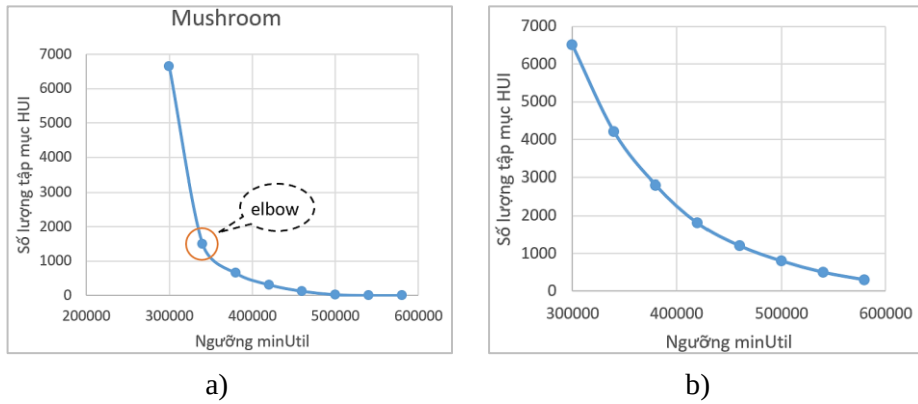
Mặc dù phương pháp chọn ngưỡng thủ công giúp đơn giản hóa quá trình khai thác, nhưng cả hai nhóm tiếp cận trong phương pháp này đều tồn tại những hạn chế đáng kể. Việc xác định minUtil theo tỷ lệ phần trăm của tổng độ hữu ích giúp đảm bảo tính tương quan với quy mô dữ liệu, nhưng không đảm bảo rằng ngưỡng này phù hợp với từng tập dữ liệu cụ thể. Trong khi đó, việc chọn một giá trị cố định lại có tính cứng nhắc, dẫn đến khó khăn khi áp dụng cho các bộ dữ liệu có đặc điểm khác nhau.

Hơn nữa, vì mỗi thuật toán có thể sử dụng các giá trị ngưỡng khác nhau, việc so sánh hiệu suất giữa các thuật toán trên cùng một bộ dữ liệu trở nên phức tạp. Sự khác biệt về giá trị minUtil giữa các nghiên cứu có thể làm sai lệch kết quả đánh giá, gây khó khăn trong việc xác định thuật toán nào thực sự hiệu quả hơn trong thực tế.

Phương pháp chọn ngưỡng thủ công tuy đơn giản nhưng lại thiếu tính tự động hóa và dễ dẫn đến lựa chọn ngưỡng không tối ưu. Vì vậy, các phương pháp khác như phương pháp Elbow và phương pháp Top-K đã được đề xuất nhằm khắc phục hạn chế này.

2.2. Phương pháp Elbow

Đây là một phương pháp phân tích đồ thị, trong đó đường cong thể hiện số lượng tập hữu ích được khai thác theo các giá trị minUtil khác nhau. Điểm gấp khúc (elbow point) trên đồ thị thường được chọn làm ngưỡng tối ưu, vì đây là điểm mà việc giảm giá trị minUtil không còn mang lại sự gia tăng đáng kể về số lượng tập hữu ích. Phương pháp Elbow ban đầu được sử dụng phổ biến trong các bài toán phân cụm, đặc biệt là thuật toán K-Means, để xác định số lượng cụm tối ưu. Trong bài toán phân cụm, đường cong biểu diễn tổng phương sai trong cụm (*within-cluster sum of squares* - WCSS) theo số lượng cụm, và điểm gấp khúc cho thấy số lượng cụm hợp lý trước khi hiệu quả của việc tăng số cụm bắt đầu giảm dần. Năm 2014, Bholowalia và cộng sự sử dụng thuật toán Kmean và phương pháp elbow để tìm ra số cụm một cách tự động [6]. Năm 2019, Saputra và cộng sự trình bày so sánh việc xác định số cụm dữ liệu bằng 2 phương pháp là elbow và silhouette dựa trên 3 độ đo là Manhattan, Euclidian và Minkowski [7]. Nhận thấy tính hiệu quả của phương pháp này, các nhà nghiên cứu đã mở rộng ứng dụng của nó sang các lĩnh vực khác. Năm 2018, Nguyen và cộng sự đề xuất sử dụng phương pháp elbow để xác định ngưỡng độ hỗ trợ trong bài toán khai thác tập phổ biến [8]. Năm 2022, Belgith và cộng sự sử dụng đề xuất sử dụng phương pháp elbow trong việc xác định ngưỡng minUtil trong đề khai thác tập hữu ích cao. Thay vì số cụm, phương pháp Elbow trong HUIM phân tích sự thay đổi của số lượng tập hữu ích theo minUtil để tìm điểm tối ưu, giúp giảm thiểu thử nghiệm thủ công và cải thiện hiệu suất khai thác.



Hình 1. Đồ thị phương pháp elbow trường hợp a) xác định được điểm gãy khúc và trường hợp b) không xác định được điểm gãy khúc

Mặc dù phương pháp elbow cung cấp một cách tiếp cận trực quan để xác định ngưỡng minUtil bằng cách phân tích đồ thị mối quan hệ giữa số lượng tập hữu ích và giá trị minUtil, nhưng phương pháp này vẫn tồn tại một số hạn chế đáng kể.

Thứ nhất, phương pháp này có thể yêu cầu nhiều lần chạy thuật toán khai thác tập hữu ích cao với các giá trị minUtil khác nhau để vẽ được đồ thị, dẫn đến tốn nhiều thời gian tính toán, đặc biệt khi áp dụng trên các tập dữ liệu lớn. Do tính chất của HUIM, mỗi lần thay đổi ngưỡng minUtil có thể ảnh hưởng đáng kể đến số lượng tập kết quả, khiến thời gian chạy của thuật toán biến động không dự đoán trước được. Điều này làm tăng chi phí tính toán, nhất là khi số lượng thử nghiệm quá lớn hoặc dữ liệu có độ phức tạp cao.

Thứ hai, không phải lúc nào cũng xác định được rõ ràng điểm gãy khúc (elbow point) trên đồ thị như trong Hình 1b). Trong một số trường hợp, đường cong có thể không xuất hiện một điểm gãy rõ rệt mà thay vào đó thay đổi dần dần, khiến việc xác định minUtil tối ưu trở nên khó khăn. Điều này đặc biệt xảy ra khi tập dữ liệu không có sự thay đổi đáng kể về số lượng tập hữu ích theo ngưỡng hoặc khi các giá trị minUtil ảnh hưởng đến kết quả một cách tuyến tính hơn là phi tuyến. Khi đó, việc áp dụng phương pháp elbow có thể trở nên kém chính xác, đòi hỏi phải kết hợp thêm các phương pháp hỗ trợ khác để xác định ngưỡng hợp lý hơn.

2.3. Phương pháp Top-k

Phương pháp elbow tuy có thể xác định ngưỡng minUtil một cách trực quan nhưng lại tồn tại những hạn chế nhất định, đặc biệt là chi phí tính toán cao và khó xác định điểm gãy khúc trong một số trường hợp. Để khắc phục những hạn chế này, một cách tiếp cận khác được đề xuất là phương pháp Top-K, trong đó thay vì đặt một ngưỡng minUtil cố định, phương pháp này sẽ tìm ra K tập hữu ích cao nhất mà không cần xác định trước giá trị ngưỡng.

Phương pháp Top-K giúp tự động điều chỉnh ngưỡng minUtil theo dữ liệu đầu vào và yêu cầu của người dùng, giảm thiểu tình trạng phải thử nghiệm nhiều lần với các giá trị minUtil khác nhau. Hơn nữa, phương pháp này đảm bảo rằng số lượng tập hữu ích thu được luôn ở mức mong muốn, tránh việc chọn ngưỡng quá cao dẫn đến mất thông tin hoặc ngưỡng quá thấp khiến kết quả trở nên dư thừa. Một số thuật toán điển hình sử dụng phương pháp Top-K để khai thác tập hữu ích cao bao gồm: TKU & TKO [10], kHMC [11], THUI [12], FTKHUIM [13]. Điểm nổi bật của phương pháp Top-k là sử dụng các chiến lược chọn ngưỡng khởi đầu và chiến lược nâng ngưỡng trong quá trình khai thác. Chiến lược chọn ngưỡng khởi đầu giúp thuật toán xác định được ngưỡng minUtil cao nhất có thể để thuật toán khai thác nhanh hơn. Trong khi đó, chiến lược nâng ngưỡng nhằm đảm bảo ngưỡng minUtil được tăng lên sao cho luôn có K tập mục với độ hữu ích lớn hơn hoặc bằng minUtil.

Bảng 3. Các chiến lược khởi tạo ngưỡng và nâng ngưỡng

STT	Thuật toán	Chiến lược khởi tạo ngưỡng	Chiến lược nâng ngưỡng
1	TKU và TKO	RUC	PE, RUC
2	kHMC	RIU	CUD, COV, RUC
3	THUI	RIU	LIU-E, LIU-LB, RUC
4	FTKHUIM	RIU	RTU, LIU, COD, COV

Phương pháp Top-K mang lại nhiều lợi ích so với các phương pháp khai thác tập hữu ích cao (HUIM) truyền thống, đặc biệt ở khả năng tự động xác định ngưỡng minUtil và tối ưu hóa quá trình khai thác. Tuy nhiên, phương pháp này cũng tồn tại một số hạn chế nhất định. Do thuật toán phải thực hiện nâng ngưỡng và đồng thời kiểm tra số lượng tập mục thỏa điều kiện k tập nên chi phí tính toán sẽ lớn đối với khi giá trị k càng lớn. Bên cạnh đó, các thuật toán này cần các cấu trúc dữ liệu bổ sung để theo dõi và nâng ngưỡng thường xuyên trong quá trình khai thác do thuật toán phải duy trì một danh sách chứa K tập hữu ích cao nhất hiện tại, đồng thời cập nhật giá trị minUtil liên tục. Điều này có thể làm tăng chi phí bộ nhớ và xử lý so với các phương pháp có ngưỡng cố định.

2.4. So sánh và đánh giá các phương pháp trên các tập dữ liệu khảo sát

Qua việc phân tích ba phương pháp xác định ngưỡng minUtil, có thể thấy mỗi cách tiếp cận có những ưu điểm và hạn chế khác nhau tùy theo đặc điểm của tập dữ liệu.

Đối với phương pháp chọn ngưỡng thủ công, kết quả khảo sát trên các tập dữ liệu như Chess, Mushroom, Retail, Chainstore cho thấy cách tiếp cận này khá linh hoạt về mặt triển khai nhưng phụ thuộc nhiều vào kinh nghiệm người dùng. Ví dụ, trên tập Retail và Kosarak, khi áp dụng ngưỡng thấp (0.01% – 0.02%), số lượng tập kết quả rất lớn, khiến thuật toán tiêu tốn nhiều tài nguyên. Trong khi đó, các tập như Accident hoặc Chainstore yêu cầu ngưỡng ban đầu rất nhỏ mới cho kết quả khai thác hiệu quả, điều này thể hiện rõ sự thiếu nhất quán của phương pháp này khi áp dụng cho các tập dữ liệu có quy mô và phân bố độ hữu ích khác nhau.

Phương pháp Elbow có ưu thế trong việc xác định ngưỡng tự động dựa trên phân tích đồ thị. Trên các tập như Chess và Mushroom, đồ thị số lượng tập hữu ích theo minUtil cho thấy điểm gãy khúc tương đối rõ ràng, giúp xác định ngưỡng tối ưu một cách trực quan. Tuy nhiên, với các tập dữ liệu lớn hơn như Kosarak hoặc Chainstore, đường cong có xu hướng thay đổi dần dần mà không xuất hiện điểm gãy rõ ràng, dẫn đến khó xác định ngưỡng một cách chính xác. Ngoài ra, việc phải chạy thuật toán nhiều lần để xây dựng đồ thị khiến phương pháp này trở nên tốn kém về thời gian và tài nguyên.

Phương pháp Top-K khắc phục hạn chế trong việc lựa chọn ngưỡng ban đầu bằng cách đặt trực tiếp số lượng tập cần tìm. Trên các tập dữ liệu như Kosarak và Retail, nơi có độ phân tán độ hữu ích cao, phương pháp này tỏ ra hiệu quả khi giảm được số lần thử nghiệm, đồng thời đảm bảo kết quả phù hợp với yêu cầu của người dùng. Tuy nhiên, với những tập nhỏ như Chess, chi phí tính toán có thể không tương xứng, và cấu trúc dữ liệu bổ sung để theo dõi các tập hiện tại có thể gây thừa tài nguyên so với phương pháp thủ công đơn giản.

Tóm lại, không có phương pháp nào là tối ưu tuyệt đối cho mọi loại dữ liệu. Phương pháp thủ công phù hợp với các tập nhỏ hoặc khi người dùng có kinh nghiệm. Elbow thích hợp với tập dữ liệu trung bình có biến động rõ rệt về số lượng tập theo ngưỡng. Top-K phát huy hiệu

quả trong các tập lớn, phức tạp và yêu cầu tự động hóa cao. Việc lựa chọn phương pháp phù hợp nên dựa trên quy mô dữ liệu, độ phân bố độ hữu ích, và mục tiêu khai thác cụ thể.

2.5. Đề xuất giải pháp kết hợp

Từ kết quả khảo sát và phân tích ở các mục trước, có thể thấy rằng mỗi phương pháp xác định ngưỡng minUtil đều có phạm vi phù hợp nhất định. Do đó, thay vì chỉ lựa chọn một phương pháp cố định, một giải pháp kết hợp thích ứng có thể mang lại hiệu quả tốt hơn trong thực tế. Giải pháp đề xuất gồm ba bước chính:

Bước 1: Phân tích sơ bộ dữ liệu đầu vào. Trước khi khai thác, tiến hành phân tích đặc trưng dữ liệu như số lượng giao dịch, phân bố độ hữu ích, mật độ các tập mục. Thông tin này giúp chọn hướng tiếp cận phù hợp (chẳng hạn: dữ liệu lớn và phân bố rộng thì ưu tiên Top-K; dữ liệu nhỏ với độ biến thiên rõ thì ưu tiên Elbow).

Bước 2, Ước lượng ngưỡng ban đầu bằng phương pháp Elbow hoặc thống kê tần suất/lợi ích. Nếu đường cong có điểm gãy rõ, dùng ngay ngưỡng Elbow làm minUtil. Nếu không, chọn ngưỡng theo tỷ lệ phần trăm tổng lợi ích như phương pháp thủ công (có thể tích hợp với học máy để đề xuất ngưỡng theo lịch sử khai thác).

Bước 3: Áp dụng phương pháp Top-K như một bước tinh chỉnh. Sau khi có ngưỡng khởi đầu, sử dụng khai thác Top-K để tìm tập kết quả tối ưu theo số lượng mong muốn. Cách này vừa tận dụng khả năng kiểm soát của Top-K, vừa giảm thiểu chi phí lặp lại từ đầu.

Giải pháp kết hợp trên cho phép tận dụng ưu điểm của từng phương pháp, vừa đảm bảo tính tự động hóa, hiệu quả, vừa giảm thiểu độ phức tạp và thử nghiệm thủ công.

3. KẾT LUẬN

Xác định ngưỡng minUtil là một vấn đề quan trọng trong khai thác tập hữu ích cao (HUIM), ảnh hưởng trực tiếp đến chất lượng kết quả khai thác cũng như hiệu suất thuật toán. Trong bài báo này, chúng tôi đã phân tích ba phương pháp phổ biến để xác định ngưỡng minUtil: chọn ngưỡng thủ công, phương pháp Elbow, và phương pháp Top-K. Phương pháp chọn ngưỡng thủ công đơn giản nhưng phụ thuộc nhiều vào kinh nghiệm người dùng và thiếu tính tự động hóa. Phương pháp Elbow cung cấp một cách tiếp cận trực quan dựa trên phân tích đồ thị, nhưng có thể gặp khó khăn trong việc xác định điểm gãy khúc rõ ràng, đặc biệt đối với những tập dữ liệu phức tạp. Trong khi đó, phương pháp Top-K mang lại sự linh hoạt hơn bằng cách xác định trực tiếp số lượng tập hữu ích mong muốn, tuy nhiên, nó đòi hỏi các chiến lược nâng ngưỡng và quản lý bộ nhớ hiệu quả để đảm bảo hiệu suất.

Mỗi phương pháp đều có ưu và nhược điểm riêng, tùy thuộc vào yêu cầu bài toán và đặc điểm dữ liệu. Do đó, việc lựa chọn phương pháp phù hợp cần xem xét nhiều yếu tố như tính tự động hóa, độ chính xác, hiệu suất tính toán và khả năng mở rộng. Trong tương lai, các nghiên cứu có thể tập trung vào việc cải tiến hoặc kết hợp các phương pháp này để tăng cường độ chính xác và tối ưu hóa hiệu năng trong khai thác HUIM.

TÀI LIỆU THAM KHẢO

1. Liu M., Qu J. - Mining high utility itemsets without candidate generation," In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (2012) 55-64.

2. Krishnamoorthy S. - Pruning strategies for mining high utility itemsets, *Expert Systems with Applications* 42 (5) (2018) 2371-2381.
3. Zida S., Fournier-Viger P., Lin J. C. W., Wu C. W. and Tseng V. S. - EFIM: a fast and memory efficient algorithm for high-utility itemset mining, *Knowledge and Information Systems* 51 (2) (2017) 595-625.
4. Fournier-Viger P., C. Wu W., Zida S. and Tseng V. S. - FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning, *International Symposium on Methodologies for Intelligent Systems*, 8502 (2014) 83-92.
5. Krishnamoorthy S. - HMiner: Efficiently mining high utility itemsets, *Expert Systems with Applications*, 90 (2017) 168-183.
6. Bholowalia P. and Kumar A. - EBK-means: A clustering technique based on elbow method and k-means in WSN, *International Journal of Computer Applications* 105 (9) (2014).
7. Saputra D. M., Saputra D. and Oswari L. D. - Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method, In *Sriwijaya international conference on information technology and its applications* (2020) 341-346.
8. Nguyen D., Nguyen T. D. W., Luo and S. Venkatesh - Trans2vec: learning transaction embedding via items and frequent itemsets, In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference 22* (2018), 361-372.
9. Belghith K., Fournier-Viger P. and Jawadi J. - Hui2Vec: learning transaction embedding through high utility itemsets, In *International Conference on Big Data Analytics* (2022) 211-224.
10. Tseng V. S., Wu C. W., Fournier-Viger P. and Philip S. Y. - Efficient algorithms for mining top-k high utility itemsets, *IEEE Transactions on Knowledge and Data Engineering* 28 (1) (2015) 54-67.
11. Duong Q.-H., Liao B., Fournier-Viger P. and Dam T.-L. - An efficient algorithm for mining the top-k high utility itemsets, using novel threshold raising and pruning strategies, *Knowledge-Based Systems* 104 (2016) 106–122.
12. Krishnamoorthy S. - Mining top-k high utility itemsets with effective threshold raising strategies, *Expert Systems with Applications* 117 (2019) 148-165.
13. Vu V. V., Lam M. T., Duong T. T., Manh L. T., Nguyen T. T., Nguyen L. V. and Vo B. - Ftkhuim: a fast and efficient method for mining top-k high-utility itemsets, *IEEE Access* 11 (2023) 104789-104805.
14. Lin J. C. W., Fournier-Viger P. and Gan W. - FHN: An efficient algorithm for mining high-utility itemsets with negative unit profits, *Knowledge-Based Systems*, 111 (2016) 283-298.

ABSTRACT

**A SURVEY ON THRESHOLD DETERMINATION METHODS FOR HIGH UTILITY
ITEMSET MINING**

Nguyen Van Le

Ho Chi Minh City University of Industry and Trade

Email: lenv@huit.edu.vn

High Utility Itemset Mining (HUIM) is a critical problem in data mining that aims to discover itemsets with high utility values from large transaction databases. One of the key challenges in HUIM is determining an appropriate minimum utility threshold (minUtil), as it directly affects the number and quality of the extracted high-utility itemsets. This paper provides a comprehensive review of various approaches for setting the minUtil threshold in HUIM, including manual threshold selection, the elbow method, and top-k mining-based approaches. We analyze the advantages, limitations, and applicability of each method while discussing trends and challenges in automating and adapting minUtil selection. Our study highlights potential research directions for developing more flexible and efficient HUIM techniques in the future.

Keywords: High Utility Itemset, Utility Threshold, Transactional Database, Data Mining, Threshold Determination

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Nguyễn Văn Lễ

Số điện thoại liên hệ: 0378890133

XÁC THỰC VÀ BẢO MẬT HÌNH ẢNH BẰNG CẤP DÙNG BLOCKCHAIN HYPERLEDGER FABRIC

**Lê Quang Vinh¹, Nguyễn Thị Thanh Thảo¹, Huỳnh Minh An¹,
Lâm Nguyễn Anh Hào¹, Nguyễn Ngọc Nghĩa¹, Đặng Ngọc Tài¹,
Hà Huỳnh Ánh Ngân¹, Huỳnh Công Tiến¹, Phan Công Hiệu¹,
Nguyễn Thị Bích Ngân^{1,*}**

¹Khoa Công nghệ Thông tin, Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: 1001240021@huit.edu.vn, 2033216553@hufi.edu.vn, 2001220056@huit.edu.vn,
2001221180@huit.edu.vn, 2001223033@huit.edu.vn, 2001224231@huit.edu.vn,
2001222912@huit.edu.vn, 2001224389@huit.edu.vn, 2001215781@hufi.edu.vn,
nganntb@huit.edu.vn

Ngày nhận bài 31/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Hiện nay, tình trạng làm giả bằng cấp ngày càng nghiêm trọng, ảnh hưởng tiêu cực đến uy tín của các tổ chức giáo dục và doanh nghiệp. Điều này không chỉ gây tổn thất tài chính khi doanh nghiệp tuyển dụng sai người mà còn làm giảm niềm tin vào hệ thống giáo dục. Trước đây, việc xác thực văn bằng chủ yếu dựa vào tem chống giả hoặc kiểm tra qua cơ sở dữ liệu truyền thống, nhưng các phương pháp này dễ bị làm giả, khó truy xuất nguồn gốc và thiếu tính minh bạch. Để khắc phục những vấn đề, trong bài báo này, chúng tôi đưa ra giải pháp ứng dụng Blockchain Hyperledger Fabric kết hợp với InterPlanetary File System (IPFS) nhằm đảm bảo tính bảo mật và xác thực văn bằng theo cách phi tập trung. Do Blockchain có chi phí cao và dung lượng lưu trữ hạn chế, IPFS đóng vai trò trung gian, lưu trữ hình ảnh bằng cấp và tạo mã định danh duy nhất (CID – Content Identifier), sau đó CID này được ghi vào Blockchain để đảm bảo tính toàn vẹn và khả năng truy xuất dữ liệu. Bên cạnh đó, để đảm bảo tính xác thực của văn bằng và danh tính các bên liên quan, hệ thống sử dụng Fabric Certificate Authority (Fabric CA) để cấp chứng chỉ số theo chuẩn X.509, giúp xác thực tổ chức cấp bằng và người nhận. Qua quá trình nghiên cứu và thực thi, chúng tôi đã triển khai thành công ứng dụng di động Fabric Certificate Authority (Fabric CA) để đăng ký và quản lý danh tính của các bên liên quan, bao gồm tổ chức cấp bằng, người xét duyệt và người nhận bằng cấp. Bên cạnh đó, ứng dụng cũng đã tích hợp InterPlanetary File System (IPFS) để hỗ trợ việc chuyển đổi hình ảnh bằng cấp thành mã định danh duy nhất (CID – Content Identifier) trước khi ghi nhận lên blockchain Hyperledger Fabric.

Từ khóa: Blockchain Hyperledger Fabric, IPFS, Fabric CA, Smart Contract, chứng chỉ X.509.

1. MỞ ĐẦU

1.1. Đặt vấn đề

Hoạt động làm giả bằng cấp, chứng chỉ,... đã có từ nhiều năm qua, và khi các công nghệ in ấn, chỉnh sửa được nâng cao thì việc giả mạo ngày càng tinh vi. Hoạt động này diễn ra ở nhiều nơi, nhiều nước và ở Việt Nam thì cũng không ngoại lệ. Theo The New York Times, công ty Axact đã vận hành ít nhất 370 trang web cung cấp bằng cấp giả trên toàn cầu, thu về hàng triệu đô la Mỹ [1]. Tại Việt Nam, thống kê của Bộ Giáo dục và Đào tạo cho thấy từ tháng 01/2017 đến tháng 11/2023, Cục Quản lý chất lượng đã tiếp nhận 37.436 hồ sơ công nhận văn bằng nước ngoài, trong đó 1.774 hồ sơ (4,74%) bị từ chối do không đáp ứng điều kiện hoặc có dấu hiệu giả mạo [2]. Điều này gây ra nhiều hậu quả nghiêm trọng đối với uy tín của các cơ sở giáo dục, làm ảnh hưởng chất lượng nhân sự, đặc biệt là các vị trí đòi hỏi nhân sự có chuyên môn cao.

Những năm trước, việc xác minh văn bằng chủ yếu phụ thuộc vào tem chống giả hoặc tra cứu qua các cơ sở dữ liệu truyền thống; tuy nhiên, những cách này dễ bị sao chép, khó xác định nguồn gốc, và không đảm bảo tính rõ ràng, tin cậy. Để giải quyết các hạn chế đó, chúng tôi nghiên cứu và đề xuất một phương pháp sử dụng blockchain Hyperledger Fabric kết hợp với Hệ thống Tập Liên hành tinh (IPFS) nhằm tăng cường bảo mật và xác thực văn bằng theo mô hình phi tập trung.

Blockchain Hyperledger Fabric là một nền tảng blockchain được thiết kế dành cho các ứng dụng doanh nghiệp, thuộc loại permissioned blockchain (chuỗi khối có cấp quyền), nghĩa là chỉ những thực thể được cấp phép mới có thể tham gia vào hệ thống, giúp đảm bảo tính bảo mật, quyền riêng tư và khả năng kiểm soát dữ liệu. Không giống như Bitcoin hay Ethereum là các blockchain công khai (public blockchain), nơi mọi giao dịch đều minh bạch và bất kỳ ai cũng có thể tham gia, Hyperledger Fabric cung cấp một môi trường có kiểm soát, phù hợp với các hệ thống yêu cầu bảo mật cao như xác thực bằng cấp. Một điểm khác biệt quan trọng là Hyperledger Fabric hỗ trợ nhiều cơ chế đồng thuận như Kafka, Raft, IBFT, giúp tối ưu hóa hiệu suất cho các ứng dụng doanh nghiệp, trong khi các blockchain công khai chủ yếu sử dụng Proof of Work (PoW) hoặc Proof of Stake (PoS), vốn tiêu tốn tài nguyên và không phù hợp với hệ thống có yêu cầu hiệu suất cao. Ngoài ra, Hyperledger Fabric tích hợp hệ thống quản lý danh tính Fabric CA, hỗ trợ xác thực và cấp quyền chi tiết cho từng người tham gia, đảm bảo tính minh bạch và chống giả mạo trong việc quản lý văn bằng, điều mà các blockchain công khai không có do mọi giao dịch đều ẩn danh. Một ưu điểm nữa là Fabric sử dụng Chaincode, một dạng Smart Contract có thể viết bằng nhiều ngôn ngữ lập trình phổ biến như Go, Java, Node.js, giúp dễ dàng triển khai vào hệ thống doanh nghiệp, trong khi Ethereum chỉ hỗ trợ Solidity, yêu cầu môi trường đặc biệt để chạy và phát sinh phí giao dịch (gas fee) cao. Nhờ kiến trúc tối ưu, Hyperledger Fabric có thể xử lý hàng nghìn giao dịch mỗi giây, trong khi Bitcoin và Ethereum bị hạn chế tốc độ do phụ thuộc vào cơ chế xác thực phi tập trung [4].

Nghiên cứu của chúng tôi đã xây dựng một ứng dụng di động thực hiện xác thực hai cấp của blockchain, bao gồm Người quản lý và Người duyệt, nhằm đảm bảo mỗi văn bằng trải qua quy trình kiểm duyệt nghiêm ngặt trước khi được ghi nhận vào blockchain. Người quản lý có trách nhiệm kiểm tra tính hợp lệ của thông tin đầu vào, xác minh danh tính tổ chức cấp bằng và đảm bảo rằng văn bằng đáp ứng các tiêu chuẩn quy định. Sau đó, Người duyệt sẽ thực hiện kiểm tra chéo, đối chiếu thông tin và xác nhận lần cuối trước khi gửi dữ liệu vào blockchain thông qua Smart Contract. Với mô hình này, ứng dụng không chỉ đảm bảo tính minh bạch và giảm thiểu rủi ro gian lận mà còn giúp ngăn chặn hành vi cấp bằng trái phép và tối ưu hóa chi phí quản lý so với các phương thức xác thực truyền thống.

1.2. Tình hình nghiên cứu liên quan

Qua quá trình khảo sát các nghiên cứu liên quan vấn đề này ở trong và ngoài nước, chúng tôi ghi nhận được như sau:

- Năm 2024, Sutradhar và cộng sự [3] nghiên cứu và kết hợp blockchain Hyperledger Fabric và OAuth 2.0 để tăng cường quyền riêng tư, bảo mật và khả năng mở rộng trong lĩnh vực chăm sóc sức khỏe. Nhóm tác giả xây dựng hệ thống quản lý danh tính và truy cập dạng sổ cái phi tập trung và không thể thay đổi để đảm bảo tính bảo mật và toàn vẹn của dữ liệu bệnh nhân đồng thời tạo điều kiện cho khả năng kiểm toán minh bạch và kiểm soát truy cập chi tiết. Hệ thống đề xuất giải quyết các thách thức về khả năng tương tác dữ liệu, tuân thủ quy định và phòng ngừa gian lận, hứa hẹn sẽ chuyển đổi các hoạt động quản lý dữ liệu chăm sóc sức khỏe.
- Năm 2024, Jena, S. K và các cộng sự [4] đề xuất giải pháp tiên tiến dựa trên blockchain sử dụng Hyperledger Fabric để theo dõi các khiếu nại gian lận trong ngành bảo hiểm chăm sóc sức khỏe. Việc tích hợp công nghệ blockchain và các kỹ thuật Machine Learning, chẳng hạn như Support Vector Machines và Random Forest Regression, nhằm mục đích tăng cường tính bảo mật và hiệu quả trong việc xử lý các khiếu nại bảo hiểm, đồng thời đảm bảo tính bảo mật dữ liệu bệnh nhân và thiết lập một hệ sinh thái lấy bệnh nhân làm trung tâm. Bằng cách tự động hóa các khuyến nghị về hợp đồng bảo hiểm và sử dụng dữ liệu sức khỏe theo thời gian thực từ các thiết bị đeo, khuôn khổ này không chỉ làm giảm các trường hợp gian lận mà còn thúc đẩy tính minh bạch và hiệu quả trong ngành bảo hiểm.
- Năm 2024, Muhammad Saad và cộng sự [4] trình bày nghiên cứu về việc tích hợp hệ thống trên blockchain sử dụng Hyperledger Fabric với camera an ninh truyền hình mạch kín (CCTV) được lắp cố định tại một bất động sản cho thuê. Nhóm tác giả nhấn mạnh tầm quan trọng của việc giải quyết các thách thức về bảo mật do việc sử dụng ngày càng tăng các thiết bị IoT trong khi cân bằng các phức tạp liên quan đến khả năng mở rộng và triển khai. Các cải tiến chính bao gồm thuật toán mã chuỗi độc đáo, xác thực mã thông báo web để tương tác liền mạch trong các tình huống cho thuê và tăng cường tính ẩn danh thông qua các địa chỉ ẩn và chữ ký vòng đã sửa đổi, cải thiện đáng kể tính bảo mật và độ tin cậy của hệ thống được đề xuất.
- Năm 2024, Jummai Enare Abang và cộng sự [4] đánh giá toàn diện về mô hình hiệu suất độ trễ trong chuỗi khối Hyperledger Fabric (HLF), đặc biệt là trong bối cảnh các ứng dụng Internet vạn vật (IoT). Nghiên cứu nêu bật tầm quan trọng của việc giải quyết các thách thức về độ trễ khi tích hợp HLF vào hệ sinh thái IoT, đề xuất các tiêu chí đánh giá mới và khám phá các nỗ lực nghiên cứu và mô hình hóa hiện có. Bài báo xác định các thông số chính ảnh hưởng đến độ trễ, phân loại bối cảnh nghiên cứu hiện tại và thảo luận các hướng nghiên cứu trong tương lai để tối ưu hóa độ trễ của HLF trong các thiết lập IoT.

Trong nước: một số nghiên cứu liên quan Blockchain Hyperledger Fabric

- Năm 2024, Nguyễn Văn Hòa và Trần Minh Khoa [5] nghiên cứu và xây dựng mạng Blockchain dựa trên framework Hyperledger Fabric và thiết kế Chaincode (hợp đồng thông minh) đáp ứng các nghiệp vụ hộ tịch cần thiết.
- Năm 2019, Trương Minh Tuyền và cộng sự [6] đề xuất giải pháp quản lý tài sản ngăn chặn bằng công nghệ blockchain thông qua nền tảng Hyperledger Fabric. Giải pháp đề xuất góp phần nâng cao hiệu quả công tác quản lý, điều hành và đơn giản thủ tục công chứng tại các phòng công chứng và cơ quan quản lý nhà nước. Kết quả thử nghiệm giải pháp đề xuất có hiệu suất cao và có khả năng ứng dụng rộng rãi trong lĩnh vực công chứng.

Qua quá trình khảo sát và đánh giá các nghiên cứu liên quan cho thấy công nghệ blockchain đang là hướng nghiên cứu hấp dẫn vì những giá trị mang lại trong lĩnh vực bảo mật thông tin dựa trên nền tảng dữ liệu phi tập trung và cấu trúc lưu trữ theo dạng chuỗi khối. Đây chính là cơ sở để chống giả mạo thông tin và nâng cao hiệu suất xác thực dựa trên kiến trúc phi tập trung và chia sẻ thông tin sổ cái. Công nghệ blockchain có thể tạo ra một hệ sinh thái an toàn, linh hoạt và thích hợp trong nhiều lĩnh vực.

Bài báo này đề xuất sử dụng Fabric Certificate Authority (Fabric CA) để đăng ký và xác thực danh tính của người nhận bằng cấp, đảm bảo rằng chỉ những cá nhân và tổ chức hợp lệ mới có thể tham gia vào quy trình xác thực. Song song đó, nhằm lưu trữ văn bằng an toàn và hiệu quả, InterPlanetary File System (IPFS) được áp dụng để tạo mã định danh duy nhất cho hình ảnh trước khi liên kết với blockchain. Cách tiếp cận này không chỉ giúp tối ưu hóa dung lượng mà còn đảm bảo tính toàn vẹn của dữ liệu. Bên cạnh đó, hệ thống được xây dựng trên nền tảng Hyperledger Fabric, tích hợp hợp đồng thông minh (Smart Contract) để tự động hóa quá trình kiểm tra và xác thực, qua đó nâng cao tính minh bạch và giảm thiểu rủi ro gian lận.

Cấu trúc còn lại của báo cáo gồm : phần 2 của bài báo giới thiệu chi tiết về các công nghệ được sử dụng và phân tích tác động của chúng trong quá trình xác thực văn bằng; phần 3 trình bày các bước thực hiện, mô tả cách dữ liệu được xử lý và ghi nhận lên blockchain; phần 4 tổng kết nội dung nghiên cứu và đề xuất hướng phát triển trong tương lai. Tài liệu tham khảo được liệt kê ở cuối bài báo.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Hyperledger Fabric Blockchain

Hyperledger Fabric là một nền tảng blockchain mã nguồn mở do Linux Foundation phát triển [7], tập trung vào nhu cầu của doanh nghiệp, đặc biệt là các tổ chức cần một hệ thống giao dịch bảo mật, minh bạch nhưng vẫn giữ quyền kiểm soát dữ liệu. Không giống như blockchain công khai, Hyperledger Fabric cho phép xây dựng mạng riêng tư (permissioned blockchain), nơi chỉ những bên được ủy quyền mới có thể tham gia, đảm bảo quyền riêng tư và tính bảo mật cao. Một trong những điểm nổi bật của nền tảng này là khả năng triển khai hợp đồng thông minh (chaincode) để tự động hóa các quy trình kinh doanh mà không cần bên thứ ba trung gian. Ngoài ra, nó hỗ trợ nhiều cơ chế đồng thuận khác nhau, giúp doanh nghiệp có thể linh hoạt lựa chọn phương thức phù hợp với mô hình hoạt động của mình, từ đó tối ưu hiệu suất và giảm chi phí giao dịch. Nhờ khả năng mở rộng tốt, dễ dàng tích hợp với hệ thống hiện có và khả năng xử lý giao dịch nhanh chóng, Hyperledger Fabric trở thành lựa chọn lý tưởng cho nhiều lĩnh vực như tài chính, chuỗi cung ứng, chăm sóc sức khỏe và quản lý danh tính [8].

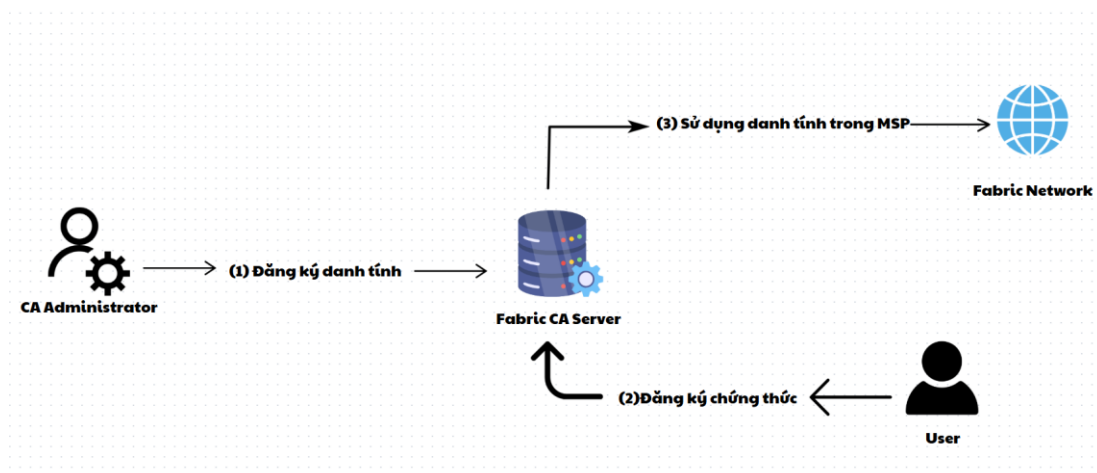
Kiến trúc Hyperledger Fabric được xây dựng dựa trên ba yếu tố cốt lõi: **Client**, **Peer Organization** và **Orderer Organization**. Mỗi phần đóng một vai trò riêng biệt, phối hợp nhịp nhàng để duy trì hoạt động của mạng lưới.

2.2. Cơ chế đăng ký và quản lý danh tính bằng Fabric CA

Fabric CA là một Certificate Authority (CA) riêng tư được tích hợp trong Hyperledger Fabric [9], có nhiệm vụ phát hành và quản lý các chứng chỉ X.509 cho các danh tính trong mạng. Fabric CA hỗ trợ việc tạo ra các danh tính có quyền hạn cụ thể (như admin, peer, orderer, client) và đảm bảo rằng chỉ các danh tính được ủy quyền mới có thể tương tác với mạng. X.509 là một tiêu chuẩn cho cơ sở hạ tầng khóa công khai (Public Key Infrastructure - PKI), quy định định dạng của các chứng chỉ công khai [9]. Trong Hyperledger Fabric, mỗi danh tính được biểu diễn bằng một chứng chỉ X.509, bao gồm thông tin như chủ sở hữu (subject), cơ

quan phát hành (issuer), khóa công khai, và thời hạn hiệu lực. Chứng chỉ này được sử dụng để xác thực danh tính khi tương tác với mạng, chẳng hạn như khi gửi giao dịch hoặc truy vấn dữ liệu.

Quy trình đăng ký và quản lý danh tính trong Fabric CA bao gồm hai bước chính: **đăng ký danh tính (registration)** và **đăng ký tham gia (enrollment)**. Dưới đây là giải thích chi tiết về quy trình này.



Hình 1. Minh họa quy trình đăng ký và quản lý danh tính trong Fabric CA

Đăng Ký Danh Tính (Registration)

Bước 1: Sử dụng Fabric CA để đăng ký danh tính cho Admin

Hệ thống (có thể là quản trị viên CA hoặc một danh tính có quyền registrar) sử dụng Fabric CA để đăng ký danh tính cho Admin của một trường học cụ thể, ví dụ: "admin_universityA". Khi được đăng ký, Admin sẽ nhận được Enroll ID và Secret, đồng thời có quyền thực hiện các tác vụ quản lý danh tính trong phạm vi trường của mình.

Bước 2: Admin đăng ký danh tính cho sinh viên

Sau khi được cấp danh tính, Admin sử dụng Fabric CA để đăng ký danh tính cho sinh viên thuộc trường đó, ví dụ: "student1_universityA". Trong quá trình này, admin trường học cấp cho mỗi sinh viên một Enroll ID và Secret riêng biệt. Đây là bước chuẩn bị để sinh viên có thể tham gia vào hệ thống.

Đăng Ký Tham Gia (Enrollment)

Bước 3: Sinh viên đăng ký tham gia để nhận chứng chỉ và đăng nhập ứng dụng

Sinh viên sử dụng Enroll ID và Secret đã được Admin cấp để thực hiện quá trình đăng ký tham gia (enrollment) thông qua Fabric CA. Kết quả của quá trình này là sinh viên nhận được chứng chỉ X.509 và khóa riêng tư. Với chứng chỉ này, sinh viên có thể chính thức đăng nhập vào ứng dụng xuất trình bằng cấp và thực hiện các hành động như xem, quản lý hoặc yêu cầu xác thực bằng cấp của mình.

Quản Lý Danh Tính

Bước 4: Admin quản lý danh tính của sinh viên

Admin có quyền quản lý danh tính của sinh viên, bao gồm việc thu hồi chứng chỉ X.509 nếu sinh viên không còn thuộc trường (ví dụ: đã tốt nghiệp, chuyển trường hoặc không còn hợp lệ). Việc thu hồi này đảm bảo rằng sinh viên không còn quyền truy cập hoặc thực hiện bất kỳ hành động nào trong ứng dụng xuất trình bằng cấp sau khi danh tính của họ không còn hiệu lực.

Các danh tính được lưu trữ trong **Membership Service Provider (MSP)** [10] được sử dụng để xác thực và ủy quyền trong mạng. Khi danh sách CRL được cập nhật, tất cả các nút trong mạng (**Peer, Orderer, Client, Admin, v.v.**) sẽ tham chiếu đến danh sách này để kiểm tra trạng thái chứng chỉ. Nếu một danh tính có chứng chỉ nằm trong danh sách CRL, hệ thống sẽ từ chối giao dịch hoặc yêu cầu từ danh tính đó.

2.3 Vai trò của IPFS và AES trong hệ thống xuất trình bằng dựa trên Hyperledger Fabric

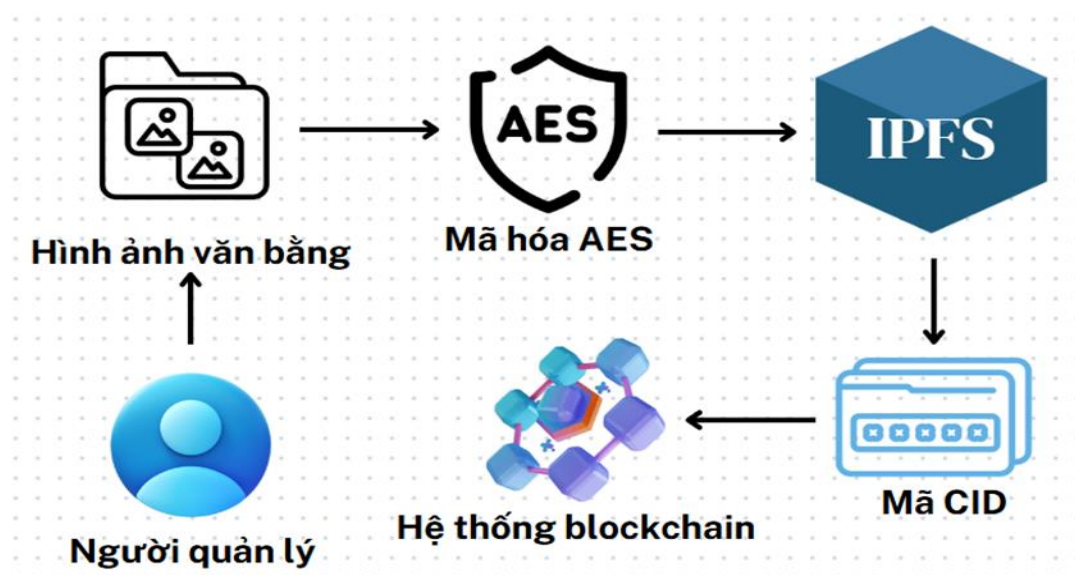
Trong hệ thống này, Hyperledger Fabric được sử dụng để quản lý danh tính, quyền truy cập và ghi nhận giao dịch trên sổ cái phân tán (DLT). Trong khi đó, IPFS hỗ trợ việc lưu trữ tài liệu bằng cấp như hình ảnh và chứng nhận mà không làm tăng dung lượng của blockchain. Khi một tài liệu được tải lên IPFS, nó sẽ được mã hóa bằng AES để đảm bảo tính bảo mật. Sau khi mã hóa, tài liệu được tải lên IPFS và nhận mã băm duy nhất (CID). CID này sau đó sẽ được ghi vào Hyperledger Fabric, giúp dễ dàng truy xuất khi cần thiết mà không ảnh hưởng đến bảo mật.

Thuật toán AES được áp dụng trước khi lưu trữ dữ liệu trên IPFS nhằm bảo vệ tính riêng tư và bảo mật. Chỉ những người sử dụng có khóa giải mã hợp lệ mới có thể truy cập nội dung bằng cấp, ngay cả khi CID bị lộ. Cách kết hợp này đảm bảo rằng dữ liệu vẫn có thể được truy xuất bất cứ lúc nào mà không ảnh hưởng đến bảo mật.

Hệ thống xuất trình bằng cấp dựa trên Hyperledger Fabric và IPFS hoạt động theo quy trình như sau:

- Khi một bằng cấp được phát hành, nó sẽ được mã hóa bằng AES trước khi tải lên IPFS bằng phương thức **'encryptData'**.
- Tài liệu đã mã hóa được tải lên IPFS thông qua hàm **'uploadToIPFS'**, sau đó nhận CID duy nhất.
- CID của tài liệu được ghi vào blockchain Hyperledger Fabric để đảm bảo tính nguyên vẹn và truy xuất nhanh chóng.
- Khi cần xác minh bằng cấp, hệ thống sẽ truy xuất CID từ blockchain và tải về từ IPFS bằng phương thức **'getFileFromIPFS'**.
- Sau đó, tài liệu được giải mã bằng AES để hiển thị, sử dụng hàm **'decryptData'**.

Việc kết hợp AES và IPFS trong hệ thống đảm bảo rằng ngay cả khi CID bị lộ, nội dung thực tế của tài liệu vẫn được bảo vệ an toàn. Chỉ những người có khóa giải mã hợp lệ mới có thể truy cập dữ liệu gốc. Cách tiếp cận này giúp giảm tải cho blockchain, bảo đảm quyền riêng tư và nâng cao khả năng lưu trữ dữ liệu trên hệ thống phân tán.



Hình 2. Minh họa quy trình mã hoá hình ảnh AES trước khi lưu trữ văn bằng theo giao thức IPFS tích hợp với Hyperledger Fabric

Ứng dụng cơ cấu của hợp đồng thông minh để xác thực bằng cấp

Hợp đồng thông minh đóng vai trò quan trọng trong việc xác thực và bảo mật dữ liệu bằng công nghệ blockchain. Với khả năng tự động hóa các quy trình kiểm tra và xác nhận thông tin, hợp đồng thông minh giúp giảm thiểu gian lận và tăng tính minh bạch. Một ứng dụng thực tiễn là hệ thống xác thực bằng cấp sử dụng Hyperledger Fabric và IPFS để lưu trữ và bảo mật dữ liệu.

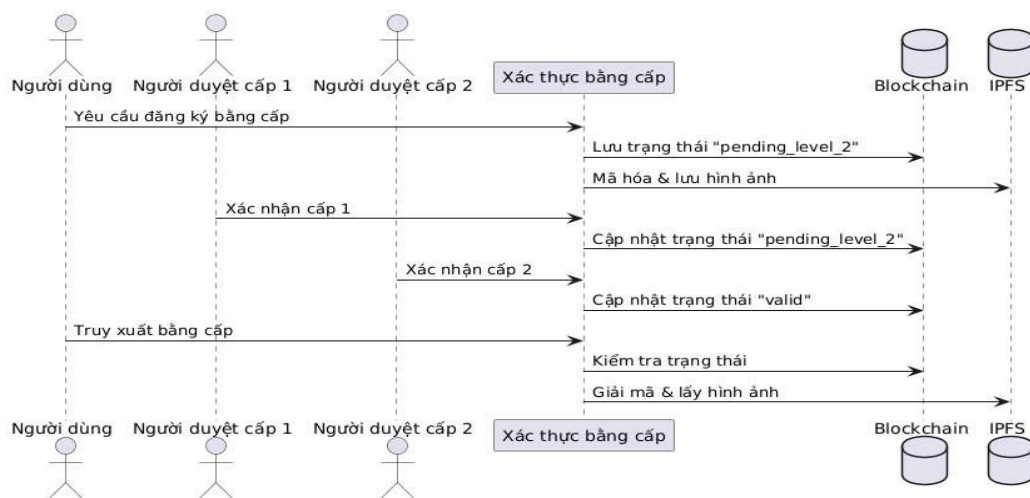
Hệ thống này sử dụng hợp đồng thông minh "DegreeVerification" để quản lý thông tin bằng cấp. Khi một người dùng đăng ký bằng cấp, dữ liệu được lưu trữ trong blockchain với trạng thái "pending_level_2". Sau khi được cấp duyệt bởi hai cấp xác nhận, bằng cấp mới được xem là hợp lệ. Điều này đảm bảo tính toàn vẹn của dữ liệu và ngăn chặn việc giả mạo.

Quy trình hoạt động của hệ thống như sau:

- Khi một bằng cấp mới được tạo, hệ thống kiểm tra xem ID có tồn tại chưa. Nếu chưa, nó sẽ được lưu trữ với trạng thái chờ duyệt.
- Người duyệt cấp 1 phê duyệt thông tin, sau đó bằng cấp được chuyển sang trạng thái "pending_level_2".
- Người duyệt cấp 2 xác thực và thay đổi trạng thái của bằng cấp thành "valid".
- Toàn bộ quá trình lưu trữ và thay đổi trạng thái đều được ghi lại trên blockchain, đảm bảo tính bất biến và minh bạch.

Ngoài ra, hệ thống còn kết hợp IPFS để lưu trữ hình ảnh bằng cấp dưới dạng mã hóa AES-256-CBC [11]. Khi người dùng tải lên hình ảnh, dữ liệu sẽ được mã hóa trước khi lưu vào IPFS. Điều này giúp bảo vệ thông tin khỏi truy cập trái phép. Khi cần truy xuất, dữ liệu sẽ được giải mã để hiển thị thông tin gốc.

Hệ thống cũng sử dụng thuật toán băm SHA-256 [12] để tạo hash cho hình ảnh bằng cấp, đảm bảo rằng dữ liệu không bị thay đổi trong quá trình lưu trữ. Bằng cách so sánh hash, hệ thống có thể xác định tính toàn vẹn của dữ liệu khi truy xuất từ IPFS.



Hình 3. Minh họa quy trình kiểm tra và xác thực dữ liệu

Với cách tiếp cận này, việc xác thực và bảo vệ dữ liệu trong hệ thống trở nên an toàn và hiệu quả. Hợp đồng thông minh không chỉ giúp tự động hóa quy trình kiểm tra mà còn đảm bảo rằng thông tin không bị thay đổi hay giả mạo trong quá trình xử lý.

Kết quả ứng dụng xác thực bằng cấp nhóm đã từng thực hiện



Hình 4. Hình ảnh văn bằng được xuất trình từ phần mềm

3. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Với phiên bản hiện tại, ứng dụng đã triển khai thành công Fabric Certificate Authority (Fabric CA) để đăng ký và quản lý danh tính của các bên liên quan, bao gồm tổ chức cấp bằng, người xét duyệt và người nhận bằng cấp. Bên cạnh đó, ứng dụng cũng đã tích hợp InterPlanetary File System (IPFS) để hỗ trợ việc chuyển đổi hình ảnh bằng cấp thành mã định danh duy nhất (CID – Content Identifier) trước khi ghi nhận lên blockchain Hyperledger Fabric. Do blockchain không thể trực tiếp lưu trữ hình ảnh vì hạn chế về dung lượng và chi phí, IPFS đóng vai trò trung gian, giúp lưu trữ hình ảnh một cách phân tán và chỉ lưu CID trên blockchain nhằm đảm bảo tính toàn vẹn và khả năng truy xuất dữ liệu khi cần thiết. Đồng thời, Smart Contract (hợp đồng thông minh) tự động hóa quá trình xác thực, kiểm tra các điều kiện cấp bằng và chỉ chấp nhận các giao dịch hợp lệ trước khi ghi vào blockchain. Sự kết hợp giữa Hyperledger Fabric, IPFS và Smart Contract không chỉ tối ưu hóa quy trình xác thực văn bằng theo cách phi tập trung mà còn giảm thiểu rủi ro giả mạo, đảm bảo tính minh bạch và giúp doanh nghiệp, tổ chức giáo dục dễ dàng kiểm tra, xác nhận văn bằng mà không cần phụ thuộc vào cơ sở dữ liệu tập trung.

Và trong thời gian tới, hệ thống sẽ được cải thiện và nâng cấp nhằm đáp ứng tốt hơn nhu cầu thực tế. Một trong những hướng phát triển quan trọng là áp dụng trí tuệ nhân tạo (AI) vào quá trình kiểm soát cấp bằng, giúp nhận diện và phát hiện các sai sót hoặc gian lận trong quá trình cấp phát văn bằng. Các kỹ thuật học sâu (Deep Learning) sẽ được sử dụng để nhận dạng khuôn mặt, giúp xác thực danh tính người nhận một cách chính xác và nhanh chóng, hạn chế tối đa tình trạng giả mạo. Ngoài ra, việc tích hợp chip điện tử vào bằng cấp có thể giúp tạo ra một cơ chế bảo mật vật lý bổ sung, đảm bảo tính duy nhất của từng văn bằng. Hệ thống cũng sẽ được mở rộng về mặt hạ tầng, với ít nhất sáu máy chủ chạy song song nhằm đảm bảo khả năng xử lý dữ liệu lớn và duy trì hiệu suất cao ngay cả khi có nhiều người dùng truy cập cùng lúc. Bên cạnh đó, một hướng đi quan trọng là triển khai cơ chế phân tán bằng cấp theo mô hình phi tập trung, giúp người dùng có thể xác thực văn bằng mà không cần phụ thuộc vào một cơ quan trung gian duy nhất. Những cải tiến này không chỉ giúp nâng cao tính bảo mật và minh bạch của hệ thống mà còn góp phần tối ưu hóa chi phí vận hành, tăng cường khả năng mở rộng và ứng dụng công nghệ blockchain trong lĩnh vực giáo dục. Một kế hoạch triển khai demo tại Trường Đại học Công Thương Thành phố Hồ Chí Minh sẽ được thực hiện để kiểm thử và hoàn thiện hệ thống. Sau khi hệ thống được ổn định và hoàn chỉnh, dự kiến sẽ triển khai thực tế cho các trường đại học và cơ sở giáo dục khác trong môi trường giáo dục Việt Nam.

TÀI LIỆU THAM KHẢO

1. Wikipedia, "Axact," Wikipedia, The Free Encyclopedia, Mar. 21, 2024. [Online]. Available: <https://en.wikipedia.org/wiki/Axact>. [Accessed: Mar. 22, 2025].
2. Đại Biểu Nhân Dân, "Gần 1.800 hồ sơ văn bằng do cơ sở nước ngoài cấp chưa được công nhận tại Việt Nam, Bộ GD-ĐT nói gì?", Báo Đại Biểu Nhân Dân, Nov. 24, 2023. [Online]. Available: <https://daibieunhandan.vn/gan-1-800-ho-so-van-bang-do-co-so-nuoc-ngoai-cap-chua-duoc-cong-nhan-tai-viet-nam-bo-gd-dt-noi-gi-post354108.html>. [Accessed: Mar. 22, 2025].
3. Saad, M., Haidery, S. A., Bhandari, A., Bhutta, M. R., Park, D. J., & Chung, T. S. (2024). *An Efficient Privacy and Anonymity Setup on Hyperledger Fabric for Blockchain-Enabled Internet of Things (IoT) Devices*. *Electronics*, 13(13), 2652.
4. Abang, J. E., Takruri, H., Al-Zaidi, R., & Al-Khalidi, M. (2024). *Latency performance modelling in hyperledger fabric blockchain: Challenges and directions with an IoT perspective*. *Internet of Things*, 101217.

5. GeeksforGeeks, "Transaction Flow in Hyperledger Fabric," *GeeksforGeeks*, 2024. [Online]. Available: <https://www.geeksforgeeks.org/transaction-flow-in-hyperledger-fabric/>. [Accessed: 23-Mar-2025].
6. Hyperledger Fabric Documentation, "Ledger," *Hyperledger Fabric Read the Docs*, 2024. [Online]. Available: <https://hyperledger-fabric.readthedocs.io/en/latest/ledger/ledger.html>. [Accessed: 23-Mar-2025].
7. Hyperledger Fabric Documentation, *Hyperledger Fabric: A Blockchain Platform for the Enterprise*, 2023. [Online]. Available: <https://hyperledger-fabric.readthedocs.io/en/release-2.2/>. [Accessed: Mar. 22, 2025].
8. Hyperledger Fabric Documentation, "Gossip Data Dissemination," *Hyperledger Fabric Read the Docs*, 2024. [Online]. Available: <https://hyperledger-fabric.readthedocs.io/en/latest/gossip.html>. [Accessed: 23-Mar-2025].
9. S. Sutradhar, S. Karforma, R. Bose, S. Roy, S. Djebali, and D. K. Bhattacharyya, "Enhancing identity and access management using Hyperledger Fabric and OAuth 2.0," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 49–67, 2024.
10. S. K. Jena, B. Kumar, B. Mohanty, A. Singhal, and R. C. Barik, "An advanced blockchain-based hyperledger fabric solution for tracing fraudulent claims in the healthcare industry," *Decision Analytics Journal*, vol. 10, p. 100411, 2024.
11. A. S. Dewi and H. Setiawan, "Implementation of SHA-256 and AES-256 for Securing Digital Al Quran Verification System," in 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1–6.
12. M. Crosby, P. Pattanayak, S. Verma, and V. Kalyanaraman, "Blockchain technology: Beyond bitcoin," *Applied Innovation Review*, vol. 2, pp. 6–10, 2016.

ABSTRACT

AUTHENTICATION AND SECURITY OF DEGREE CERTIFICATES USING BLOCKCHAIN HYPERLEDGER FABRIC

Le Quang Vinh¹, Nguyen Thi Thanh Thao¹, Huynh Minh An¹, Lam Nguyen Anh Hao¹,
Nguyen Ngoc Nghia¹, Dang Ngoc Tai¹, Ha Huynh Anh Ngan¹, Huynh Cong Tien¹,
Phan Cong Hieu¹, Nguyen Thi Bich Ngan^{1,*}.

¹*Faculty of Information Technology,*

Ho Chi Minh City University of Industry and Trade

*Email: 1001240021@huit.edu.vn, 2033216553@hufi.edu.vn, 2001220056@huit.edu.vn,
2001221180@huit.edu.vn, 2001223033@huit.edu.vn, 2001224231@huit.edu.vn,
2001222912@huit.edu.vn, 2001224389@huit.edu.vn, 2001215781@hufi.edu.vn,
nganntb@huit.edu.vn

Currently, the issue of degree counterfeiting is becoming increasingly serious, negatively impacting the reputation of educational institutions and businesses. This not only results in financial losses when businesses recruit unqualified individuals but also undermines trust in the education system. Previously, degree authentication primarily relied on anti-counterfeiting stamps or verification through traditional databases; however, these methods are easily counterfeited, difficult to verify the source, and lack transparency. To address these challenges, this paper proposes a solution that integrates the Hyperledger Fabric blockchain

with the InterPlanetary File System (IPFS) to ensure security and authenticate diplomas in a decentralized manner. Due to the high cost and limited storage capacity of blockchain technology, IPFS serves as an intermediary by storing the digital diploma and generating a unique identifier (CID - Content Identifier), which is then recorded on the Hyperledger Fabric blockchain to ensure data integrity and traceability. Furthermore, to verify the authenticity of the diploma and the identities of the involved parties, the system employs the Fabric Certificate Authority (Fabric CA) to issue digital certificates based on the X.509 standard, enabling authentication of the issuing institution and the recipient. Through the research and implementation process, we have successfully developed a Fabric CA mobile application to register and manage the identities of the involved parties, including the institution issuing the diploma, the reviewer, and the recipient. Moreover, the application integrates IPFS to facilitate the conversion of the digital diploma into a unique identifier (CID) before recording it on the Hyperledger Fabric blockchain.

Keywords: Blockchain Hyperledger Fabric, IPFS, Fabric CA, Smart Contract, X.509 Certificate.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Lê Quang Vinh

Số điện thoại liên hệ: 0933 549 469

NGHIÊN CỨU THUẬT TOÁN KHAI THÁC TẬP HỮU ÍCH CAO TRÊN CƠ SỞ DỮ LIỆU ÂM DƯƠNG THAY ĐỔI

Nguyễn Huy Hoàng¹, Nguyễn Công Dũng¹, Hà Trọng Thắng¹, Vũ Văn Vinh^{1,*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: vinhvv@huit.edu.vn

Ngày nhận bài 26/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Khai thác tập hữu ích cao là bài toán có nhiều ứng dụng trong thực tế và được các nhà khoa học trong và ngoài nước quan tâm. Nhiều thuật khai thác tập hữu ích cao hiệu quả đã được đề xuất và ứng dụng rộng rãi. Tuy nhiên, phần lớn các phương pháp này tập trung vào cơ sở dữ liệu mà độ hữu ích của các mục (item) là số dương và cố định. Các nghiên cứu trên cơ sở dữ liệu chứa các mục có lợi ích âm và thay đổi còn hạn chế. Thuật toán EMHUN đã được đề xuất để phát hiện các tập hữu ích cao trên cơ sở dữ liệu này, nhưng vẫn còn tồn tại những hạn chế về thời gian xử lý và mức tiêu thụ bộ nhớ, đặc biệt khi làm việc với tập dữ liệu lớn. Để khắc phục vấn đề này, nghiên cứu đề xuất một thuật toán cải tiến có tên iEMHUN, được thiết kế nhằm tối ưu hóa hiệu suất khai thác bằng cách rút ngắn thời gian xử lý và giảm thiểu lượng bộ nhớ sử dụng.

Từ khóa: EMHUN, iEMHUN, khai thác tập hữu ích cao, lợi nhuận thay đổi.

1. GIỚI THIỆU

Khai thác tập hữu ích cao (HUIM - High Utility Itemset Mining) là một trong những hướng nghiên cứu quan trọng trong khai thác dữ liệu giao dịch, giúp xác định các tập hợp sản phẩm mang lại giá trị cao cho doanh nghiệp. Các phương pháp truyền thống [1, 2, 3, 4, 5] thường tập trung vào việc khai thác tập hữu ích cao trên cơ sở dữ liệu có lợi nhuận dương, nhưng gặp nhiều hạn chế khi áp dụng trên dữ liệu có lợi nhuận âm dương biến động. Điều này xuất phát từ việc các sản phẩm có thể mang lại lợi nhuận âm do chiết khấu, phí vận chuyển hoặc các giao dịch bị hoàn trả, khiến quá trình khai thác gặp khó khăn và làm giảm độ chính xác của mô hình.

Thuật toán EMHUN [6] được đề xuất để giải quyết bài toán khai thác tập hữu ích cao trong môi trường dữ liệu có lợi nhuận không ổn định. EMHUN sử dụng các chỉ số như RTWU, RSU để cắt tỉa không gian tìm kiếm, giúp cải thiện hiệu suất so với các phương pháp trước đó. Tuy nhiên, qua thực nghiệm, thuật toán này vẫn tồn tại hạn chế về thời gian thực thi và mức tiêu thụ bộ nhớ, đặc biệt khi áp dụng trên các tập dữ liệu lớn. Điều này đặt ra bài toán tối ưu hóa hiệu suất khai thác HUIM mà vẫn đảm bảo độ chính xác của kết quả.

Trong nghiên cứu này, chúng tôi đề xuất thuật toán cải tiến iEMHUN, nhằm nâng cao hiệu quả khai thác HUIM trên dữ liệu có lợi nhuận âm dương thay đổi. Khác với EMHUN, iEMHUN áp dụng cơ chế lọc trước dữ liệu, giúp giảm không gian tìm kiếm và tăng tốc độ xử lý. Ngoài ra, chúng tôi còn ứng dụng thuật toán vào hệ thống bán sản phẩm điện tử, giúp nhà

quản lý đưa ra quyết định kinh doanh tối ưu hơn bằng cách xác định các sản phẩm có giá trị hữu ích cao, hỗ trợ tối ưu hóa danh mục hàng hóa và chiến lược bán hàng.

2. CƠ SỞ LÝ THUYẾT

2.1. Khai phá dữ liệu (Data Mining)

Khai thác dữ liệu (Data Mining) là quá trình phát hiện các mẫu, xu hướng hoặc thông tin hữu ích từ các tập dữ liệu lớn. Quá trình này sử dụng các thuật toán phân tích, học máy và thống kê để trích xuất thông tin tiềm ẩn nhằm hỗ trợ ra quyết định. Khai thác dữ liệu đóng vai trò quan trọng trong nhiều lĩnh vực như thương mại điện tử, tài chính, y tế và khoa học xã hội.

Trong bối cảnh dữ liệu lớn (Big Data), việc khai thác dữ liệu tốt đã giúp doanh nghiệp tổ chức và xử lý được khối lượng dữ liệu khổng lồ một cách hiệu quả hơn. Một số ứng dụng điển hình của khai thác dữ liệu bao gồm: Phân tích hành vi khách hàng; Dự đoán xu hướng mua sắm, tối ưu hóa hệ thống gợi ý sản phẩm; Phát hiện gian lận tài chính; Nhận diện giao dịch bất thường trong hệ thống ngân hàng; Trong y tế và chăm sóc sức khỏe: Hỗ trợ chẩn đoán bệnh dựa trên dữ liệu lịch sử bệnh án.

Một trong những hướng nghiên cứu quan trọng trong khai thác dữ liệu là khai thác tập hữu ích cao (High Utility Itemset Mining - HUIM), nhằm tìm ra các tập hợp mục có giá trị hữu ích cao trong cơ sở dữ liệu giao dịch.

2.2. Khai thác tập hữu ích cao (High Utility Itemset Mining – HUIM)

HUIM là một bài toán trong khai thác dữ liệu nhằm tìm ra các tập hợp mục có tổng lợi nhuận cao hơn một ngưỡng tối thiểu $minU$ được xác định bởi người dùng. Khác với khai thác tập phổ biến (Frequent Itemset Mining - FIM), HUIM không chỉ dựa trên tần suất xuất hiện của các mục mà còn xem xét giá trị lợi nhuận của chúng (chẳng hạn như doanh thu, chi phí, lợi nhuận ròng).

Bài toán HUIM trở nên phức tạp hơn FIM do không thỏa mãn tính chất bao đóng giảm (Downward Closure property), tức là một tập con của một tập hữu ích cao không nhất thiết phải là một tập hữu ích cao.

Giả sử một cửa hàng bán các sản phẩm điện tử với các giao dịch được thể hiện trong Bảng 1 :

Bảng 1. Ví dụ về cơ sở dữ liệu giao dịch

Giao dịch	Sản phẩm	Số lượng	Giá bán	Lợi nhuận
T1	Laptop	1	20 triệu	5 triệu
T1	Chuột	2	500 ngàn	200 ngàn
T2	Bàn Phím	1	1 triệu	300 ngàn

Tập {Laptop, Chuột} có tổng lợi nhuận là 5.2 triệu. Nếu $minU = 4.5$ triệu, thì tập này là tập hữu ích cao (HUI).

2.3. Tập hữu ích cao (High Utility Itemsets - HUIs)

Tập hữu ích cao (High Utility Itemset - HUI) là một tập hợp mục (itemset) có giá trị hữu ích tổng thể vượt qua một ngưỡng tối thiểu (min-utility) do người dùng đặt ra.

Khác với khai thác tập phổ biến (Frequent Itemset Mining - FIM), nơi chỉ quan tâm đến tần suất xuất hiện của tập hợp mục, khai thác tập hữu ích cao còn xem xét giá trị hữu ích của tập hợp mục, phản ánh yếu tố kinh tế thực tế của chúng trong giao dịch.

Ví dụ: Trong một cửa hàng, có những sản phẩm ít được mua nhưng lợi nhuận cao và một số sản phẩm có lượt mua cao nhưng lợi nhuận thấp. HUIM giúp xác định tập hợp sản phẩm có tần suất mua cao với sản phẩm ít được mua nhưng mang lại lợi nhuận cao tạo nên các chiến dịch mua kèm để hỗ trợ ra quyết định kinh doanh hiệu quả hơn.

2.3. Trọng số lợi nhuận giao dịch được định nghĩa lại (Redefined Transaction Weight Utility - RTWU)

Trong các cơ sở dữ liệu có giá trị lợi nhuận âm, việc sử dụng trọng số lợi nhuận giao dịch (Transaction Weight Utility - TWU) thông thường có thể làm mất các tập hữu ích tiềm năng. Để khắc phục điều này, RTWU được giới thiệu nhằm xác định tổng giá trị lợi nhuận của các giao dịch chứa một tập hợp mục, nhưng chỉ tính các giá trị lợi nhuận dương để tránh loại bỏ nhầm các tập hữu ích cao.

TWU [2] của một tập hợp mục (X) trong cơ sở dữ liệu D được tính bằng tổng giá trị hữu ích của tất cả các giao dịch chứa tập hợp mục đó:

$$TWU(X) = \sum_{X \subseteq T_k \wedge T_k \in D} TU(T_k)$$

Trong đó, $TU(T_k)$ là tổng giá trị hữu ích của tất cả các mục trong giao dịch T_k .

Tuy nhiên, khi một giao dịch chứa giá trị hữu ích âm, TWU có thể không phản ánh chính xác giá trị thực của tập hợp mục. Do đó, RTWU được định nghĩa để khắc phục hạn chế này:

$$RTWU(X) = \sum_{X \subseteq T_k \wedge T_k \in D} RTU(T_k)$$

Trong đó, $RTU(T_k)$ là tổng giá trị hữu ích của tất cả các mục có lợi nhuận dương trong giao dịch T_k .

2.4. Lợi nhuận cận trên của cây con được định nghĩa lại (Redefined Sub-tree Utility - RSU).

RSU là một giới hạn trên của giá trị hữu ích thực tế của một tập hợp mục, giúp tối ưu hóa quá trình khai thác bằng cách loại bỏ các tập hợp mục không tiềm năng. RSU được sử dụng để đánh giá giá trị hữu ích của một tập hợp mục và các tập hợp con mở rộng của nó trong cây tìm kiếm.

RSU của một tập hợp mục X với một mục z được tính như sau:

$$RSU(X, z) = \sum_{T_k \in D \wedge \{X \cup z\} \subseteq T_k} [u(X, T_k) + u(z, T_k) + RRU(z, T_k)]$$

Trong đó:

- $u(X, T_k)$ là giá trị hữu ích của tập hợp X trong giao dịch T_k .
- $u(z, T_k)$ là giá trị hữu ích của mục z trong giao dịch T_k .

$RRU(z, T_k)$ là Tiềm năng còn lại được xác định lại của mục z trong giao dịch T_k .

2.5. Tiềm ích còn lại được xác định lại (Redefined Remaining Utility – RRU)

RRU là một biến thể của Remaining Utility (RU), phản ánh giá trị hữu ích dương còn lại phía sau item X của một tập hợp mục trong cơ sở dữ liệu giao dịch.

RRU của một tập hợp mục X trong giao dịch T_k được tính như sau:

$$RRU(X, T_k) = \sum_{i \in RE(X, T_k)} u(i, T_k)$$

Trong đó, $RE(X, T_k)$ là tập hợp các mục xuất hiện sau X trong giao dịch T_k và có giá trị hữu ích dương.

3. THUẬT TOÁN

3.1 Thuật toán EMHUN

Năm 2024, N.T. Tung và cộng sự đã giới thiệu thuật toán EMHUN [6] là thuật toán khai thác tập hữu ích cao được phát triển để giải quyết những hạn chế trong các phương pháp khai thác tập hữu ích cao (HUIM) truyền thống. Điều đặc biệt ở thuật toán EMHUN là khả năng xử lý dữ liệu giao dịch có lợi ích âm, dương thay đổi, giúp tìm kiếm các bộ vật phẩm có giá trị lợi ích cao trong các tình huống phức tạp như khi lợi ích của các vật phẩm thay đổi theo thời gian hoặc khi các bộ vật phẩm có lợi ích âm có thể mang lại lợi ích dương khi kết hợp với nhau.

Các thuật toán HUIM truyền thống thường không hỗ trợ tốt dữ liệu có lợi ích âm, vì chúng giả định rằng tất cả các giá trị hữu ích đều dương. EMHUN được thiết kế để giải quyết vấn đề này, giúp khai thác các tập hữu ích cao một cách chính xác hơn ngay cả khi dữ liệu giao dịch có sự biến động về lợi ích.

Thuật toán EMHUN không chỉ đơn giản là tìm kiếm các bộ vật phẩm có tần suất xuất hiện cao, mà nó còn tính toán lợi ích thực tế mà các bộ vật phẩm này mang lại cho doanh nghiệp, giúp tối ưu hóa lợi nhuận trong môi trường có sự thay đổi về giá cả, khuyến mãi, hoặc các yếu tố chiến lược khác.

Nguyên lý hoạt động của EMHUN:

Đầu tiên, nó phân loại các mục trong cơ sở dữ liệu thành ba nhóm: mục có lợi ích dương, mục có cả lợi ích dương và âm, và mục có lợi ích âm.

Sau đó, thuật toán tính toán RTWU để xác định các mục có khả năng trở thành tập hữu ích cao. Dựa vào đó, EMHUN loại bỏ các tập hợp mục không tiềm năng nhằm giảm không gian tìm kiếm.

Tiếp theo, RSU được sử dụng để đánh giá giá trị hữu ích của các tập hợp mục trong không gian tìm kiếm.

Cuối cùng, thuật toán tìm kiếm đệ quy các tập hợp mục có giá trị hữu ích cao, loại bỏ các tập hợp không đáp ứng ngưỡng min-utility và xuất kết quả.

Ví dụ. Xét cơ sở dữ liệu trong Bảng 2:

Bảng 2. Cơ sở dữ liệu giao dịch

Giao dịch	Sản phẩm	Lợi nhuận sản phẩm
T1	a, b, d, e, f, g	-4, 2, 4, 3, -2, -2
T2	b, c	-1, 5
T3	b, c, d, e, f	-2, 1, 12, 2, -1
T4	c, d, e	2, 4, 3
T5	a, f	4, -3
T6	a, b, c, d, e, f, g	2, 1, 4, 8, 1, -3, -2
T7	b, c, e	3, 4, 4

Dựa vào ví dụ ta thấy sản phẩm a có giá trị lợi nhuận qua các giao dịch là $\{-4, 4, 2\}$ sẽ thuộc nhóm lai. Từ đó ta có thể sắp xếp các sản phẩm theo các nhóm riêng biệt như sau:

Nhóm dương (ρ - Positive Utility Only): chứa các sản phẩm luôn có lợi nhuận dương, là tập chứa $\{c, d, e\}$.

Nhóm lai (δ - Both Positive & Negative Utility) chứa các sản phẩm có lợi nhuận dương trong giao dịch này và lợi nhuận âm trong giao dịch khác, là tập $\{a, b\}$.

Nhóm âm (η - Negative Utility Only): Chỉ chứa các sản phẩm có lợi nhuận âm, là tập $\{f, g\}$.

Các bước thực hiện của thuật toán EMHUN

Bảng 3. Bảng thể hiện cách tính RTWU(z)

Sản phẩm	Giao dịch chứa sản phẩm	Tổng lợi nhuận dương trong giao dịch đó	RTWU(z)
a	T1, T5, T6	(T1: $2+4+3=9$) + (T5: 4) + (T6: $2+1+4+8+1=16$)	$9 + 4 + 16 = 29$
b	T1, T2, T3, T6, T7	(T1: 9) + (T2: 5) + (T3: $1+12+2=15$) + (T6: 16) + (T7: $3+4+4=11$)	$9 + 5 + 15 + 16 + 11 = 56$
c	T2, T3, T4, T6, T7	(T2: 5) + (T3: 15) + (T4: $2+4+3=9$) + (T6: 16) + (T7: 11)	$5 + 15 + 9 + 16 + 11 = 56$
d	T1, T3, T4, T6	(T1: 9) + (T3: 15) + (T4: 9) + (T6: 16)	$9 + 15 + 9 + 16 = 49$
e	T1, T3, T4, T6, T7	(T1: 9) + (T3: 15) + (T4: 9) + (T6: 16) + (T7: 11)	$9 + 15 + 9 + 16 + 11 = 60$
f	T1, T3, T5, T6	(T1: 9) + (T3: 15) + (T5: 4) + (T6: 16)	$9 + 15 + 4 + 16 = 44$
g	T1, T6	(T1: 9) + (T6: 16)	$9 + 16 = 25$

Bước 1: $X \leftarrow \emptyset$: Khởi tạo tập hợp mục rỗng X. Sau đó, Tính toán RTWU cho tất cả các mục trong tập hợp $\rho \cup \delta$.

Áp dụng công thức của RTWU: Sản phẩm a tồn tại ở giao dịch T1, T5, T6. Tại T1 ta có tổng lợi nhuận dương $RTU(T1) = 9$, tương tự $RTU(T5) = 4$, $RTU(T6) = 16$.

Khi đó, $RTWU(a) = RTU(T1) + RTU(T5) + RTU(T6) = 29$.

Từ đó Bảng 3 thể hiện kết quả của RTWU của tất cả sản phẩm trong cơ sở dữ liệu:

Bước 2: Lọc danh sách các mục có khả năng trở thành ứng viên tiếp tục khai thác bằng cách chỉ giữ lại những mục có $RTWU \geq minU$.

Giả sử: $minU = 25$. Dựa vào RTWU khi so sánh với giá trị $minU$, tìm được tập các sản phẩm có tiềm năng cao là $Secondary(X) = [a, b, c, d, e]$ trong tập hợp $\rho \cup \delta$ và $\eta = [f, g]$.

Bước 3: Loại bỏ các mục không tiềm năng khỏi tập dữ liệu: tất cả các sản phẩm đều nằm trong tập hữu ích cao nên không có sản phẩm nào bị loại khỏi tập dữ liệu.

Bước 4: Sắp xếp lại giao dịch theo giá trị RTWU từ thấp đến cao: mục có lợi nhuận dương trước, mục lai tiếp theo và mục có lợi nhuận âm cuối cùng.

Kết quả: $Secondary(X) = [d, c, e, a, b]$ trong tập hợp $\rho \cup \delta$ và $\eta = [g, f]$.

Bước 5: Sắp xếp lại các giao dịch theo danh sách các sản phẩm trong giao dịch đó.

Tiếp tục sắp xếp theo sản phẩm xuất hiện nhiều nhất từ cuối giao dịch trở lên. Ví dụ trong tất cả các giao dịch có sản phẩm f xuất hiện ở cuối cùng nhiều nhất (T6, T1, T3, T5), sẽ ưu tiên sắp xếp bốn giao dịch này. Tiếp theo xét tới sản phẩm g xuất hiện 2 lần ở (T6, T1). Tiếp tục với sản phẩm b, sản phẩm a, sản phẩm e. Khi tới sản phẩm khác nhau (T6: 3), (T1: 4) ta sẽ xét RTWU của c và d: $56 > 49$ nên T6 xếp trước T1. Tương tự với T3 và T5. Tiếp theo tới T7 và T2. Cuối cùng là T4.

Bảng 4. Cơ sở dữ liệu sau khi sắp xếp các giao dịch

Giao dịch	Sản phẩm	Lợi nhuận
T6	d, c, e, a, b, g, f	8, 4, 1, 2, 1, - 2, - 3
T1	d, e, a, b, g, f	4, 3, - 4, 2, - 2, - 2
T3	d, c, e, b, f	12, 1, 2, - 2, - 1
T5	a, f	4, - 3
T7	c, e, b	4, 4, 3
T2	c, b	5, - 1
T4	d, c, e	4, 2, 3

Bước 6: Tính toán RSU cho từng sản phẩm.

Áp dụng công thức tính RSU: Sản phẩm a tồn tại ở giao dịch T1, T5, T6 tại T1 ta có $u(X, T1) = 0$. Tương tự, kết quả các RSU(X,z) được thể hiện trong Bảng 5.

Bước 7: Lọc danh sách các mục chắc chắn có độ hữu ích cao bằng cách chỉ giữ lại những mục có $RSU(X, z) \geq minU$.

Dựa vào RSU khi so sánh với giá trị $minU$ ta tìm được tập các sản phẩm chắc chắn có độ hữu ích cao là $Primary = [c, d]$.

Kết quả: $Secondary(X) = [d, c, e, a, b]$. $Primary(X) = [d, c]$.

Bảng 5. Cách tính RSU(X,z)

Sản phẩm	Giao dịch chứa sản phẩm	Cách tính	RSU(X,z)
a	T1, T5, T6	(T1: $(-4) + 2 = -2$) + (T5: 4) + (T6: $2 + 1 = 3$)	$-2 + 4 + 3 = 5$
b	T1, T2, T3, T6, T7	(T1: 2) + (T2: -1) + (T3: -2) + (T6: 1) + (T7: 3)	$2 - 1 - 2 + 1 + 3 = 3$
c	T2, T3, T4, T6, T7	(T2: 5) + (T3: $1 + 2 = 3$) + (T4: $2 + 3 = 5$) + (T6: $4 + 1 + 2 + 1 = 8$) + (T7: $4 + 4 + 3 = 11$)	$5 + 3 + 5 + 8 + 11 = 32$
d	T1, T3, T4, T6	(T1: $4 + 3 + 2 = 9$) + (T3: $12 + 1 + 2 = 15$) + (T4: $4 + 2 + 3 = 9$) + (T6: $8 + 4 + 1 + 2 + 1 = 16$)	$9 + 15 + 9 + 16 = 49$
e	T1, T3, T4, T6, T7	(T1: $3 + 2 = 5$) + (T3: 2) + (T4: 3) + (T6: $1 + 2 + 1 = 4$) + (T7: $4 + 3 = 7$)	$5 + 2 + 3 + 4 + 7 = 21$
f	T1, T3, T5, T6	(T1: -3) + (T3: -1) + (T5: -3) + (T6: -3)	
g	T1, T6	(T1: 9) + (T6: 16)	

Bước 8: Áp dụng tìm kiếm đệ quy để phát hiện tập hữu ích cao.

Trong thuật toán EMHUN, khi mở rộng tập hữu ích cao (HUI), nó ưu tiên ghép cặp với các sản phẩm có lợi ích âm trước khi ghép với các sản phẩm có lợi ích dương.

Ví dụ: Ta có $Primary = [d, c]$ là hai sản phẩm có độ hữu ích cao sẽ được mở rộng với sản phẩm có lợi ích âm như g hoặc f.

Ghép $\{d\}$ với $\{g\}$

- $\{d\}$ có độ hữu ích là 28.
- $\{d, g\}$ có độ hữu ích là 8 (nhỏ hơn minU). Loại bỏ $\{d, g\}$

Ghép $\{d\}$ với $\{f\}$

- $\{d, f\}$ có độ hữu ích là 18 (nhỏ hơn minU). Loại bỏ $\{d, f\}$

Nếu $\{d, g\}$ hoặc $\{d, f\}$ có độ hữu ích lớn hơn hoặc bằng $minU$, thuật toán sẽ tiếp tục mở rộng. Vì cả hai đều không thỏa, thuật toán sẽ bỏ qua các nhánh này.

Khi không thể mở rộng với sản phẩm có lợi ích âm nào nữa, thuật toán chuyển sang mở rộng với sản phẩm có lợi ích dương (ρ và δ).

Lúc này $\{d\}$ sẽ ghép với $\{c\}, \{e\}, \{b\}$ để tạo $\{d, c\}, \{d, e\}, \{d, b\}$, v.v.

Các tổ hợp này tiếp tục được kiểm tra độ hữu ích để xác định xem có được mở rộng tiếp hay không.

- $\{d, c\}$ có độ hữu ích là 31 (lớn hơn minU). Được giữ lại
- $\{d, e\}$ có độ hữu ích là 37 (lớn hơn minU). Được giữ lại

Sau khi $\{d, c\}$ được chấp nhận, thuật toán tiếp tục mở rộng $\{d, c, e\}$, $\{d, c, e, b\}$, v.v.

Quá trình này tiếp tục được mở rộng theo chiều sâu (DFS) cho đến khi không thể mở rộng thêm.

Bảng 6. Danh sách các tập hữu ích cao

Tập sản phẩm hữu ích	Tổng lợi nhuận
d	28
d c	31
d c e	37
d c e b	27
d e	37
d e b	31
d e b f	25
d b	25

Hiệu suất của thuật toán EMHUN:

Thời gian thực thi nhanh: EMHUN hoạt động hiệu quả trên nhiều loại cơ sở dữ liệu khác nhau, tối ưu hóa quá trình khai thác tập hữu ích cao.

Giảm mức sử dụng bộ nhớ: Nhờ cơ chế loại bỏ các tập hợp mục không tiềm năng, thuật toán giảm đáng kể lượng bộ nhớ tiêu thụ trong quá trình xử lý.

Khả năng mở rộng tốt: EMHUN có thể áp dụng linh hoạt trên cả cơ sở dữ liệu dày đặc và thưa, đảm bảo hiệu suất ổn định khi quy mô dữ liệu tăng.

Hạn chế của thuật toán EMHUN:

Thuật toán EMHUN dù có khả năng khai thác tập hữu ích cao trong cơ sở dữ liệu có lợi ích âm, dương thay đổi nhưng vẫn tồn tại nhiều hạn chế cần cải tiến.

Thứ nhất, chi phí tính toán cao khi xử lý tập dữ liệu lớn do thuật toán phải quét toàn bộ dữ liệu nhiều lần để tính toán RTWU và RSU.

Thứ hai, mức tiêu thụ bộ nhớ lớn vì EMHUN lưu trữ quá nhiều dữ liệu trung gian, đặc biệt khi số lượng giao dịch và mục tăng.

Thứ ba, chiến lược cắt tĩa chưa tối ưu, khiến nhiều tập hợp mục không có tiềm năng vẫn phải kiểm tra, làm tăng thời gian xử lý. Cuối cùng, chưa tối ưu khi dữ liệu có lợi ích âm không ổn định, dẫn đến giảm hiệu suất khai thác

3.2. Thuật toán iEMHUN

Để khắc phục các hạn chế của thuật toán EMHUN truyền thống trong việc khai thác các tập mục lợi nhuận cao (High Utility Itemset Mining — HUIM), thuật toán iEMHUN được đề xuất với các cải tiến đáng kể về tốc độ xử lý, hiệu quả bộ nhớ và khả năng mở rộng trên cơ sở dữ liệu lớn.

Một trong những điểm nâng cấp nổi bật của iEMHUN chính là việc tích hợp B-Set — một cấu trúc dữ liệu dạng tập hợp (set) được sử dụng để lưu truy vết các giao dịch (transaction) của từng sản phẩm (item) trong cơ sở dữ liệu. Cơ chế này giúp tăng tốc quá trình lọc và quét

giao dịch, đồng thời giảm số lượng giao dịch và sản phẩm phải xử lý trong từng bước thuật toán.

Các bước chính của thuật toán iEMHUN:

Thuật toán iEMHUN kế thừa quy trình khai thác tập mục lợi nhuận cao từ EMHUN nhưng được cải tiến bằng cách tích hợp thêm cấu trúc dữ liệu B-Set, giúp tăng tốc độ xử lý và giảm số lượng giao dịch cần quét trong từng bước của thuật toán.

1. Đầu tiên, nó phân loại các mục trong cơ sở dữ liệu thành ba nhóm: mục có lợi ích dương, mục có cả lợi ích dương và âm, và mục có lợi ích âm. Đồng thời, xây dựng ItemTransactionMap (dựa trên phương pháp B-Set) — đây là một tập hợp lưu trữ danh sách vị trí các Transaction tương ứng với từng sản phẩm (item) mà sản phẩm đó xuất hiện.

2. Sau đó, thuật toán tính toán RTWU để xác định các mục có khả năng trở thành tập hữu ích cao. Việc sử dụng B-Set giúp giảm thiểu đáng kể số lượt tra cứu và quét dữ liệu trong các bước tiếp theo, thay vì phải duyệt tuần tự từng giao dịch như cách truyền thống. Dựa vào đó, iEMHUN loại bỏ các tập hợp mục không tiềm năng nhằm giảm không gian tìm kiếm.

3. Tiếp theo, RSU được sử dụng để đánh giá giá trị hữu ích của các tập hợp mục trong không gian tìm kiếm..

4. Cuối cùng, thuật toán tìm kiếm đệ quy các tập hợp mục có giá trị hữu ích cao sử dụng B-Set để truy xuất nhanh các giao dịch liên quan, từ đó tính toán giá trị lợi ích thực tế. , loại bỏ các tập hợp không đáp ứng ngưỡng min-utility và xuất kết quả.

Ví dụ. Giả sử cơ sở dữ liệu giao dịch như sau:

Bảng 7. Cơ sở dữ liệu giao dịch

Giao dịch	Sản phẩm	Lợi nhuận sản phẩm
T1	a, b, d, e, f, g	-4, 2, 4, 3, -2, -2
T2	b, c	-1, 5
T3	b, c, d, e, f	-2, 1, 12, 2, -1
T4	c, d, e	2, 4, 3
T5	a, f	4, -3
T6	a, b, c, d, e, f, g	2, 1, 4, 8, 1, -3, -2
T7	b, c, e	3, 4, 4

Dựa vào thông tin của cơ sở dữ liệu giao dịch iEMHUN tạo ra ItemTransactionMap(dựa trên phương pháp B-Set):

Bảng 8. Bảng ItemTransactionMap (Dựa trên B-Set)

Sản phẩm (Item)	Giao dịch chứa sản phẩm đó (Transaction)
a	T1, T5, T6
b	T1, T2, T3, T6, T7
c	T2, T3, T4, T6, T7
d	T1, T3, T4, T6
e	T1, T3, T4, T6, T7
f	T1, T3, T5, T6
g	T1, T6

Các bước thực hiện của thuật toán iEMHUN

Bước 1: $X \leftarrow \emptyset$: Khởi tạo tập hợp mục rỗng X . xây dựng ItemTransactionMap dựa trên B-Set.

Ở bước này, tất cả các mục trong cơ sở dữ liệu được phân loại thành ba nhóm:.

Nhóm dương (ρ - Positive Utility Only): chứa các sản phẩm luôn có lợi nhuận dương, là tập chứa $\{c, d, e\}$.

Nhóm lai (δ - Both Positive & Negative Utility) chứa các sản phẩm có lợi nhuận dương trong giao dịch này và lợi nhuận âm trong giao dịch khác, là tập $\{a, b\}$.

Nhóm âm (η - Negative Utility Only): Chỉ chứa các sản phẩm có lợi nhuận âm, là tập $\{f, g\}$.

Đồng thời, thuật toán xây dựng ItemTransactionMap(Bảng 8) — một tập hợp (set) lưu lại danh sách vị trí các giao dịch mà mỗi sản phẩm xuất hiện. Nhờ vậy, việc truy xuất các giao dịch chứa sản phẩm bất kỳ trở nên nhanh chóng và chính xác mà không cần duyệt tuần tự toàn bộ cơ sở dữ liệu như các thuật toán truyền thống.

Bước 2: Tính toán RTWU cho tất cả các sản phẩm(item) trong tập hợp $\rho \cup \delta$ dựa vào itemTransactionMap(Bảng 8) thay vì phải duyệt tuần tự từng giao dịch để tính toán theo như cách truyền thống.

Áp dụng công thức của RTWU: Sản phẩm a tồn tại ở giao dịch $T1, T5, T6$. Tại $T1$ ta có tổng lợi nhuận dương $RTU(T1) = 9$, tương tự $RTU(T5) = 4, RTU(T6) = 16$.

Khi đó, $RTWU(a) = RTU(T1) + RTU(T5) + RTU(T6) = 29$.

Từ đó Bảng 3 thể hiện kết quả của RTWU của tất cả sản phẩm trong cơ sở dữ liệu:

Bước 3: Lọc danh sách các mục có khả năng trở thành ứng viên tiếp tục khai thác bằng cách chỉ giữ lại những mục có $RTWU \geq minU$.

Giả sử: $minU = 25$. Dựa vào RTWU khi so sánh với giá trị $minU$, tìm được tập các sản phẩm có tiềm năng cao là $Secondary(X) = [a, b, c, d, e]$ trong tập hợp $\rho \cup \delta$ và $\eta = [f, g]$.

Bước 4: Loại bỏ các mục không tiềm năng khỏi tập dữ liệu và cập nhật lại itemTransactionMap(Bảng 8): tất cả các sản phẩm đều nằm trong tập hữu ích cao nên không có sản phẩm nào bị loại khỏi tập dữ liệu.

Bước 5: Sắp xếp lại giao dịch theo giá trị RTWU từ thấp đến cao: mục có lợi nhuận dương trước, mục lai tiếp theo và mục có lợi nhuận âm cuối cùng.

Kết quả: $Secondary(X) = [d, c, e, a, b]$ trong tập hợp $\rho \cup \delta$ và $\eta = [g, f]$.

Bước 6: Sắp xếp lại các giao dịch theo danh sách các sản phẩm trong giao dịch đó.

Tiếp tục sắp xếp theo sản phẩm xuất hiện nhiều nhất từ cuối giao dịch trở lên. Ví dụ trong tất cả các giao dịch có sản phẩm f xuất hiện ở cuối cùng nhiều nhất ($T6, T1, T3, T5$), sẽ ưu tiên sắp xếp bốn giao dịch này. Tiếp theo xét tới sản phẩm g xuất hiện 2 lần ở ($T6, T1$). Tiếp tục với sản phẩm b , sản phẩm a , sản phẩm e . Khi tới sản phẩm khác nhau ($T6: 3$), ($T1: 4$) ta sẽ xét RTWU của c và d : $56 > 49$ nên $T6$ xếp trước $T1$. Tương tự với $T3$ và $T5$. Tiếp theo tới $T7$ và $T2$. Cuối cùng là $T4$.

Bước 7: Tính toán RSU cho từng sản phẩm.

Áp dụng công thức tính RSU và sử dụng itemTransactionMap(Bảng 8): Sản phẩm a tồn tại ở giao dịch T1, T5, T6 tại T1 ta có $u(X, T1) = 0$. Tương tự, kết quả các $RSU(X, z)$ được thể hiện trong Bảng 5.

Bước 8: Lọc danh sách các mục chắc chắn có độ hữu ích cao bằng cách chỉ giữ lại những mục có $RSU(X, z) \geq minU$.

Dựa vào RSU khi so sánh với giá trị minU ta tìm được tập các sản phẩm chắc chắn có độ hữu ích cao là $Primary = [c, d]$.

Kết quả: $Secondary(X) = [d, c, e, a, b]$. $Primary(X) = [d, c]$.

Bước 9: Áp dụng tìm kiếm đệ quy để phát hiện tập hữu ích cao.

Trong thuật toán EMHUN, khi mở rộng tập hữu ích cao (HUI), nó ưu tiên ghép cặp với các sản phẩm có lợi ích âm trước khi ghép với các sản phẩm có lợi ích dương.

Các bước tính toán đều sẽ dùng itemTransactionMap(Bảng 8) thay vì phải duyệt tuần tự từng giao dịch để tính toán theo như cách truyền thống.

Ví dụ: Ta có $Primary = [d, c]$ là hai sản phẩm có độ hữu ích cao sẽ được mở rộng với sản phẩm có lợi ích âm như g hoặc f.

Ghép $\{d\}$ với $\{g\}$

- $\{d\}$ có độ hữu ích là 28.
- $\{d, g\}$ có độ hữu ích là 8 (nhỏ hơn minU). Loại bỏ $\{d, g\}$

Ghép $\{d\}$ với $\{f\}$

- $\{d, f\}$ có độ hữu ích là 18 (nhỏ hơn minU). Loại bỏ $\{d, f\}$

Nếu $\{d, g\}$ hoặc $\{d, f\}$ có độ hữu ích lớn hơn hoặc bằng minU, thuật toán sẽ tiếp tục mở rộng. Vì cả hai đều không thỏa, thuật toán sẽ bỏ qua các nhánh này.

Khi không thể mở rộng với sản phẩm có lợi ích âm nào nữa, thuật toán chuyển sang mở rộng với sản phẩm có lợi ích dương (ρ và δ).

Lúc này $\{d\}$ sẽ ghép với $\{c\}, \{e\}, \{b\}$ để tạo $\{d, c\}, \{d, e\}, \{d, b\}$, v.v.

Các tổ hợp này tiếp tục được kiểm tra độ hữu ích để xác định xem có được mở rộng tiếp hay không.

- $\{d, c\}$ có độ hữu ích là 31 (lớn hơn minU). Được giữ lại
- $\{d, e\}$ có độ hữu ích là 37 (lớn hơn minU). Được giữ lại

Sau khi $\{d, c\}$ được chấp nhận, thuật toán tiếp tục mở rộng $\{d, c, e\}, \{d, c, b\}$, v.v.

Quá trình này tiếp tục được mở rộng theo chiều sâu (DFS) cho đến khi không thể mở rộng thêm. Kết thúc trả về danh sách tập hữu ích cao Bảng 6.

Ưu điểm của thuật toán iEMHUN:

Thời gian thực thi vượt trội: iEMHUN cải thiện rõ rệt tốc độ thực thi so với EMHUN nhờ tích hợp cấu trúc dữ liệu B-Set. Việc truy xuất giao dịch chứa sản phẩm trở nên nhanh chóng mà không cần quét tuần tự toàn bộ cơ sở dữ liệu. Điều này đặc biệt hiệu quả trên các cơ sở dữ liệu lớn hoặc dữ liệu dày đặc, giúp giảm số lượt quét và thao tác đọc dữ liệu.

Giảm mức sử dụng bộ nhớ: Nhờ cơ chế loại bỏ các tập hợp mục không tiềm năng, thuật toán giảm đáng kể lượng bộ nhớ tiêu thụ trong quá trình xử lý.

Khả năng mở rộng tốt: iEMHUN có thể áp dụng linh hoạt trên cả cơ sở dữ liệu dày đặc và thưa, đảm bảo hiệu suất ổn định khi quy mô dữ liệu tăng.

4. THỰC NGHIỆM

4.1. Bộ dữ liệu thực nghiệm

Thực nghiệm được tiến hành trên tập dữ liệu Foodmart, Chess, Mushroom, Liquor, BMS một tập dữ liệu giao dịch phổ biến trong khai thác dữ liệu.

Bảng 9. Thông tin cơ sở dữ liệu thực nghiệm

Tập dữ liệu	Số giao dịch	Số item	Độ dài trung bình	Loại dữ liệu
Foodmart	54.357	1.559	4,61	Thưa
BMS	59.602	497	2.51	Thưa
Mushroom	8.124	119	23	Dày
Chess Dynamic	3.196	75	37	Dày
Liquor	90.826	4.603	10,28	Dày

Định dạng dữ liệu: Mỗi giao dịch chứa một tập hợp các sản phẩm được mua kèm theo tổng lợi nhuận của giao dịch đó.

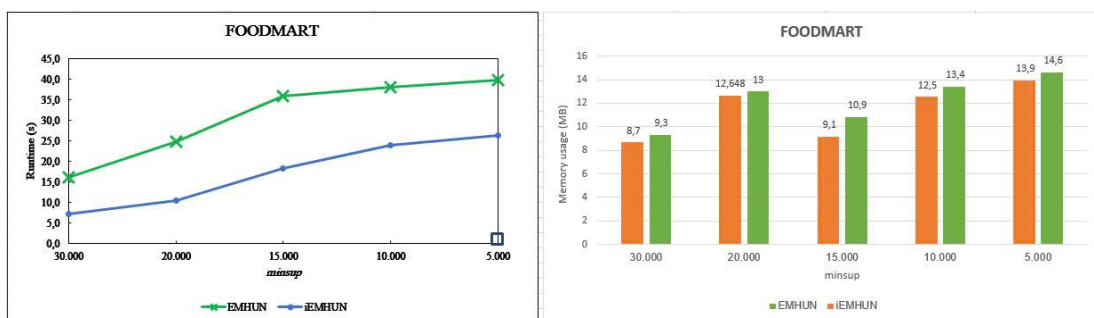
4.2. Tiêu chí đánh giá

Quá trình thực nghiệm được đánh giá dựa trên hai tiêu chí chính:

- Thời gian thực thi (Runtime - s)
 - Đo lường thời gian thực thi của thuật toán trên tập dữ liệu với các mức minsup khác nhau.
 - So sánh thời gian giữa EMHUN và iEMHUN để xác định mức độ cải thiện.
- Bộ nhớ sử dụng (Memory - MB)
 - Đo lường bộ nhớ tiêu thụ trong quá trình thực thi thuật toán.
 - Phân tích sự thay đổi trong mức tiêu thụ bộ nhớ giữa EMHUN và iEMHUN.

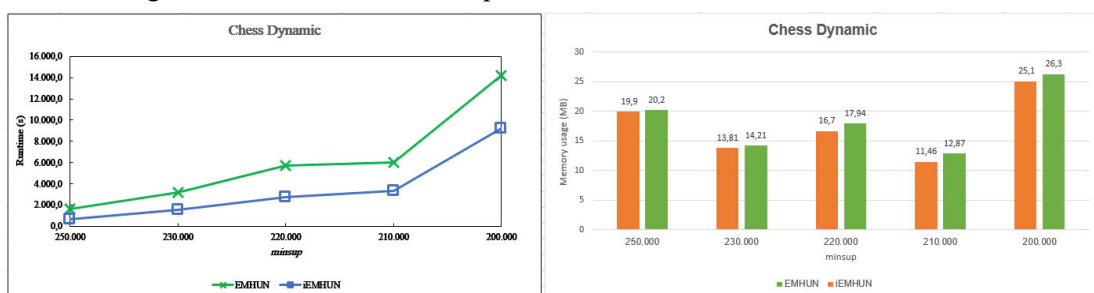
4.3. So sánh hiệu suất

Thực nghiệm được tiến hành trên tập dữ liệu Foodmart



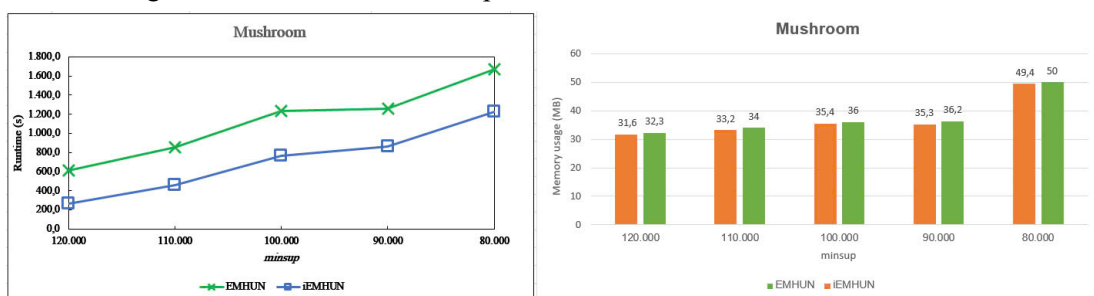
Hình 1. Hiệu suất thực nghiệm trên tập dữ liệu Foodmart

Thực nghiệm được tiến hành trên tập dữ liệu Chess:



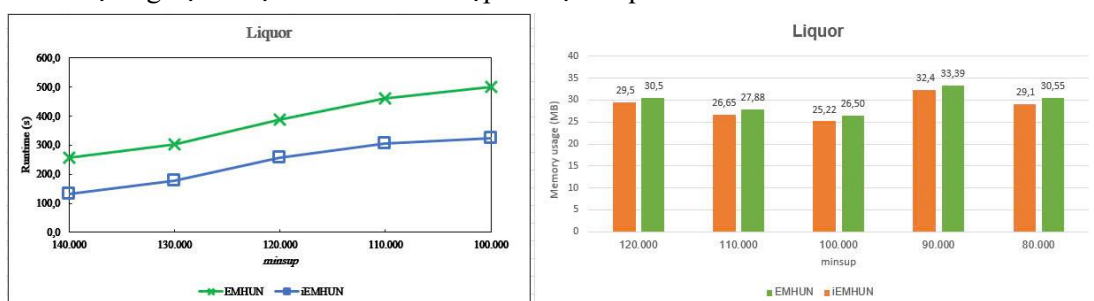
Hình 2. Hiệu suất thực nghiệm trên tập dữ liệu Chess

Thực nghiệm được tiến hành trên tập dữ liệu Mushroom:



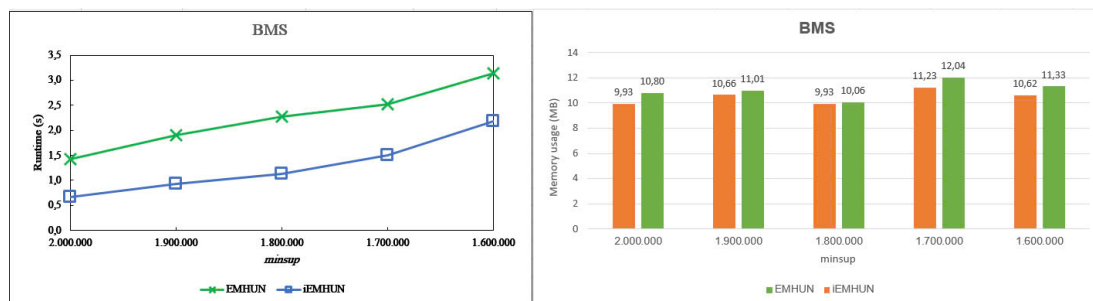
Hình 3. Hiệu suất thực nghiệm trên tập dữ liệu Mushroom

Thực nghiệm được tiến hành trên tập dữ liệu Liquor:



Hình 4. Hiệu suất thực nghiệm trên tập dữ liệu Liquor

Thực nghiệm được tiến hành trên tập dữ liệu BMS:



Hình 5. Hiệu suất thực nghiệm trên tập dữ liệu BMS

Thuật toán iEMHUN đã cho thấy cải thiện đáng kể về thời gian thực thi, với mức giảm trung bình từ 9.17% đến 31.37%, tùy thuộc vào tập dữ liệu. Đặc biệt, trên tập dữ liệu Liquor, thuật toán đạt mức cải thiện cao nhất (31.37%), cho thấy hiệu quả tối ưu hóa rõ rệt trong việc xử lý loại dữ liệu này. Ngược lại, trên tập Mushroom, mức cải thiện chỉ đạt 9.17%, điều này có thể do đặc điểm của tập dữ liệu không tạo ra sự khác biệt lớn khi áp dụng phương pháp lọc trước. Nhìn chung, việc giới hạn phạm vi tìm kiếm giúp iEMHUN tăng tốc độ khai thác tập dữ liệu hiệu quả, đặc biệt trên các tập dữ liệu có cấu trúc phù hợp với cơ chế cắt tỉa của thuật toán.

Về bộ nhớ sử dụng, iEMHUN giúp tiết kiệm từ 1.97% đến 7.34%, tuy mức cải thiện ít hơn so với thời gian thực thi nhưng vẫn đáng kể. Trên tập Foodmart, mức tiết kiệm bộ nhớ đạt 7.34%, cao nhất trong các tập dữ liệu thử nghiệm, chứng tỏ thuật toán hoạt động hiệu quả khi dữ liệu có cấu trúc phù hợp với quá trình lọc trước. Tuy nhiên, trên tập Mushroom, mức tiết kiệm chỉ đạt 1.97%, cho thấy iEMHUN có thể không mang lại lợi ích lớn về bộ nhớ trong mọi trường hợp. Dù vậy, tổng thể, iEMHUN vẫn tối ưu hóa không gian lưu trữ hơn so với EMHUN, giúp cải thiện hiệu suất khai thác trên các tập dữ liệu thực tế.

5. KẾT LUẬN

Bài báo đã nghiên cứu và phân tích thuật toán EMHUN trong khai thác tập dữ liệu hiệu quả trên cơ sở dữ liệu giao dịch có lợi nhuận âm dương thay đổi. Thông qua thực nghiệm, chúng tôi nhận thấy thuật toán EMHUN vẫn còn tồn tại những hạn chế về hiệu suất khi xử lý trên các tập dữ liệu lớn.

Để khắc phục vấn đề này, chúng tôi đã đề xuất thuật toán iEMHUN với cơ chế lọc trước dữ liệu giao dịch dựa trên danh sách sản phẩm quan tâm. Kết quả thực nghiệm trên tập dữ liệu Foodmart cho thấy iEMHUN giúp cải thiện thời gian thực thi trung bình từ 9.17% đến 31.37% và tiết kiệm bộ nhớ trung bình từ 1.97% đến 7.34% so với EMHUN, nhờ vào việc giới hạn phạm vi tìm kiếm và tối ưu hóa không gian bộ nhớ. Tuy nhiên, iEMHUN có thể tiêu thụ bộ nhớ nhiều hơn trong một số trường hợp, điều này cần được cải thiện trong các nghiên cứu tiếp theo.

Kiến nghị:

- Nghiên cứu tiếp các phương pháp giảm bộ nhớ sử dụng của iEMHUN để tối ưu hiệu suất tổng thể.
- Thử nghiệm thuật toán trên các tập dữ liệu lớn hơn và đa dạng hơn để đánh giá tính tổng quát.
- Ứng dụng thuật toán vào hệ thống gợi ý sản phẩm thực tế nhằm kiểm chứng hiệu quả trong thương mại điện tử.

Kết quả nghiên cứu của bài báo có thể được áp dụng trong hệ thống bán lẻ để xác định các sản phẩm có lợi nhuận cao, từ đó hỗ trợ ra quyết định kinh doanh chính xác hơn.

TÀI LIỆU THAM KHẢO

- [1]. Y. Liu, W. Liao, A. Choudhary, "A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets," in *PAKDD 2005*, HaNoi, VietNam, 2005.
- [2]. J. Liu, K. Wang, B.C.M. Fung, "Mining High Utility Patterns in One Phase without Generating Candidates," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1245-1257, 2016.
- [3]. P. Fournier-Viger, C.W. Wu, S. Zida, V.S. Tseng, "FHM: Faster High-Utility Itemset Mining Using Estimated Utility Co-occurrence Pruning," in *ISMIS*, 2014.
- [4]. S. Zida, P. Fournier-Viger, J.C. Lin, C. Wu, V.S. Tseng, "EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 595-625, 2017.
- [5]. S. Krishnamoorthy, "HMiner: Efficiently mining high utility itemsets," *Expert Systems with Applications*, vol. 90, pp. 168-183, 2017
- [6]. N.T. Tung, Trinh D.D. Nguyen, Loan T.T. Nguyen, Bay Vo, "An efficient method for mining High-Utility itemsets from unstable negative profit databases," *Expert Systems with Applications*, vol. 237, no. part C, 2024.

ABSTRACT

RESEARCH ON HIGH-UTILITY ITEMSET MINING IN DATABASES WITH DYNAMIC POSITIVE AND NEGATIVE UTILITY

Nguyễn Huy Hoàng¹, Nguyễn Công Dũng¹, Hà Trọng Thắng¹, Vũ Văn Vinh^{1,*}

¹HO CHI MINH CITY UNIVERSITY OF INDUSTRY AND TRADE

*Email: vinhvv@huit.edu.vn

High utility itemset mining is a problem with many practical applications, so it has attracted the attention of domestic and foreign researchers. Many studies have been published, and many valuable algorithms for mining high utility itemsets have been proposed. However, most of these methods focus on databases where the utility of items is positive and fixed. Research on databases containing items with variable utility and negative values is still limited. Although the EMHUN algorithm [3] has discovered high utility sets on this database, there are still limitations in processing time and memory consumption, especially when working with large datasets. iEMHUN algorithm enhances efficiency by reducing execution time and optimizing memory usage through a data pre-filtering approach, which narrows the search space

Keywords: iEMHUN, EMHUN, High-Utility Itemset Mining (HUIM)

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Vũ Văn Vinh

Số điện thoại liên hệ: 0909 443 886

HỆ THỐNG ĐIỂM DANH SINH VIÊN BẰNG KHUÔN MẶT TẠI KÝ TÚC XÁ TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH

**Bùi Hùng Phương, Huỳnh Thanh Triệu, Huỳnh Vĩ Khang,
Trần Thị Vân Anh*.**

Trường Đại học Công Thương Thành phố Hồ Chí Minh

**Email: anhttv@huit.edu.vn*

Ngày nhận bài: 31/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Để hỗ trợ cho việc quản lý sinh viên tại ký túc xá được chặt chẽ và đảm bảo an ninh, nhóm chúng tôi đề xuất một phương pháp điểm danh sinh viên bằng cách nhận diện qua khuôn mặt của sinh viên đó và tích hợp vào ứng dụng quản lý ký túc xá tại trường Đại học Công thương TP Hồ Chí Minh. Hệ thống tích hợp phương pháp nhận diện khuôn mặt sử dụng OpenCV và Haar Cascade để kiểm soát truy cập và điểm danh tự động. Phương pháp Haar Cascade được áp dụng để phát hiện khuôn mặt nhanh chóng trong thời gian thực, kết hợp với các kỹ thuật tiền xử lý ảnh như chuyển đổi thang độ xám và cân bằng sáng. Thử nghiệm trên tập dữ liệu Yale Face và sinh viên tại ký túc xá HUIT đạt độ chính xác lần lượt là 93.33 % và 90.67%. Từ đó tích hợp vào hệ thống quản lý ký túc xá để điểm danh sinh viên bằng khuôn mặt.

Từ khóa: nhận diện khuôn mặt, OpenCV, điểm danh

1. MỞ ĐẦU

Hiện nay, với sự phát triển mạnh mẽ của công nghệ thông tin, nhiều trường đại học đã áp dụng các giải pháp công nghệ vào công tác quản lý. Trong đó việc điểm danh tự động sinh viên tại ký túc xá thì vẫn còn hạn chế là điểm danh bằng tay và tốn nhiều công sức khi tần suất điểm danh được yêu cầu gia tăng để đảm bảo an ninh và trật tự tại ký túc xá sinh viên. Do đó, hệ thống điểm danh sinh viên tự động tại ký túc xá là một nhu cầu cấp thiết tại Trường Đại học Công Thương thành phố Hồ Chí Minh để đảm bảo kết quả điểm danh một cách chính xác và nhanh chóng. Đồng thời đảm bảo công tác quản lý được an ninh hơn với việc kiểm soát, nhận dạng đúng đối tượng.

Nhận diện khuôn mặt là một trong những ứng dụng quan trọng của thị giác máy tính, được ứng dụng rộng rãi trong giám sát an ninh, kiểm soát truy cập và quản lý danh tính. Với sự phát triển mạnh mẽ của trí tuệ nhân tạo và học máy, nhiều thuật toán nhận diện khuôn mặt đã được đề xuất và triển khai hiệu quả [1].

Năm 2001, Viola-Jones và cộng sự đề xuất phương pháp phát hiện khuôn mặt dựa trên Adaboost kết hợp đặc trưng haar – like để phát hiện khuôn mặt cho tốc độ phát hiện nhanh chóng 0,07 giây và độ chính xác cao lên đến 90% [2]. Phương pháp này cũng được

sử dụng rộng rãi trong các công bố về các phương pháp nhận diện khuôn mặt trong nước và quốc tế.

Đóng góp chính của bài báo: (1) Sử dụng thư viện mã nguồn mở OpenCV kết hợp với haar cascade để phát hiện khuôn mặt và nhận diện sinh viên trong thời gian thực, (2) Xây dựng ứng dụng quản lý ký túc xá với các chức năng hỗ trợ quản lý như: quản lý phòng, đăng ký lưu trú, điểm danh, quản lý dịch vụ phòng, (3) Tích hợp mô hình nhận diện ảnh sinh viên để tự động điểm danh sinh viên có mặt hay không có mặt tại ký túc xá. Hệ thống giúp giảm thiểu sai sót trong quá trình quản lý, tiết kiệm thời gian và chi phí.

Trong phần kế tiếp của bài báo sẽ trình bày như sau: phần 2 cơ sở lý thuyết các thuật toán, phần 3 phương pháp thực hiện, phần 4 kết quả thực nghiệm và cuối cùng là phần kết luận.

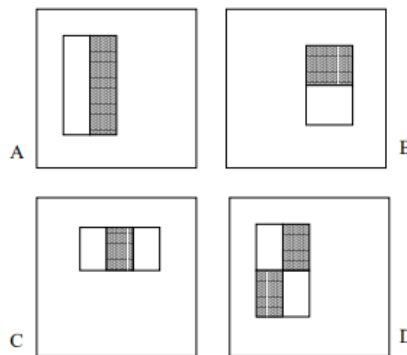
2. CƠ SỞ LÝ THUYẾT

Nhận diện khuôn mặt, một ứng dụng quan trọng của thị giác máy tính, được ứng dụng rộng rãi trong giám sát an ninh và quản lý danh tính, đã đạt được nhiều tiến bộ nhờ sự phát triển của trí tuệ nhân tạo. Nghiên cứu này tập trung vào việc sử dụng OpenCVSharp, một thư viện .NET wrapper của OpenCV, kết hợp với Haar Cascade Classifier để phát hiện khuôn mặt trong thời gian thực. OpenCVSharp hoạt động như một cầu nối giữa ngôn ngữ C# và thư viện OpenCV gốc, tận dụng nền tảng .NET để quản lý bộ nhớ tự động, giúp đơn giản hóa quá trình lập trình và xử lý ảnh.

2.1. Kỹ thuật phát hiện khuôn mặt với đặc trưng Haar-like

Kỹ thuật phát hiện khuôn mặt với đặc trưng Haar-like là một phương pháp hiệu quả và phổ biến trong lĩnh vực thị giác máy tính, đặc biệt là trong các ứng dụng yêu cầu tốc độ xử lý nhanh, chẳng hạn như nhận diện khuôn mặt trong thời gian thực. Phương pháp này được phát triển bởi Paul Viola và Michael Jones vào năm 2001, và đã trở thành một trong những nền tảng quan trọng cho các hệ thống nhận diện khuôn mặt [1].

Trích xuất đặc trưng Haar là phương pháp sử dụng các bộ lọc Haar để phát hiện sự khác biệt về cường độ sáng giữa các vùng trong ảnh. Các bộ lọc này có dạng mẫu hình chữ nhật như hình 1, gồm các vùng sáng và tối với số lượng vùng có thể là 2, 3 hoặc 4, tạo nên sự tương phản rõ ràng. Những mẫu đặc trưng Haar-like được thiết kế để nhận diện các đặc điểm quan trọng trên khuôn mặt, như vùng mắt, mũi hoặc miệng, dựa trên sự chênh lệch cường độ ánh sáng giữa các khu vực.



Hình 1 Minh họa phương pháp haar-like

Ở hình 1 đưa ra ví dụ về các đặc điểm hình chữ nhật được hiển thị liên quan đến cửa sổ phát hiện bao quanh. Tổng các điểm ảnh nằm trong các hình chữ nhật màu trắng được trừ khỏi tổng các điểm ảnh trong các hình chữ nhật màu xám. Các đặc điểm hai hình chữ nhật được hiển thị trong (A) và (B). Hình (C) hiển thị một đặc điểm ba hình chữ nhật và (D) một đặc điểm bốn hình chữ nhật.

2.2. Trích xuất đặc trưng và phát hiện khuôn mặt.

Để tăng tốc độ tính toán, phương pháp tích hợp ảnh (Integral Image) được áp dụng, giúp nhanh chóng tính tổng giá trị pixel trong một vùng bất kỳ theo công thức (1). Tiếp theo, thuật toán AdaBoost được sử dụng để lựa chọn những đặc trưng quan trọng nhất, giúp giảm số lượng đặc trưng cần xét, nâng cao hiệu suất nhận diện. Cuối cùng, quá trình phát hiện theo dạng Cascade được triển khai, áp dụng các tầng bộ lọc tuần tự để loại bỏ dần những vùng ảnh không cần thiết, giúp tăng tốc độ và độ chính xác trong việc nhận diện khuôn mặt sinh viên từ hình ảnh thu thập tại cổng kiểm soát.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (1)$$

Haar Cascade Classifier có ưu điểm là tốc độ xử lý nhanh, có thể nhận diện khuôn mặt trong thời gian thực ngay cả trên phần cứng có hiệu suất thấp. Tuy nhiên, hạn chế của phương pháp này là độ chính xác có thể bị ảnh hưởng bởi ánh sáng, góc nhìn và nhiễu trong ảnh.

Việc kết hợp OpenCVSharp và Haar Cascade Classifier trong hệ thống nhận diện khuôn mặt mang lại nhiều lợi ích. OpenCVSharp cung cấp giao diện lập trình mạnh mẽ cho việc thao tác hình ảnh và triển khai thuật toán, trong khi Haar Cascade Classifier giúp phát hiện khuôn mặt một cách nhanh chóng và hiệu quả. Trong thực tế, hệ thống nhận diện khuôn mặt dựa trên hai công nghệ này có thể được triển khai để kiểm soát an ninh, điểm danh tự động trong các tổ chức, trường học hoặc tích hợp vào các hệ thống nhà thông minh.

Mặc dù Haar Cascade Classifier là một phương pháp hiệu quả để phát hiện khuôn mặt, nhưng để tăng độ chính xác, có thể kết hợp với các phương pháp hiện đại hơn như HOG (Histogram of Oriented Gradients), CNN (Convolutional Neural Networks) hoặc deep learning. Việc tối ưu hóa mô hình và sử dụng các tập dữ liệu đa dạng hơn cũng là một hướng phát triển quan trọng trong tương lai.

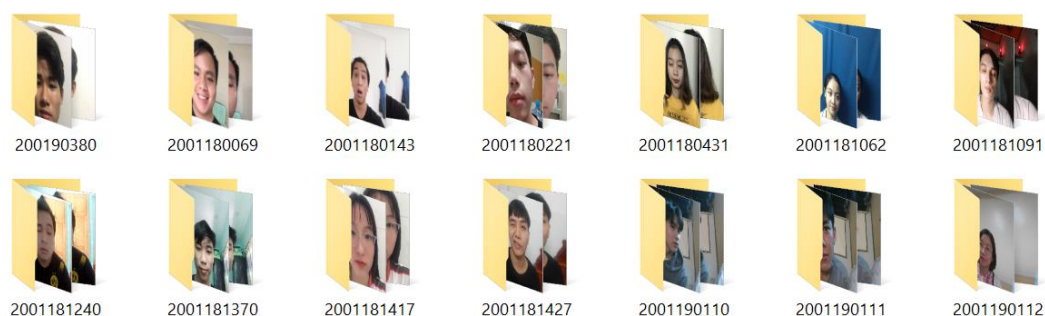
3. PHƯƠNG PHÁP THỰC HIỆN

Trong nghiên cứu này, việc triển khai hệ thống nhận diện khuôn mặt dựa trên OpenCVSharp và Haar Cascade Classifier được tiến hành theo một quy trình khoa học chặt chẽ, nhằm tối ưu hóa khả năng phát hiện và nhận diện khuôn mặt trong các điều kiện thực tế. Quá trình thực hiện bao gồm nhiều giai đoạn liên kết chặt chẽ, từ thu thập dữ liệu, tiền xử lý, phát hiện đến nhận diện và đánh giá hiệu suất. Dưới đây là mô tả chi tiết về từng bước, được trình bày theo phong cách bài báo khoa học.

3.1. Thu thập và tiền xử lý dữ liệu đầu vào

Quá trình thu thập dữ liệu nhằm xây dựng tập ảnh khuôn mặt đa dạng về ánh sáng, góc chụp và biểu cảm, giúp đánh giá hiệu quả phát hiện của thuật toán Haar Cascade trong thực tế. Sau khi thu thập, ảnh được tiền xử lý bằng cách chuyển sang thang độ xám để loại

bỏ thông tin màu, giảm nhiễu và tăng tốc xử lý. Đồng thời, ảnh được chuẩn hóa kích thước và cân bằng sáng bằng Histogram Equalization để đảm bảo trích xuất đặc trưng khuôn mặt nhất quán và hiệu quả.



Hình 2. Minh họa bộ dữ liệu ảnh sinh viên

Sau khi xử lý dữ liệu, bước tiếp theo là phát hiện khuôn mặt bằng cách áp dụng thuật toán Haar Cascade Classifier. Thuật toán này, dựa trên nguyên lý trích xuất đặc trưng Haar và sử dụng cấu trúc tích hợp (Integral Image) để tính toán nhanh chóng các tổng pixel, đã được huấn luyện trước trên một tập dữ liệu lớn gồm các ảnh có và không có khuôn mặt. Trong quá trình quét ảnh, hệ thống áp dụng một loạt các cửa sổ với kích thước khác nhau để tìm kiếm các đặc trưng phù hợp, sau đó sử dụng cơ chế cascade để loại bỏ các cửa sổ không có khả năng chứa khuôn mặt, qua đó tối ưu hóa tốc độ xử lý. Mỗi khuôn mặt được phát hiện sẽ được đánh dấu và cắt ra để phục vụ cho quá trình nhận diện sau này.

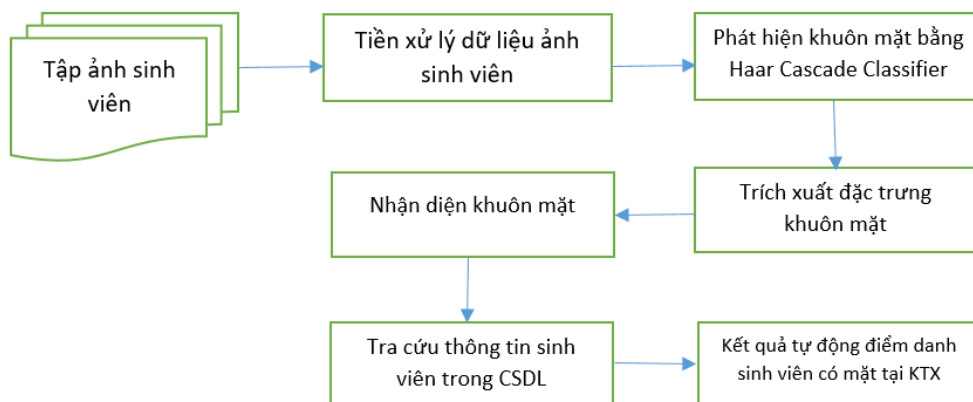
Tiếp theo, sau khi đã xác định được vị trí khuôn mặt trong ảnh, hệ thống tiến hành trích xuất đặc trưng khuôn mặt. Giai đoạn này nhằm mục đích chuyển đổi thông tin hình ảnh thành các vector đặc trưng có thể so sánh được. Các vector đặc trưng này chứa đựng những thông tin quan trọng về hình dạng, cấu trúc của khuôn mặt và là đầu vào cho giai đoạn nhận diện tiếp theo. Trong trường hợp này, ngoài việc sử dụng trực tiếp các đặc trưng từ bộ lọc Haar, hệ thống còn có thể kết hợp các phương pháp truyền thống như LBPH nhằm tăng cường khả năng phân biệt giữa các khuôn mặt khác nhau, đặc biệt trong trường hợp dữ liệu đầu vào có sự biến đổi về điều kiện ánh sáng hoặc góc nhìn.

Giai đoạn cuối cùng của quy trình là nhận diện khuôn mặt dựa trên các vector đặc trưng đã trích xuất. Tại bước này, hệ thống so sánh các vector từ ảnh đầu vào với cơ sở dữ liệu đã được lưu trữ, từ đó xác định danh tính của đối tượng. Quá trình so sánh được thực hiện thông qua các thuật toán đối sánh mẫu, giúp đánh giá mức độ tương đồng giữa các vector đặc trưng. Kết quả của bước nhận diện được ghi nhận và hiển thị, đồng thời được lưu trữ để phục vụ cho việc giám sát và ra quyết định trong các ứng dụng thực tiễn như điểm danh tự động, kiểm soát truy cập trong ký túc xá.

Bên cạnh đó, hệ thống được xây dựng với khả năng kiểm thử và đánh giá hiệu suất một cách toàn diện. Các chỉ số quan trọng như tốc độ xử lý, độ chính xác trong việc phát hiện và nhận diện khuôn mặt được đo lường dưới nhiều điều kiện khác nhau. Các thí nghiệm cho thấy, mặc dù thuật toán Haar Cascade cho phép phát hiện khuôn mặt nhanh chóng trong điều kiện ảnh chụp trực diện và ánh sáng đầy đủ, nhưng khi điều kiện ánh sáng yếu hoặc khuôn mặt bị che khuất một phần, hiệu quả của hệ thống giảm sút đáng kể. Những phát hiện này đã chỉ ra hướng cải tiến thông qua việc tinh chỉnh bộ dữ liệu huấn luyện và cân nhắc kết hợp các thuật toán deep learning tiên tiến nhằm nâng cao độ chính xác và khả năng chịu đựng của hệ thống trước những biến đổi của môi trường thực tế.

3.2. Phương pháp nhận diện khuôn mặt

Hình minh họa các bước của phương pháp nhận diện khuôn mặt bao gồm: (1)Tiền xử lý dữ liệu: Dùng Haar Cascade quét ảnh, xác định và cắt vùng khuôn mặt. (2) Phát hiện khuôn mặt bằng Haar Cascade Classifier: Dùng Haar Cascade quét ảnh, xác định và cắt vùng khuôn mặt. (3) Trích xuất đặc trưng khuôn mặt: Biến ảnh thành vector đặc trưng, kết hợp Haar-like & LBPH. (4) Nhận diện khuôn mặt: So sánh vector với cơ sở dữ liệu, xác định danh tính. (5) Tra cứu thông tin sinh viên nhận diện được ở bước 4 có trong cơ sở dữ liệu lưu trữ hay không. (6) Tự động chuyển trạng thái điểm danh cho sinh viên là có mặt tại KTX nếu kiểm tra sinh viên đang có trong hệ thống.



Hình 3. Minh họa phương pháp nhận diện khuôn mặt

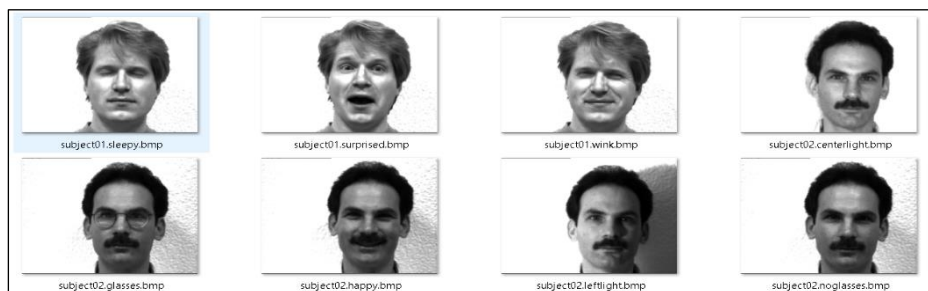
Tóm lại, phương pháp thực hiện được triển khai trong nghiên cứu không chỉ đảm bảo một quy trình xử lý có hệ thống mà còn mở ra khả năng ứng dụng linh hoạt trong các môi trường thực tiễn. Qua đó, các bước từ thu thập dữ liệu, tiền xử lý, phát hiện, trích xuất đặc trưng đến nhận diện được tích hợp một cách chặt chẽ, tạo thành một hệ thống nhận diện khuôn mặt hiệu quả và có tiềm năng mở rộng trong tương lai. Những kết quả thu được từ quá trình thực hiện sẽ là cơ sở để phát triển các hệ thống giám sát và kiểm soát danh tính trong môi trường kỷ túc xá và các ứng dụng an ninh khác.

4. THỰC NGHIỆM

4.1 Dữ liệu thực nghiệm

Trong nghiên cứu này, hệ thống nhận diện khuôn mặt dựa trên OpenCVSharp và Haar Cascade Classifier được triển khai và đánh giá trên hai tập dữ liệu YaleFace và tập ảnh sinh viên HUIT.

Hình 4 minh họa tập dữ liệu YaleFace bao gồm 165 ảnh xám của 15 người khác nhau. Mỗi người có 11 ảnh khác nhau bao gồm các trạng thái: ánh sáng trung tâm, có đeo kính, hạnh phúc, ánh sáng trái, không đeo kính, bình thường, ánh sáng phải, buồn, buồn ngủ, ngạc nhiên và nháy mắt. thực tế đa dạng:



Hình 4. Tập ảnh YaleFace

Tập ảnh sinh viên khoa Công nghệ Thông tin trường Đại học Công thương TP. Hồ Chí Minh gồm 50 bạn mỗi bạn từ 10-20 ảnh với các biểu cảm khác nhau: vui, buồn, hạnh phúc, ngạc nhiên, như hình 5.



Hình 5. Tập ảnh sinh viên HUIT

Ban đầu, để xây dựng bộ dữ liệu, các bức ảnh khuôn mặt được thu thập từ nhiều nguồn khác nhau. Bộ dữ liệu này bao gồm ảnh được chụp dưới các điều kiện ánh sáng khác nhau: từ ánh sáng tự nhiên rực rỡ, ánh sáng nhân tạo ổn định đến những điều kiện ánh sáng yếu, có thể có hiệu ứng ngược sáng hoặc đổ bóng. Các bức ảnh cũng được chụp với nhiều góc nhìn đa dạng – từ ảnh trực diện, nghiêng nhẹ cho đến nghiêng mạnh – nhằm mô phỏng các tình huống mà đối tượng không luôn luôn xuất hiện theo hướng chuẩn. Hơn nữa, để đánh giá mức độ chịu nhiễu của hệ thống, bộ dữ liệu bao gồm các trường hợp khuôn mặt bị che khuất một phần do phụ kiện như khẩu trang, kính mắt hoặc do tự nhiên như tóc che một phần khuôn mặt. Độ phân giải của ảnh cũng được điều chỉnh đa dạng, từ các ảnh có độ phân giải cao cho đến các ảnh có chất lượng thấp, nhằm kiểm tra khả năng xử lý của hệ thống dưới nhiều điều kiện thực tế khác nhau. Việc xây dựng một tập dữ liệu phong phú và đa dạng này nhằm đảm bảo rằng quá trình đánh giá không chỉ mang tính lý thuyết mà còn phản ánh đúng những thách thức gặp phải trong môi trường ứng dụng thực tế như quản lý ký túc xá hay giám sát an ninh.



Hình 6. Bộ dữ liệu đã được tiền xử lý

Sau khi bộ dữ liệu được thu thập, tất cả các ảnh đầu vào trải qua quá trình tiền xử lý nghiêm ngặt. Trước tiên, các bức ảnh được chuyển đổi sang thang độ xám để loại bỏ thông tin màu không cần thiết, giúp giảm thiểu nhiễu và giảm tải cho các thuật toán xử lý. Việc chuyển đổi này cũng góp phần chuẩn hóa dữ liệu, làm cho quá trình phát hiện khuôn mặt trở nên nhất quán hơn. Tiếp theo, các ảnh được resize về kích thước đồng nhất, một bước quan trọng giúp đảm bảo rằng các cửa sổ quét trong thuật toán Haar Cascade có thể hoạt động hiệu quả trên tất cả các ảnh. Ngoài ra, các kỹ thuật xử lý như Histogram Equalization được áp dụng nhằm cân bằng độ sáng và tương phản, làm nổi bật các đặc trưng quan trọng của khuôn mặt, từ đó tạo điều kiện thuận lợi cho giai đoạn phát hiện.

Trong giai đoạn phát hiện, hệ thống sử dụng Haar Cascade Classifier đã được huấn luyện trên một tập dữ liệu lớn để nhận diện các đặc trưng khuôn mặt. Thuật toán này hoạt động bằng cách quét từng vùng nhỏ của ảnh với nhiều kích thước cửa sổ khác nhau, nhằm phát hiện các đặc trưng đặc trưng của khuôn mặt, như vùng mắt, mũi và miệng. Quá trình này được tăng tốc thông qua cơ chế Integral Image, cho phép tính toán nhanh tổng số điểm ảnh trong mỗi cửa sổ với độ phức tạp tính toán không phụ thuộc vào kích thước cửa sổ. Sau khi các cửa sổ có khả năng chứa khuôn mặt được xác định, kỹ thuật Non-Maximum Suppression (NMS) được áp dụng để loại bỏ các phát hiện trùng lặp, đảm bảo mỗi khuôn mặt chỉ được xác định một lần. Giai đoạn này không chỉ chứng minh tính hiệu quả của thuật toán trong việc phát hiện khuôn mặt mà còn cho thấy tốc độ xử lý ấn tượng, với thời gian phát hiện dao động từ 0.05 đến 0.1 giây trong điều kiện ánh sáng đầy đủ.

4.2 Kết quả thực nghiệm

Thực nghiệm được thực thi trên máy AMD Ryzen 3 2200G with Radeon Vega Graphics 3.50 GHz, 8.00 GB RAM, hệ điều hành Windows 10 Pro 64 bit và sử dụng ngôn ngữ lập trình python, C# và .NET Framework 4.6.

Bảng 1 trình bày hiệu suất của phương pháp đề xuất trên bộ ảnh YaleFace và trên bộ ảnh sinh viên HUIT với tổng số ảnh lần lượt là 165 và 980 (85% dùng để huấn luyện, 15% dùng để kiểm thử). Thực nghiệm trên bộ ảnh chuẩn Yale Face cho kết quả là 93.33%. Đối với bộ ảnh sinh viên Huit độ chính xác đạt 90.67% thấp hơn do chất lượng ảnh thu thập được.

Bảng 1. Kết quả thực nghiệm

Tập dữ liệu ảnh	Tổng số ảnh	Số ảnh huấn luyện	Số ảnh kiểm thử	Độ chính xác trung bình
Yale Face	165	135	30	93.33%
Sinh viên HUIT	980	833	147	90.67%

Sau khi xác định vị trí khuôn mặt, các vùng chứa khuôn mặt được cắt ra để trích xuất đặc trưng. Việc trích xuất này chuyển đổi thông tin hình ảnh thành các vector đặc trưng, giúp so sánh và xác định danh tính của đối tượng. Trong nghiên cứu này, mặc dù tập trung vào việc phát hiện khuôn mặt, các vector đặc trưng được trích xuất sau đó cũng được sử dụng trong các thuật toán đối sánh truyền thống như LBPH nhằm đánh giá độ tương đồng giữa ảnh đầu vào và dữ liệu mẫu có sẵn trong cơ sở dữ liệu. Giai đoạn nhận diện dựa trên so sánh vector đặc trưng cho thấy rằng, đối với các ảnh chụp trực diện với điều kiện ánh sáng đầy đủ, hệ thống đạt độ chính xác nhận diện trên 90.67%. Tuy nhiên, khi đối mặt với các ảnh có góc nhìn không chuẩn hoặc bị che khuất, độ chính xác thấp hơn trong điều kiện ánh sáng yếu, khi mà nhiễu ảnh làm tăng tỷ lệ dương tính giả.

Thêm vào đó, quá trình đo lường hiệu suất của hệ thống được thực hiện một cách kỹ lưỡng. Thời gian xử lý của từng bước được ghi nhận bằng mili-giây, từ đó đánh giá tốc độ phản hồi của hệ thống. Các chỉ số đánh giá như độ chính xác, tỷ lệ dương tính giả và tỷ lệ âm tính giả được tính toán để cung cấp một bức tranh toàn diện về hiệu suất hoạt động của hệ thống dưới các điều kiện thử nghiệm khác nhau. Qua đó, các yếu tố ảnh hưởng như điều kiện ánh sáng, góc nhìn và mức độ che khuất được phân tích chi tiết, giúp rút ra những nhận định rõ ràng về điểm mạnh và hạn chế của hệ thống.

Phân tích kết quả cho thấy, ánh sáng là yếu tố quan trọng nhất ảnh hưởng đến chất lượng nhận diện. Trong điều kiện ánh sáng đầy đủ, hệ thống hoạt động với hiệu suất cao, trong khi ánh sáng yếu làm giảm đáng kể độ tương phản, dẫn đến tăng tỷ lệ nhiễu và sai lệch trong phát hiện. Bên cạnh đó, các bức ảnh có góc nhìn lệch hoặc bị che khuất một phần cũng gây ra sự giảm sút trong độ chính xác nhận diện, đòi hỏi phải có một bộ dữ liệu huấn luyện phong phú và các kỹ thuật nhận diện hiện đại hơn để bù đắp.

Từ các thí nghiệm trên, hệ thống nhận diện khuôn mặt dựa trên OpenCVSharp và Haar Cascade Classifier đã chứng minh được khả năng phát hiện nhanh chóng và nhận diện chính xác trong điều kiện ánh sáng đầy đủ. Tuy nhiên, trong môi trường thực tế với các biến số phức tạp, vẫn cần có những biện pháp cải tiến thêm để nâng cao độ ổn định và hiệu quả tổng thể. Những kết quả này cung cấp cơ sở vững chắc cho việc phát triển các ứng dụng nhận diện khuôn mặt trong các hệ thống giám sát an ninh, quản lý ký túc xá và kiểm soát truy cập tự động, mở ra hướng nghiên cứu và ứng dụng rộng rãi trong tương lai.

The screenshot displays a web application for student attendance. The interface includes a header, a sidebar with filters, a main content area with a student photo and personal details, and a table of student records.

Điểm Danh Sinh Viên

Phòng:

Ngày Điểm Danh: Friday, March 21, 2025

Thông Tin Sinh Viên

Mã Sinh Viên: 2001245831

Họ Và Tên: Cầm Văn Cư

Thời Gian: 2025-03-22 14:28:23

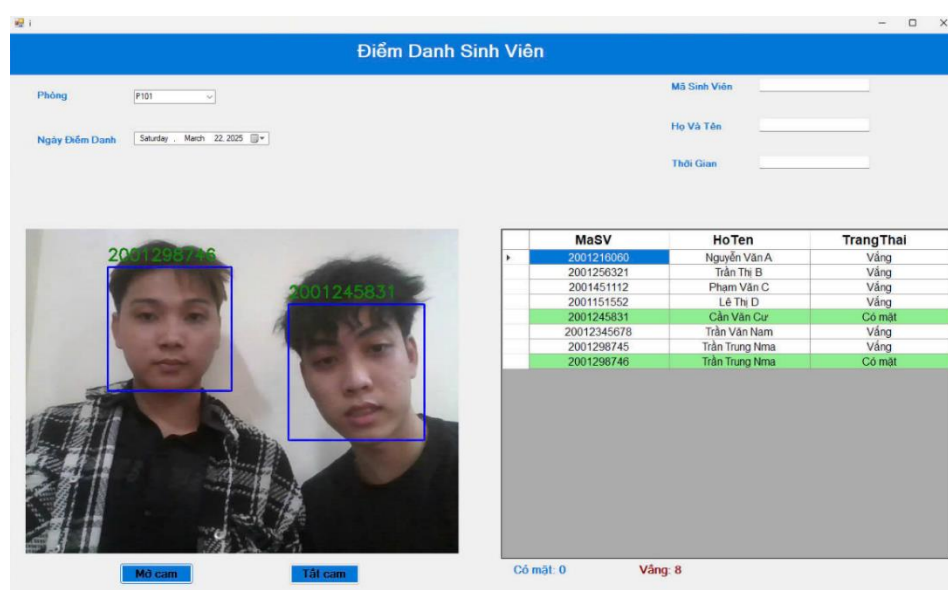
Kiểm Tra Nhận Diện Sinh Viên

2001245831

MaSV	HoTen	TrangThai
2001216060	Nguyễn Văn A	Vắng
2001256321	Trần Thị B	Vắng
2001451112	Phạm Văn C	Vắng
2001151552	Lê Thị D	Vắng
2001245831	Cầm Văn Cư	Có mặt
20012345678	Trần Văn Nam	Vắng
2001298745	Trần Trung Nma	Vắng

Mở cam Tắt cam Có mặt: 2 Vắng: 6

Hình 7. Thực nghiệm nhận diện một đối tượng



Hình 8. Thực nghiệm nhận diện đa đối tượng

5. KẾT LUẬN

Nghiên cứu đã đề xuất giải pháp nhận diện trực tiếp khuôn mặt sinh viên có mặt tại ký túc xá thông qua camera và tích hợp kết quả nhận diện vào ứng dụng quản lý điểm danh sinh viên tại ký túc xá. Khuôn mặt nhận diện được sẽ được so khớp trong cơ sở dữ liệu quản lý và kết quả điểm danh sinh viên có mặt sẽ được tự động cập nhật vào hệ thống. Kết quả thực nghiệm trong điều kiện ánh sáng đầy đủ cho kết quả chính xác 90.67%. Trong trường hợp điều kiện ánh sáng yếu, góc nhìn lệch và khi khuôn mặt bị che khuất một phần thì sẽ ảnh hưởng đến kết quả nhận dạng.

TÀI LIỆU THAM KHẢO

- [1] P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001.
- [2] S. Z. Li and A. K. Jain, Handbook of Face Recognition, Springer, 2011.
- [3] Nicolas Ortiz, et al., Survey of biometric pattern recognition via machine learning techniques, Contemporary Engineering Sciences, vol. 11, no. 34, pp. 1677-1694, 2018.
- [4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations (ICLR), 2015.
- [5] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NIPS), 2012.

ABSTRACT

STUDENT ATTENDANCE SYSTEM USING FACIAL RECOGNITION AT THE DORMITORY OF HO CHI MINH CITY UNIVERSITY OF INDUSTRY AND TRADE

Bui Hung Phuong, Huynh Thanh Trieu, Huynh Vi Khang, Tran Thi Van Anh*

Ho Chi Minh City University of Industry and Trade, Vietnam

Email : anhttv@huit.edu.vn

Received: March 31, 2025; Accepted: May 20, 2025

To enhance student management and ensure security in the dormitory, we propose a facial recognition-based attendance system integrated into the dormitory management application at Ho Chi Minh City University of Industry and Trade. The system utilizes facial recognition techniques with OpenCV and Haar Cascade to enable automatic attendance and access control. Haar Cascade is employed to detect faces quickly in real-time, combined with image preprocessing techniques such as grayscale conversion and brightness normalization. Experiments conducted on the Yale Face dataset and real-world student data from the HUIT dormitory achieved accuracy rates of 93.33% and 90.67%, respectively. This demonstrates the effectiveness of integrating facial recognition into the dormitory management system for automatic student attendance.

Keywords: facial recognition, OpenCV, attendance tracking

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Trần Thị Vân Anh

Số điện thoại liên hệ: 0909188906

XÂY DỰNG MÔ HÌNH PHÂN LOẠI CẢM XÚC DỰA TRÊN HỌC SÂU VÀ ỨNG DỤNG TRONG MẠNG XÃ HỘI ĐỂ PHÁT HIỆN, NGĂN CHẶN BÌNH LUẬN KHÔNG PHÙ HỢP

Hà Trọng Thắng, Hồ Thị Kiều Oanh, Nguyễn Thanh Vi, Nguyễn Huy Hoàng, Nguyễn
Công Dũng, Lê Thị Thùy Lan, Trần Thị Vân Anh*

¹Trường Đại học Công thương Thành phố Hồ Chí Minh

*Email: anhttv@huit.edu.vn

Ngày nhận bài: 31/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Bài báo này trình bày về phương pháp xây dựng mô hình phân loại cảm xúc dựa trên học sâu nhằm phát hiện và ngăn chặn bình luận không phù hợp trên mạng xã hội. Nhóm tác giả sử dụng mô hình PhoBERT, một biến thể của BERT được huấn luyện trên dữ liệu tiếng Việt để trích xuất đặc trưng văn bản và phân loại cảm xúc. Bộ dữ liệu huấn luyện bao gồm gần 17.814 bình luận thu thập từ các nền tảng mạng xã hội, được gán nhãn theo các trạng thái cảm xúc như tích cực, tiêu cực và trung tính. Thử nghiệm cho thấy mô hình PhoBERT đạt độ chính xác trên 90% khi phân loại cảm xúc trong bình luận tiếng Việt. Hệ thống đã được nhóm tác giả tích hợp mô hình vào ứng dụng để thực hiện phân loại cảm xúc vào một nền tảng mạng xã hội để tự động phát hiện và xử lý nội dung không phù hợp, giúp tạo ra môi trường tương tác lành mạnh.

Từ khóa: BERT, bình luận, PhoBERT, phân tích cảm xúc

1. MỞ ĐẦU

Hiện nay, mạng xã hội là nền tảng phổ biến để người dùng chia sẻ quan điểm, cảm xúc và thông tin. Tuy nhiên, sự xuất hiện của quá nhiều bình luận tiêu cực có thể gây ảnh hưởng xấu đến tâm lý cá nhân và môi trường trực tuyến. Các nội dung mang tính công kích, xúc phạm hoặc tiêu cực quá mức có thể dẫn đến những hậu quả nghiêm trọng như bạo lực mạng hoặc ảnh hưởng đến nhận thức xã hội. Vì vậy, việc các bình luận trên mạng xã hội được nhận diện và phân loại một cách tự động nhằm tạo ra một môi trường mạng lành mạnh và văn minh đang là một vấn đề quan trọng cần được quan tâm và xử lý.

Tổng quan về các phương pháp nhận diện cảm xúc nhận diện cảm xúc có thể được thực hiện bằng nhiều phương pháp khác nhau, bao gồm: Nhận diện qua khuôn mặt sử dụng kỹ thuật xử lý ảnh và thị giác máy tính để phân tích biểu cảm khuôn mặt, từ đó xác định trạng thái cảm xúc của người dùng. Phương pháp này thường áp dụng trong các hệ thống nhận diện khuôn mặt thông minh và giám sát hành vi. Nhận diện qua giọng nói phân tích đặc điểm âm thanh như âm sắc, tần số, cao độ và ngữ điệu để nhận diện cảm xúc. Phương pháp này được sử dụng phổ biến trong trợ lý ảo, tổng đài thông minh và phân tích phản hồi khách

hàng. Nhận diện qua văn bản áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để phân tích nội dung văn bản và xác định cảm xúc dựa trên từ vựng, ngữ cảnh và cú pháp. Đây là phương pháp quan trọng trong các hệ thống phân tích bình luận, đánh giá sản phẩm và mạng xã hội. Các phương pháp học sâu như CNN, LSTM, BERT và PhoBERT đã được áp dụng rộng rãi trong các bài toán phân loại cảm xúc. Một số nghiên cứu đã so sánh hiệu quả giữa các phương pháp này: CNN có khả năng trích xuất đặc trưng mạnh mẽ nhưng thiếu thông tin về ngữ cảnh của từ. LSTM có thể giữ được thông tin ngữ cảnh tốt hơn nhưng hiệu suất không cao bằng mô hình Transformer. Nghiên cứu trên tập dữ liệu IMDB cho thấy LSTM đạt độ chính xác 85%, trong khi BERT đạt 92% [7]. So sánh trên dữ liệu bình luận tiếng Việt, SVM đạt 78%, CNN đạt 83% và PhoBERT đạt 96% [8]. Trên tập dữ liệu Facebook bình luận tiêu cực, PhoBERT kết hợp với SVM đạt 97% độ chính xác [9]. BERT và PhoBERT cho kết quả chính xác cao nhất nhờ vào cơ chế tự chú ý (self-attention), giúp hiểu ngữ nghĩa toàn diện hơn. Bài nghiên cứu này đề xuất một mô hình phân loại cảm xúc dựa trên học sâu, sử dụng PhoBERT – một mô hình được tối ưu hóa cho tiếng Việt, nhằm nhận diện và ngăn chặn các bình luận không phù hợp.

Phương pháp nghiên cứu quy trình thực hiện bao gồm các bước chính: Thu thập và tiền xử lý dữ liệu dữ liệu bình luận được thu thập từ các nền tảng mạng xã hội, tiến hành làm sạch dữ liệu như loại bỏ ký tự đặc biệt, chuẩn hóa ngôn ngữ và gán nhãn cảm xúc. Phương pháp trích xuất đặc trưng các phương pháp phổ biến để trích xuất đặc trưng văn bản gồm: CNN trích xuất đặc trưng dựa trên các bộ lọc tích chập, LSTM giữ được ngữ cảnh của từ trong câu, BERT/PhoBERT mô hình transformer mạnh mẽ trong xử lý ngôn ngữ tự nhiên. Lựa chọn mô hình bài nghiên cứu sử dụng PhoBERT do khả năng hiểu ngữ nghĩa ngữ cảnh tốt hơn các phương pháp truyền thống như SVM hoặc LSTM. Huấn luyện và đánh giá mô hình kiểm tra độ chính xác trên tập dữ liệu thử nghiệm, so sánh với các mô hình truyền thống. Ứng dụng thực tế tích hợp mô hình vào hệ thống kiểm duyệt nội dung trên mạng xã hội để tự động phát hiện và xử lý bình luận tiêu cực.

Nhiều nghiên cứu trước đây đã đề xuất các phương pháp khác nhau để phân loại cảm xúc từ văn bản. Một số bài báo tiêu biểu bao gồm: [1] đề xuất mô hình LSTM để xử lý ngôn ngữ tự nhiên, [2] giới thiệu BERT – mô hình tiên tiến trong xử lý ngôn ngữ tự nhiên, [3] nghiên cứu về PhoBERT – phiên bản BERT tối ưu hóa cho tiếng Việt. Trong các phương pháp trên thì phương pháp PhoBERT cho hiệu quả vượt trội trong phân loại cảm xúc văn bản tiếng Việt nhờ khả năng học ngữ cảnh tốt, được huấn luyện trên tập dữ liệu lớn và tối ưu hóa riêng cho tiếng Việt. Do đó, mô hình này được lựa chọn để chạy cho nghiên cứu nhằm đảm bảo độ chính xác cao và phù hợp với đặc thù của ngôn ngữ này. Bên cạnh đó, các mô hình học sâu khác cũng được áp dụng trong phân tích cảm xúc như CNNs (*Convolutional Neural Networks*) [4] – có khả năng trích xuất đặc trưng hiệu quả từ văn bản, hay sự kết hợp giữa BERT và SVM (*Support Vector Machine*) [5] giúp cải thiện độ chính xác trong phân loại cảm xúc. Một số nghiên cứu [6] đã thực hiện so sánh giữa các phương pháp truyền thống như Naïve Bayes, SVM với các mô hình học sâu và nhận thấy rằng BERT và PhoBERT mang lại hiệu suất cao hơn đáng kể. Do đó, trong nghiên cứu này, PhoBERT được lựa chọn làm mô hình chính để đảm bảo độ chính xác cao và phù hợp với đặc thù của tiếng Việt.

Nghiên cứu này đề xuất phương pháp kết hợp PhoBERT với từ điển cảm xúc để cải thiện độ chính xác trong phân loại cảm xúc bình luận tiếng Việt. Từ điển cảm xúc được xây dựng từ các bộ dữ liệu gán nhãn trước đó, bao gồm danh sách các từ/cụm từ mang sắc thái tích cực, tiêu cực và trung tính. Khi một bình luận được phân tích, mô hình PhoBERT trích xuất đặc trưng ngữ nghĩa, trong khi phương pháp từ điển giúp bổ sung thông tin về sắc thái cảm xúc dựa trên các từ khóa xuất hiện trong câu.

Hệ thống được thiết kế để tích hợp vào một nền tảng mạng xã hội với các chức năng chính như đăng tải bài viết, bình luận, tương tác (thích, chia sẻ) và kiểm duyệt nội dung. Mô

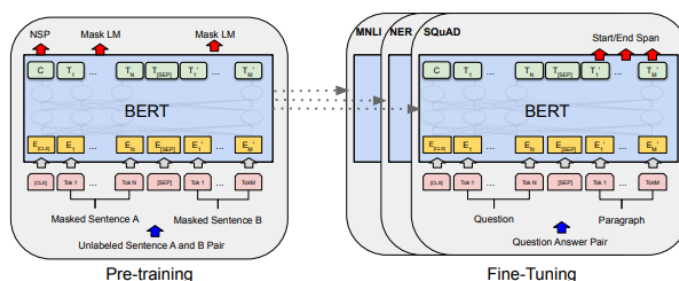
hình PhoBERT kết hợp với từ điển cảm xúc giúp nhận diện nhanh chóng các bình luận tiêu cực và hỗ trợ kiểm duyệt nội dung hiệu quả hơn. PhoBERT được triển khai trên máy chủ và kết nối với ứng dụng mạng xã hội thông qua API. Khi người dùng đăng bình luận, hệ thống sẽ tự động phân tích cảm xúc và đưa ra cảnh báo nếu bình luận có nội dung tiêu cực. Ngoài ra, mô hình có thể được cải thiện theo thời gian bằng cách huấn luyện lại với các dữ liệu mới, giúp nâng cao độ chính xác và khả năng thích ứng với ngôn ngữ tự nhiên trong môi trường mạng xã hội.

Phần tiếp theo của bài nghiên cứu sẽ trình bày như sau: Phần 2 giới thiệu cơ sở lý thuyết về các thuật toán học sâu và phương pháp phân loại cảm xúc, Phần 3 trình bày quy trình thực hiện, bao gồm thu thập dữ liệu, tiền xử lý, huấn luyện và đánh giá mô hình, Phần 4 mô tả kết quả thực nghiệm và so sánh với các phương pháp khác, và cuối cùng là phần kết luận.

2. CƠ SỞ LÝ THUYẾT

2.1. Mô hình BERT

BERT (*Bidirectional Encoder Representations from Transformers*) là một mô hình học sâu tiên tiến trong lĩnh vực NLP, được Google giới thiệu vào năm 2018. Đây là một trong những mô hình đầu tiên sử dụng kiến trúc Transformer theo hướng hai chiều, giúp nó hiểu ngữ cảnh của từ trong văn bản một cách tốt hơn so với các mô hình trước đây. BERT được huấn luyện bằng phương pháp học có giám sát tự động (*self-supervised learning*), tức là mô hình học từ dữ liệu mà không cần gán nhãn thủ công. Quá trình huấn luyện của BERT bao gồm hai giai đoạn như hình 1: Giai đoạn đầu tiên huấn luyện trước (*Pre-training*) mô hình được đào tạo trên một lượng lớn dữ liệu văn bản tổng quát để học các đặc trưng ngôn ngữ cơ bản, hai nhiệm vụ chính trong giai đoạn này là Masked Language Model (MLM) và Next Sentence Prediction (NSP). Giai đoạn thứ 2 tinh chỉnh (*Fine-tuning*) sau khi được huấn luyện trước, BERT có thể được điều chỉnh trên các tập dữ liệu cụ thể cho các tác vụ như phân loại văn bản, trả lời câu hỏi, trích xuất thực thể, v.v. Mô hình chỉ cần một vài epoch huấn luyện để đạt hiệu suất cao trên nhiều bài toán NLP khác nhau.



Hình 1. Quy trình tổng thể trước khi đào tạo và điều chỉnh cho BERT[10]

BERT được xây dựng dựa trên Transformer, một kiến trúc mạng nơ-ron bao gồm nhiều encoder xếp chồng lên nhau. Các encoder này giúp mô hình học được mối quan hệ giữa các từ trong câu thông qua cơ chế tự chú ý (*self-attention*). Multi-Head Self-Attention cho phép mô hình tập trung vào nhiều phần khác nhau của câu cùng một lúc. Position Embeddings do Transformer không có cơ chế tuần tự như RNN, vị trí của từ trong câu được mã hóa bằng vector embedding. Feedforward Layers các lớp kết nối dày đặc (*fully connected*) giúp trích xuất đặc trưng ngữ nghĩa từ dữ liệu văn bản.

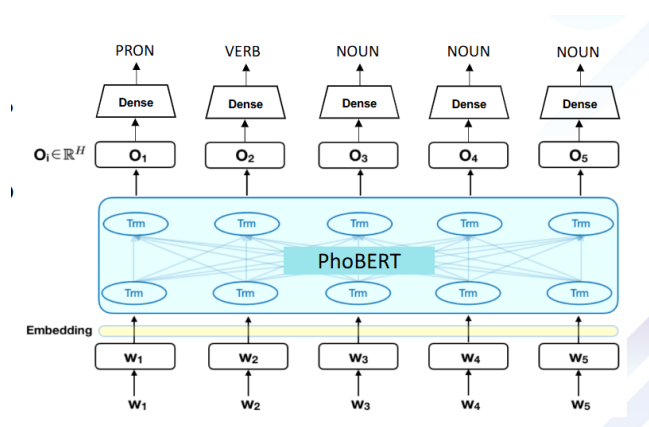
BERT có hai phiên bản chính: BERT Base: 12 lớp mã hóa (encoder layers), 768 nút ẩn (hidden units), 12 nút tập trung (attention heads), (~110 triệu tham số). BERT Large: 24 lớp mã hóa (encoder layers), 1024 nút ẩn (hidden units), 16 nút tập trung (attention heads), (~340 triệu tham số).

2.2 Mô hình PhoBERT

PhoBERT (Vietnamese BERT) là một mô hình xử lý ngôn ngữ tự nhiên (NLP) dành riêng cho tiếng Việt, được phát triển bởi nhóm nghiên cứu của Viện John von Neumann (JVN) và VinAI Research. Mô hình này dựa trên kiến trúc RoBERTa – một phiên bản cải tiến của BERT, giúp cải thiện khả năng hiểu ngữ cảnh và xử lý ngôn ngữ tiếng Việt hiệu quả hơn[8].

PhoBERT được huấn luyện trên một tập dữ liệu văn bản lớn (~20GB), thu thập từ Wikipedia, báo chí và các nguồn tài liệu tiếng Việt khác, giúp nó vượt trội hơn so với các mô hình đa ngôn ngữ như mBERT (Multilingual BERT) khi áp dụng cho tiếng Việt. PhoBERT có kiến trúc tương tự BERT nhưng được tối ưu hóa để xử lý tiếng Việt hiệu quả hơn, dựa trên Transformer Encoder với hai phiên bản: (1) PhoBERT-base: 12 lớp, 768 chiều vector, 117M tham số; (2) PhoBERT-large: 24 lớp, 1024 chiều vector, 370M tham số. PhoBERT được đánh giá trên 5 tác vụ xử lý ngôn ngữ tự nhiên (NLP) tiếng Việt, bao gồm:

Part-of-Speech (POS) Tagging: Xác định loại từ (danh từ, động từ, đại từ, v.v.) trong một câu với bộ dữ liệu sử dụng tập dữ liệu chuẩn VLSP 2013. Phương pháp sử dụng PhoBERT để trích xuất đặc trưng của câu và thêm một lớp dự đoán tuyến tính (Dense Layer) để gán nhãn loại từ cho từng từ trong câu.



Hình 2. Quá trình Part-of-Speech (POS) Tagging[11]

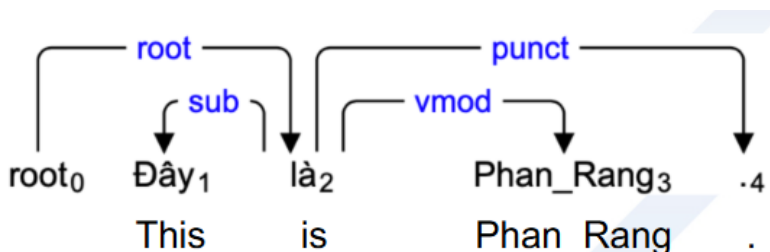
Mô hình bao gồm ba thành phần chính: Lớp Embedding các từ đầu vào ($W_1, W_2, \dots, W_5$) được ánh xạ thành các vector nhúng (embedding) để biểu diễn đặc trưng ngữ nghĩa. Mô hình PhoBERT các vector nhúng được đưa vào mô hình PhoBERT, gồm nhiều lớp Transformer (Trm), để học và trích xuất các đặc trưng ngữ cảnh. Lớp phân loại đầu ra từ PhoBERT ($O_1, O_2, \dots, O_5$) đi qua các lớp Dense để thực hiện phân loại từ loại, xác định các từ là đại từ (PRON), động từ (VERB) hay danh từ (NOUN). Mô hình này giúp hiểu ngữ cảnh và phân loại từ loại một cách chính xác dựa trên bối cảnh của câu.

Named Entity Recognition (NER): Xác định và phân loại các thực thể trong câu như tên người, địa danh, tổ chức với bộ dữ liệu sử dụng tập dữ liệu VLSP 2016 (Nguyen et al.,

2019). Phương pháp PhoBERT được sử dụng để trích xuất đặc trưng văn bản và thêm lớp dự đoán tuyến tính (Dense Layer) để gán nhãn thực thể.

Lớp Embedding các từ đầu vào ($W_1, W_2, \dots, W_5$) được chuyển thành các vector nhúng (embedding) để biểu diễn đặc trưng. Mô hình PhoBERT các vector nhúng được đưa vào mô hình PhoBERT, sử dụng nhiều lớp Transformer (Trm) để trích xuất thông tin ngữ cảnh. Lớp Phân loại đầu ra từ PhoBERT ($O_1, O_2, \dots, O_5$) được đưa qua các lớp Dense để phân loại. Các nhãn đầu ra có thể là các thực thể như tổ chức (B-ORG, I-ORG), hoặc không phải thực thể (O), "B-ORG" (*Beginning of Organization*) đánh dấu từ đầu tiên của một tổ chức, "I-ORG" (*Inside of Organization*) đánh dấu các từ tiếp theo thuộc tổ chức đó. Mô hình này giúp nhận diện và phân loại thực thể trong văn bản dựa vào ngữ cảnh của câu.

Dependency Parsing: Xác định mối quan hệ cú pháp giữa các từ trong câu với bộ dữ liệu: VLSP 2013 Dependency Parsing task. Phương pháp PhoBERT được tinh chỉnh để dự đoán quan hệ giữa các từ.



Hình 3. Quá trình Dependency Parsing[11]

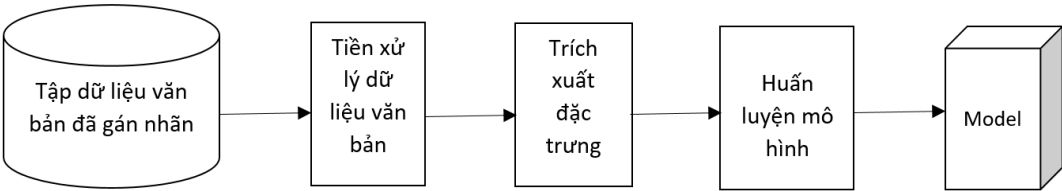
Các bước chính của quá trình Dependency Parsing: Tokenization (tách từ) chia câu thành các từ riêng lẻ (tokens). Part-of-Speech (POS) Tagging (gán nhãn từ loại) xác định loại từ của mỗi token (danh từ, động từ, tính từ, v.v.). Xây dựng quan hệ phụ thuộc xác định từ gốc (root) của câu và thiết lập các quan hệ phụ thuộc giữa các từ. Tạo cây phụ thuộc (Dependency Tree) hệ thống tổ chức các từ theo cấu trúc cây, trong đó mỗi từ có một nút (node) và các nhánh thể hiện mối quan hệ phụ thuộc. Mô hình này giúp máy hiểu câu theo cách con người phân tích ngữ pháp, từ đó có thể xử lý và suy luận dựa trên văn bản một cách hiệu quả hơn.

Natural Language Inference (NLI): Xác định mối quan hệ giữa hai câu (giả định, trung lập, mâu thuẫn) với bộ dữ liệu: ViNLI dataset. Phương pháp sử dụng PhoBERT để trích xuất đặc trưng ngữ nghĩa và một lớp phân loại trên đầu ra của mô hình.

Text-to-SQL semantic parsing: Chuyển đổi các câu truy vấn ngôn ngữ tự nhiên thành biểu diễn có ý nghĩa dưới dạng truy vấn SQL tiêu chuẩn. Truy cập thông tin trong cơ sở dữ liệu thông qua các câu lệnh ngôn ngữ tự nhiên. Người dùng không cần hiểu cú pháp SQL hoặc cấu trúc của cơ sở dữ liệu. Tạo bộ dữ liệu Text-to-SQL quy mô lớn cho bài toán phân tích cú pháp ngữ nghĩa tiếng Việt: Khoảng 10.000 cặp câu hỏi và truy vấn SQL trên hơn 200 cơ sở dữ liệu.

3. PHƯƠNG PHÁP THỰC HIỆN

Mô hình phân loại cảm xúc văn bản bao gồm các bước: tiền xử lý dữ liệu, trích xuất đặc trưng văn bản, sau đó các đặc trưng này được huấn luyện để xây dựng mô hình phân loại cảm xúc như hình 4.



Hình 4. Minh họa mô hình phân loại cảm xúc

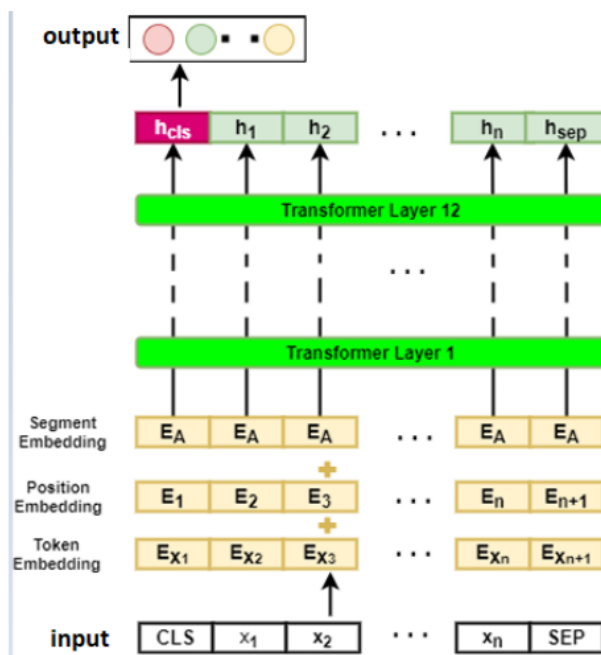
3.1. Tiền xử lí dữ liệu văn bản với PhoBERT

Tập dữ liệu thu thập được gồm 17.814 bình luận thu thập từ các nền tảng mạng xã hội, được gán nhãn theo các trạng thái cảm xúc như tích cực (0) , tiêu cực (1) và trung tính (2). Tập dữ liệu sẽ được tiền xử lí qua các bước sau: Loại bỏ giá trị Null: Các giá trị bằng không và các điểm dữ liệu không liên quan được loại bỏ để làm sạch bộ dữ liệu. Chuẩn hóa văn bản: loại bỏ dấu chấm câu, các ký tự đặc biệt, chuyển đổi văn bản thành chữ thường để đảm bảo tính đồng nhất. Tách từ theo ngữ cảnh tiếng Việt: Văn bản được phân đoạn thành các câu hoặc mã thông báo để tạo điều kiện phân tích. Bước này rất cần thiết để chia nhỏ văn bản thành các phần có thể quản lý được. Do sự khác biệt về âm tiết tiếng Việt, PhoBERT sử dụng Công cụ phân đoạn từ tiếng Việt từ VNCoreNLP

STT	Bình luận	Phân loại cảm xúc
1.	Sang lớp thầy Tiến Chung lấy lại gốc cả rề nè bạn	Tích cực
2.	Bạn đóng rồi thì cứ yên tâm, từ từ công nợ sẽ được cập nhật sau, còn ko thì bạn cầm Bill lên phòng kế hoạch tài chính nhờ các cô check và trừ công nợ cho	Tích cực
3.	Đang bình thường, đọc xong cái tus muốn cộc.	Tiêu cực
4.	Nói chuyện như kiểu dân anh chị như m ai dám ở chung	Tiêu cực
5.	Nhất trai làng anh công tử IT, Nhất gái quê em cô út English	Trung lập

3.2. Xây dựng mô hình phân loại cảm xúc

Trong giai đoạn này, nhóm nghiên cứu áp dụng mô hình Transformer đã được huấn luyện trước. Cơ chế attention (cơ chế chú ý) của Transformer cho phép xử lý đồng thời tất cả các từ trong một câu, bất kể thứ tự của chúng. Do đó, Transformer được coi là một mô hình hai chiều, mặc dù mô tả chính xác hơn thì nó là mô hình phi hướng. Đặc điểm này cho phép mô hình học được ngữ cảnh của một từ dựa trên tất cả các từ xung quanh, bao gồm cả những từ ở bên trái và bên phải. Kiến trúc mô hình gồm 3 phần như hình 5.



Hình 5. Kiến trúc mô hình PhoBert

Đầu vào: được tiền xử lý thành chuỗi ký tự X , bao gồm n token, ký hiệu là $X = \{x_1, x_2, \dots, x_n\}$, trong đó x_i biểu thị token ở vị trí thứ i trong chuỗi đầu vào. Các token này được phân chia vào từ vựng và mã hóa bằng số thông qua từ điển từ vựng (các Lớp Embedding) và tích hợp thông tin về vị trí của từng token như một thành phần đầu vào. Độ dài tối đa của chuỗi đầu vào cho các câu trong tập huấn luyện được xác định bằng cách lựa chọn chuỗi có độ dài lớn nhất. Các câu có độ dài ngắn hơn sẽ được tự động đệm bằng cách sử dụng token đệm đặc biệt.

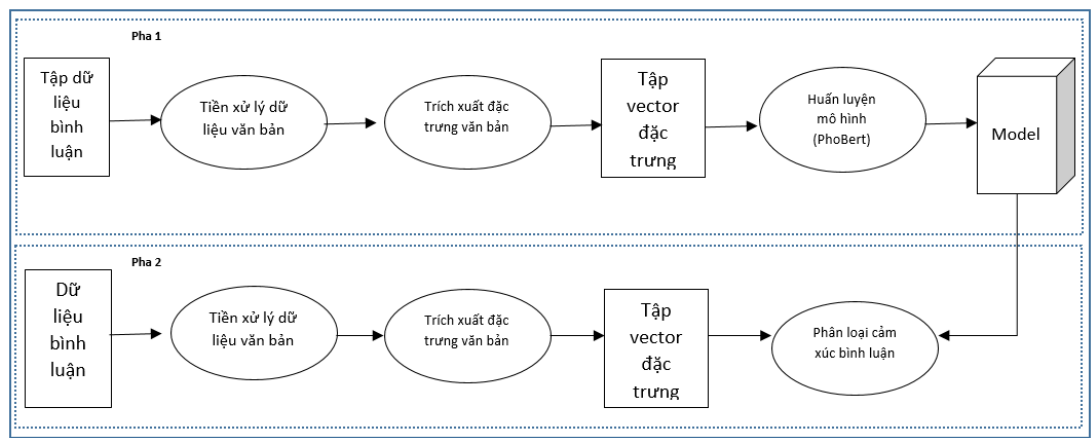
PhoBERT encoding: Nghiên cứu này sử dụng kiến trúc PhoBERT_Base, bao gồm 12 khối Transformer và 12 lớp self-attention tính toán mức độ quan trọng của mỗi từ trong câu với tất cả các từ khác để trích xuất đặc trưng từ các chuỗi đầu vào với độ dài tối đa 512 tokens. Kết quả của mô hình này là một lớp ẩn $H = \{h_1, h_2, \dots, h_n\}$, tương ứng với chuỗi đầu vào. Để phân loại các chuỗi đầu vào dựa trên nội dung cảm xúc của chúng, nhóm nghiên cứu thu được biểu diễn đặc trưng từ token $[CLS]$. Vector đặc trưng này biểu diễn toàn bộ chuỗi đầu vào.

Đầu ra: Sử dụng biểu diễn vector của chuỗi đầu vào, nhóm nghiên cứu triển khai một bộ phân loại với hàm kích hoạt softmax để tính toán các giá trị phân phối xác suất cho mỗi nhãn phân loại theo các hạng mục cảm xúc.

4. THỰC NGHIỆM

Mô hình thực nghiệm Hình 6 gồm 2 pha: pha 1 huấn luyện mô hình phân loại cảm xúc, pha 2 phân loại cảm xúc bình luận trên mô hình đã được huấn luyện ở pha 1

4.1. Mô hình thực nghiệm



Hình 6. Mô hình phân loại cảm xúc bằng văn bản

Mô hình phân loại cảm xúc dựa trên bình luận gồm có 2 pha chính. Pha 1 dùng để huấn luyện mô hình phân loại cảm xúc gồm các bước: (1) Tiền xử lý dữ liệu bình luận văn bản, (2) Trích xuất đặc trưng của văn bản bình luận, (3) Huấn luyện mô hình phân loại dựa trên PhoBERT để phân loại bình luận đầu vào. Pha 2 dùng để phân loại bình luận là tích cực, tiêu cực hoặc trung lập gồm các bước: (1) Tiền xử lý dữ liệu bình luận cần phân loại, (2) Trích xuất đặc trưng văn bản, (3) Sử dụng model phân loại PhoBERT đã huấn luyện để phân loại bình luận đầu vào là tích cực, tiêu cực hoặc trung lập.

Để đo hiệu suất của phương pháp đề xuất, nghiên cứu sử dụng các độ đo gồm: độ chính xác (precision), độ phủ (recall) và độ trung hòa (F1-score) theo các công thức (1)

$$precision = \frac{\sum_{C=1}^C TP}{\sum_{C=1}^C (TP+FP)} \quad recall = \frac{\sum_{C=1}^C TP}{\sum_{C=1}^C (TP+FN)} \quad F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

Trong đó, TP (True Positive): số ảnh thuộc lớp positive được phân loại đúng là positive; TN (True Negative): số ảnh thuộc lớp negative được phân loại đúng là negative; FP (False Posivite): số ảnh thuộc lớp negative bị phân loại nhầm là positive; FN (False Negative): số ảnh thuộc lớp positive bị phân loại nhầm là negative; C là số lớp của tập dữ liệu ảnh.

4.2. Bộ dữ liệu thực nghiệm

Nghiên cứu thực nghiệm trên bộ dữ liệu gồm 11.400 bình luận (trong đó gồm 13.155 bình luận tích cực, 3676 bình luận trung lập, 983 bình luận tiêu cực) được thu thập từ các bình luận trên các trang mạng xã hội và được gán nhãn theo các tiêu chí trong bảng 1:

Bảng 1. Hiệu suất của phương pháp được đề xuất trên các bộ dữ liệu thực nghiệm.

Nhãn	Tiêu chí	Ví dụ minh họa
Trung lập (1)	Không có cảm xúc rõ ràng, thông tin khách quan, không phản đối hay ủng hộ	Cần thêm thời gian để đánh giá rõ hơn Hôm nay trời mưa, nhưng không quá ẩm ướt.
Tích cực (0)	Thể hiện sự ủng hộ, đồng	Sự kiện này thật tuyệt vời! Mọi thứ đều hoàn

Nhãn	Tiêu chí	Ví dụ minh họa
	tình, hoặc khen ngợi.	hào và tôi đã học được rất nhiều. Cảm ơn vì dịch vụ tuyệt vời, tôi sẽ quay lại lần sau.
Tiêu cực (2)	Phản ánh sự không hài lòng, chỉ trích, hoặc phê phán.	Thật thất vọng về cách tổ chức sự kiện này. Chắc chắn sẽ không tham gia lần sau! Dịch vụ khách hàng quá tệ, tôi không hài lòng với cách giải quyết vấn đề.

4.3. Kết quả thực nghiệm

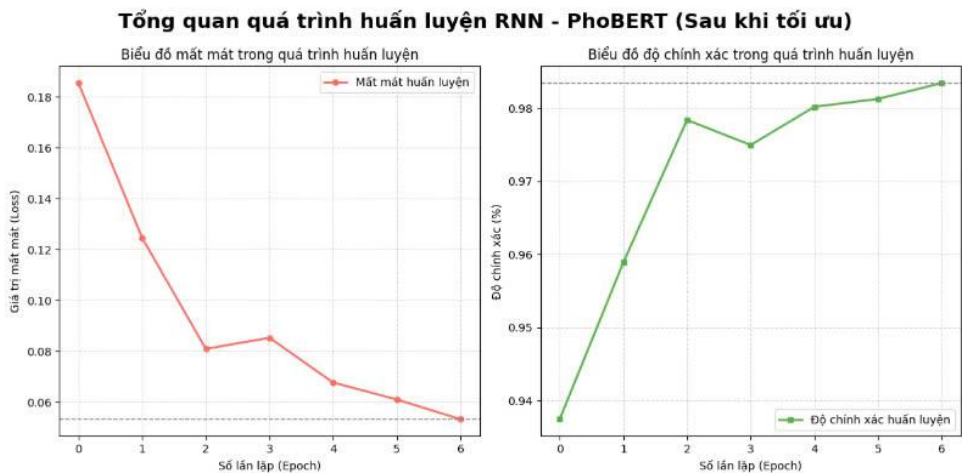
Thực nghiệm được thực thi trên máy PC CPU Intel Core i5 - 1035G1 CPU Intel Xeon 2.3GHz, 30.0GB RAM, hệ điều hành Windows 10 Home 64 bit và sử dụng GPU T4x2, bộ nhớ GPU 16GB x 2, sử dụng ngôn ngữ lập trình python, C# và .NET Framework 4.6.

Bảng 2 trình bày hiệu suất của phương pháp đề xuất trên bộ dữ liệu bình luận gồm 17814 bình luận được gán nhãn tiêu cực, tích cực và trung lập với tỷ lệ 64% dùng để huấn luyện, 20% dùng để kiểm thử và 16% dùng để đánh giá). Thực nghiệm cho kết quả độ chính xác 90%, độ phủ và độ trung hòa đạt ở ngưỡng cao là 86% và 91%.

Bảng 2. Hiệu suất của phương pháp được đề xuất trên các bộ dữ liệu thực nghiệm.

Tổng số bình luận	Số nhãn train	Số nhãn validation	Số nhãn test	Độ chính xác trung bình (precision)	Độ phủ trung bình (recall)	Độ trung hòa trung bình (f1-score)
17814	11400	2850	3564	90%	86%	91%

Sau khi sử dụng hàm `train_model_optimized` để huấn luyện mô hình bằng cách tối ưu hóa lại và kết hợp với từ điển cảm xúc để đưa ra được kết quả như hình 7,8.



Hình 7. Kết quả sau khi tối ưu và kết hợp với dict

```

Nhập câu cần phân tích cảm xúc (hoặc gõ 'exit' để thoát):
> đấm m h
Dự đoán cảm xúc: [NEGATIVE]

Nhập câu cần phân tích cảm xúc (hoặc gõ 'exit' để thoát):
> t đánh nhau vs m lun
Dự đoán cảm xúc: [NEGATIVE]

Nhập câu cần phân tích cảm xúc (hoặc gõ 'exit' để thoát):
> thấy A ok ấ mn
Dự đoán cảm xúc: [POSITIVE]

Nhập câu cần phân tích cảm xúc (hoặc gõ 'exit' để thoát):
> mn cho hỏi cái
Dự đoán cảm xúc: [POSITIVE]

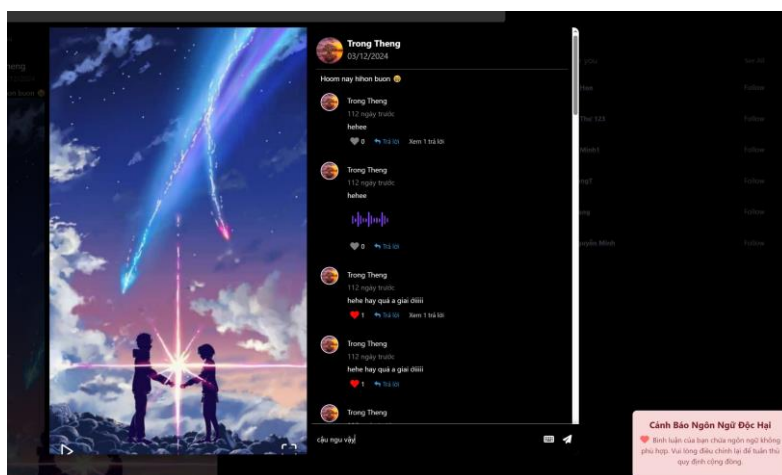
Nhập câu cần phân tích cảm xúc (hoặc gõ 'exit' để thoát):
> làm sao v
Dự đoán cảm xúc: [NEUTRAL]

```

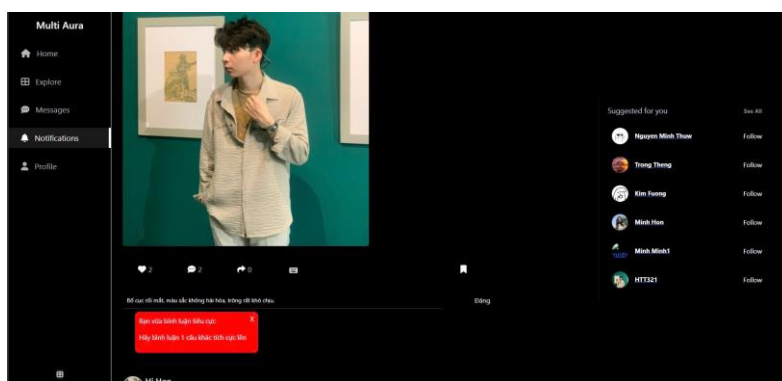
Hình 8. Kết quả phân loại bình luận

4.4. Hệ thống ngăn chặn bình luận tiêu cực

Khi mô hình nhận diện được dữ liệu bình luận từ một bài post mang tính tiêu cực thì hệ thống sẽ ngăn chặn lại trước khi đưa vào hệ thống bằng cách hiển thị thông báo cho người dùng biết là bình luận đó là bình luận tiêu cực thể hiện ở hình 9, 10.



Hình 9. Minh họa chức năng tự động kiểm duyệt bình luận trên bài đăng



Hình 10. Phát hiện và ngăn chặn khi đăng bình luận tiêu cực

5. KẾT LUẬN

Qua kết quả nghiên cứu cho thấy, việc sử dụng mô hình PhoBERT kết hợp với phương pháp từ điển cảm xúc (dict) để nhận diện bình luận của người dùng và ngăn chặn các bình luận tiêu cực ngay khi chúng xuất hiện mang lại hiệu quả cao. Cụ thể, mô hình đạt độ chính xác 90% trên tập kiểm thử, cao hơn so với các phương pháp truyền thống như LSTM hay Naive Bayes. Hệ thống nhận diện văn bản và phân loại cảm xúc trên câu bình luận một cách tự động. Đối với những bình luận tiêu cực, hệ thống sẽ tự động ngăn chặn không cho phép hiển thị trên ứng dụng.

Trong bước kế tiếp, nghiên cứu tiếp tục tăng cường dữ liệu về văn bản tiêu cực, tích cực, trung lập để áp dụng vào hệ thống nhận diện bình luận một cách tốt hơn và nguyên cứu để xử lý các bình luận NaN sau khi tiền xử lý dữ liệu để tránh ảnh hưởng tới mô hình.

TÀI LIỆU THAM KHẢO

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.
3. Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042.
4. Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 616–623.
5. Nguyen, T. H., Vu, X. T., & Pham, T. M. (2021). BERT, RoBERTa, PhoBERT, BERTweet: State-of-the-art pre-trained models for text classification. *Viblo Journal of Artificial Intelligence and Data Science*, 12(4), 45–58.
6. Le, Q. H., Tran, D. K., & Hoang, N. M. (2022). Implementing a sentiment classification model for Vietnamese text using deep learning. *Viblo AI Conference Proceedings*, 15(2), 67–79.
7. Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv preprint arXiv:2009.05451*.
8. Nguyen, H. Q., & Nguyen, Q. U. (2018). An ensemble of shallow and deep learning algorithms for Vietnamese sentiment analysis. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 171-176). IEEE.
9. Nguyen, H. Q., & Nguyen, Q. U. (2018). An ensemble of shallow and deep learning algorithms for Vietnamese sentiment analysis. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 171-176). IEEE.
10. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT, 2019*.
11. Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained Language Models for Vietnamese. *Findings of EMNLP, 2020*.
12. Zhang, R., He, H., Lin, W., & He, J. (2020). Syntax-Enhanced BERT for Text-to-SQL Generation. *Proceedings of ACL, 2020*.
13. Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2021). A Pilot Study on Text-to-SQL Semantic Parsing for Vietnamese. *Proceedings of RIVF, 2021*.

. ABSTRACT

**BUILDING A DEEP LEARNING-BASED EMOTION CLASSIFICATION MODEL
AND APPLYING IT TO SOCIAL NETWORKS TO DETECT AND PREVENT
INAPPROPRIATE COMMENTS**

*Ha Trong Thang, Ho Thi Kieu Oanh, Nguyen Thanh Vi, Nguyen Huy Hoang,
Nguyen Cong Dung, Le Thi Thuy Lan, Tran Thi Van Anh**

Ho Chi Minh City University of Industry and Trade, Vietnam

Email: anhttv@huit.edu.vn

This paper presents a method for building a deep learning-based sentiment classification model to detect and prevent inappropriate comments on social networks. The authors use the PhoBERT model, a variant of BERT trained on Vietnamese data, to extract text features and classify emotions. The training dataset consists of comments collected from social networking platforms, labeled according to emotional states such as positive, negative and neutral. Experiments show that the PhoBERT model achieves an accuracy of over 90% when classifying emotions in Vietnamese comments. The system can be integrated into social networking platforms to automatically detect and handle inappropriate content, helping to create a healthy interactive environment.

Keywords: BERT, comments, PhoBERT, sentiment analysis

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Trần Thị Vân Anh

Số điện thoại liên hệ: 0909188906

DỰ ĐOÁN GIÁ THỊ TRƯỜNG BITCOIN THÔNG QU PHÂN TÍCH CẢM XÚC TIN TỨC DÙNG MONGODB

Trần Việt Hùng^{1,*}, Ngô Ngọc Huỳnh Như²

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

²Trường Đại học Ngân Hàng Thành phố Hồ Chí Minh

*Email: hungtv@huit.edu.vn

Ngày nhận bài: 31/03/2025; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Dự báo giá Bitcoin là thách thức trong công nghệ tài chính do tính biến động cao. Nghiên cứu này đề xuất phương pháp kết hợp phân tích cảm xúc tin tức Blockchain sử dụng TextBlob và phân tích nhân tố Factor Analysis để xác định mối tương quan tuyến tính với giá Bitcoin. Dữ liệu từ 3 nguồn tin tức CoinTelegraph, CryptoNews, CryptoPotato và giá lịch sử Bitcoin được xử lý trên nền tảng MongoDB. Kết quả cho thấy mô hình đạt độ chính xác cao trên tập huấn luyện và trên tập kiểm thử so với phương pháp dựa chỉ vào dữ liệu giá. Phân tích trọng số hồi quy khẳng định cảm xúc từ tin tức đặc biệt về Bitcoin và Ethereum ảnh hưởng mạnh hơn giá lịch sử ngày trước đó. Nghiên cứu chứng minh hiệu quả của MongoDB trong quản lý dữ liệu đa dạng và tiềm năng của phân tích cảm xúc trong dự báo thị trường tiền điện tử.

Từ khóa: Bitcoin, Blockchain, sentiment analysis, MongoDB.

1. MỞ ĐẦU

Khởi nguồn từ sách trắng năm 2009 của Satoshi Nakamoto [1], thị trường tiền điện tử đã hình thành. Bitcoin, dù đối mặt với các đối thủ như Ethereum hay Altcoin, vẫn giữ vị trí thống trị với vốn hóa vượt một nghìn tỷ USD[2]. Điều này khiến việc dự báo giá Bitcoin và các loại tiền điện tử khác trở thành một bài toán vừa hấp dẫn vừa khó khăn trong nghiên cứu [3][4].

Về mặt phương pháp, công nghệ blockchain tạo ra các thị trường ngang hàng, hoạt động trên mạng lưới phi tập trung, khác biệt với hệ thống tài chính tập trung truyền thống. Vì thế, phân tích cảm tính từ mạng xã hội đã nổi lên như một phương pháp dự đoán giá Bitcoin. Tuy nhiên, đây là lĩnh vực mới và mối liên hệ cụ thể với biến động tiền điện tử còn chưa được làm rõ hoàn toàn [5].

Nghiên cứu này xem xét mối quan hệ giữa phân tích cảm tính bằng công cụ TextBlob (trên dữ liệu từ ba nguồn báo chí về bitcoin) và biến động giá bitcoin lịch sử [6]. Dữ liệu được xử lý gần như thời gian thực bằng cơ sở dữ liệu MongoDB. MongoDB là cơ sở dữ liệu NoSQL, dạng tài liệu, mang lại sự linh hoạt và khả năng mở rộng để xử lý dữ liệu đa dạng và

khối lượng lớn thường thấy trong hệ sinh thái Blockchain, ví dụ như dữ liệu giao dịch, thị trường hoặc cảm tính. Đặc tính này phù hợp cho việc phân tích dữ liệu gần thời gian thực trong lĩnh vực này [7].

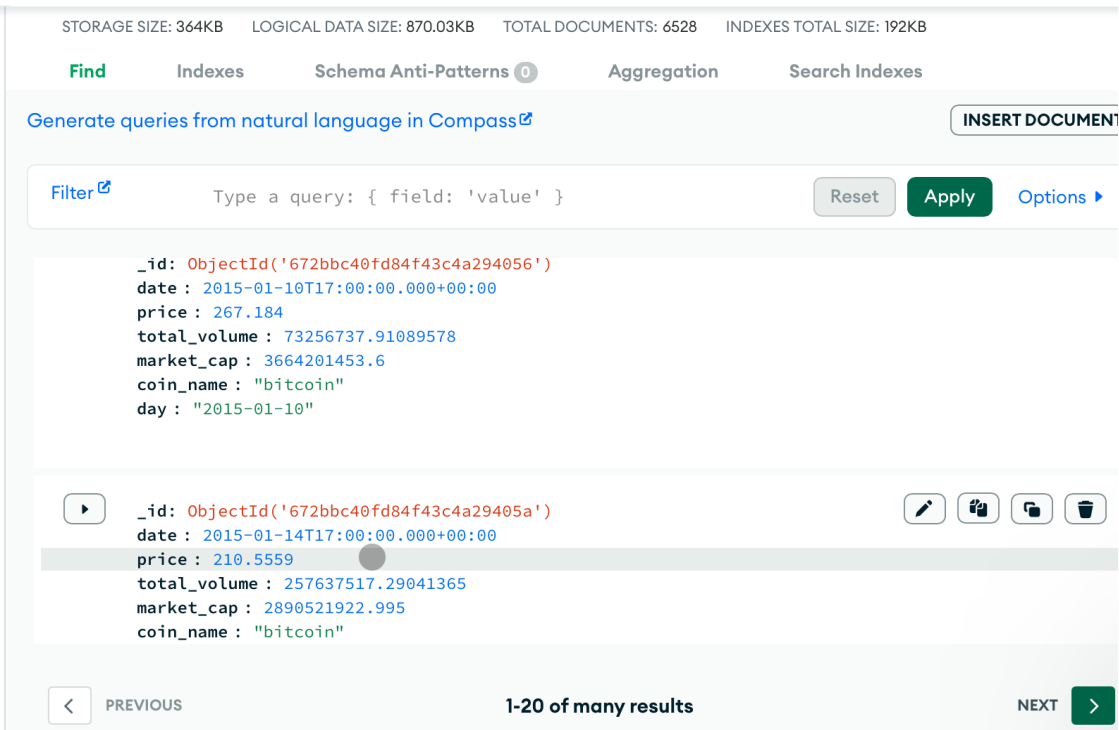
Phân tích factor analysis trong nghiên cứu này cho thấy mối liên hệ tuyến tính giữa chấm điểm cảm tính TextBlob hàng ngày và mức tăng giảm giá Bitcoin hàng ngày. Kết quả này gợi ý rằng việc tập trung vào các giá trị cảm tính tích lũy hoặc trung bình, có thể ở dạng nhị phân, sẽ hữu ích hơn cho các mô hình dự báo trong tương lai.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Dữ liệu

2.1.1 Dữ liệu tiền điện tử

Dữ liệu lịch sử giá Bitcoin và Ethereum của nghiên cứu này, trong khoảng thời gian từ ngày 1 tháng 12 năm 2021 đến ngày 1 tháng 12 năm 2023, được thu thập từ giá đóng cửa hàng ngày. Dữ liệu này lấy từ bộ dữ liệu [8] được cung cấp trên nền tảng Kaggle. Kaggle là một nền tảng thuộc Google, nổi tiếng với các cuộc thi khoa học dữ liệu và là kho lưu trữ nhiều bộ dữ liệu công khai hữu ích cho việc phân tích. Nguồn gốc của dữ liệu trong bộ này là CoinGecko, một trang web tổng hợp dữ liệu tiền điện tử rất phổ biến và uy tín, cung cấp thông tin về giá, khối lượng giao dịch, vốn hóa thị trường và nhiều chỉ số khác cho hàng nghìn loại tiền điện tử. Để phục vụ cho các phân tích sau này, nghiên cứu này cũng xem xét đến giá Bitcoin theo ngày.

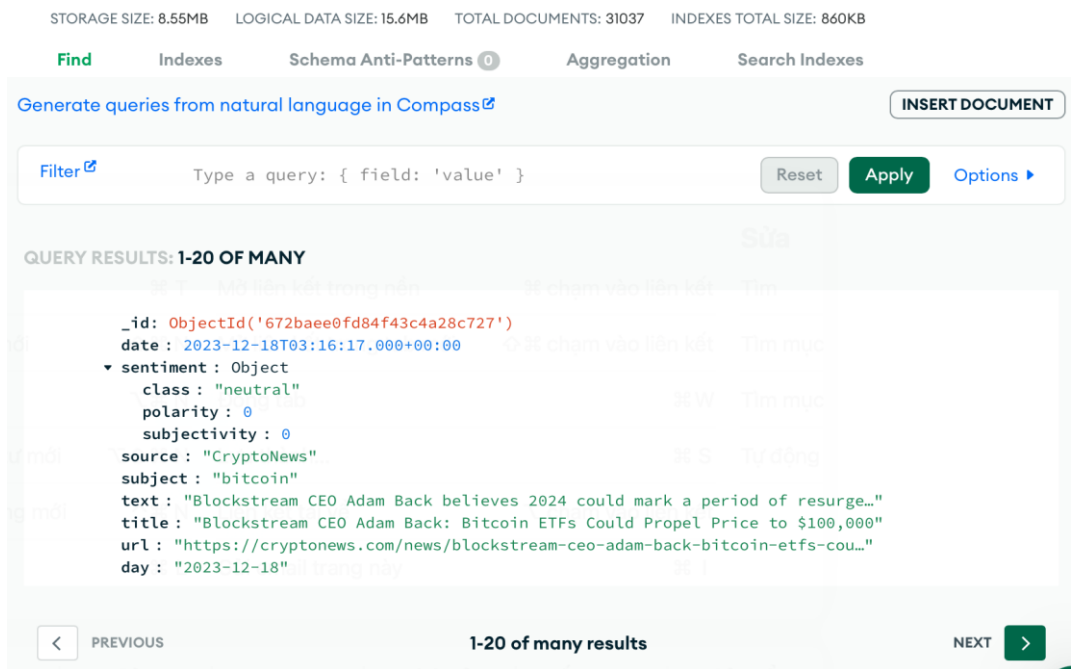


Hình 1. Dữ liệu giá Bitcoin và các đồng tiền ảo được trích xuất từ Google Kaggle [8] lên MongoDB, từ ngày 01 tháng 01 năm 2015 tới ngày 27 tháng 03 năm 2024

2.1.2 Dữ liệu tin tức về Blockchain:

Bộ dữ liệu tin tức tiền điện tử [9] trong nghiên cứu này bao gồm tiêu đề (title) và phần tóm tắt (headline) hàng ngày từ ba trang tin tức chính: CoinTelegraph, CryptoNews và CryptoPotato. Dữ liệu kéo dài từ ngày 12 tháng 10 năm 2021 đến ngày 19 tháng 12 năm 2023. Thêm vào đó, mỗi bản tin trong bộ dữ liệu đã được phân loại sẵn theo các chủ đề được gán nhãn bởi các trang tin, bao gồm: Altcoin, Bitcoin, Blockchain, DeFi, Ethereum và NFT.

Đối với mỗi bản tin, cả tiêu đề và phần tóm tắt đều được phân tích bằng TextBlob để xác định giá trị phân cực và chủ quan tương ứng trong nghiên cứu này, giá trị cảm tính TextBlob của mỗi bản tin được định nghĩa là trung bình cộng của điểm TextBlob cho tiêu đề và điểm TextBlob cho phần tóm tắt. Vì một ngày có thể có nhiều bản tin, điểm cảm tính TextBlob hàng ngày được xác định bằng cách lấy trung bình cộng của tất cả các điểm cảm tính từ các bản tin trong ngày đó. Ngoài ra, vào những ngày không có bản tin nào được ghi nhận, giá trị phân cực và chủ quan của TextBlob sẽ được đặt bằng 0, thể hiện rằng không có thông tin cảm tính cho những ngày đó.



Hình 2. Dữ liệu tin tức về tiền ảo được trích xuất từ Google Kaggle [9] lên MongoDB, từ ngày 13 tháng 10 năm 2021 đến ngày 19 tháng 12 năm 2023.

2.2 Phương pháp nghiên cứu

2.2.1 Phân tích ngôn ngữ cảm tính dùng TextBlob

TextBlob là một thư viện Python phổ biến, cung cấp phương pháp phân tích cảm tính dựa trên từ điển (lexicon-based) với mỗi đoạn văn bản, TextBlob đối chiếu từ ngữ với bộ từ vựng sẵn có và trả về hai kết quả chính: độ phân cực (polarity) và độ chủ quan (subjectivity). Giá trị độ phân cực dao động trong khoảng $[-1, 1]$, phản ánh cảm xúc tiêu cực hay tích cực được nhận diện trong văn bản. Tương tự, độ chủ quan nằm trong khoảng $[0, 1]$, ước lượng mức độ khách quan hay chủ quan của nội dung. Trong nghiên cứu này, TextBlob được sử dụng để xác định cảm tính hàng ngày từ các nguồn tin tức liên quan đến Bitcoin.

2.2.2 Phân tích hồi quy đa biến

Nghiên cứu sử dụng hồi quy đa biến [10] để xây dựng mô hình tuyến tính dự báo giá Bitcoin từ dữ liệu các đồng tiền ảo và dữ liệu cảm tính TextBlob từ thị trường. Dùng công thức toán học nghịch đảo Moore-Penrose [11] tìm trọng số cho từng nguồn dữ liệu sao cho tổng sai lệch giữa giá dự báo và thực tế là nhỏ nhất. Trọng số sau đó được chuẩn hóa để dễ so sánh: đồng tiền có trọng số cao hơn sẽ ảnh hưởng lớn hơn đến dự báo. Khi có dữ liệu mới, chỉ cần nhân với trọng số đã tính để ra giá Bitcoin dự kiến. Phương pháp này giống như phương pháp phân tích nhân tố chính từ dữ liệu nhiều nguồn để đoán xu hướng Bitcoin.

3. KẾT QUẢ VÀ THẢO LUẬN

Trong nghiên cứu này, dữ liệu được chia theo tỉ lệ 80%–20% để đánh giá khả năng tổng quát hóa của mô hình, xét từ ngày 1 tháng 12 năm 2021 đến ngày 1 tháng 12 năm 2023. Cụ thể, tập huấn luyện train (80%) là dữ liệu giá mỗi ngày từ 01/12/2021 đến 15/08/2023, dùng để xây dựng mô hình và tối ưu hóa trọng số. Tập kiểm định test (20%) là dữ liệu giá mỗi ngày từ 16/08/2023 đến 01/12/2023, dùng để kiểm tra hiệu suất mô hình trong điều kiện thực tế.

3.1 Xác định giá trị dự báo Bitcoin trên tập Train:

3.1.1 Phân tích hồi quy đa tuyến tính (Factor analysis) theo thời gian:

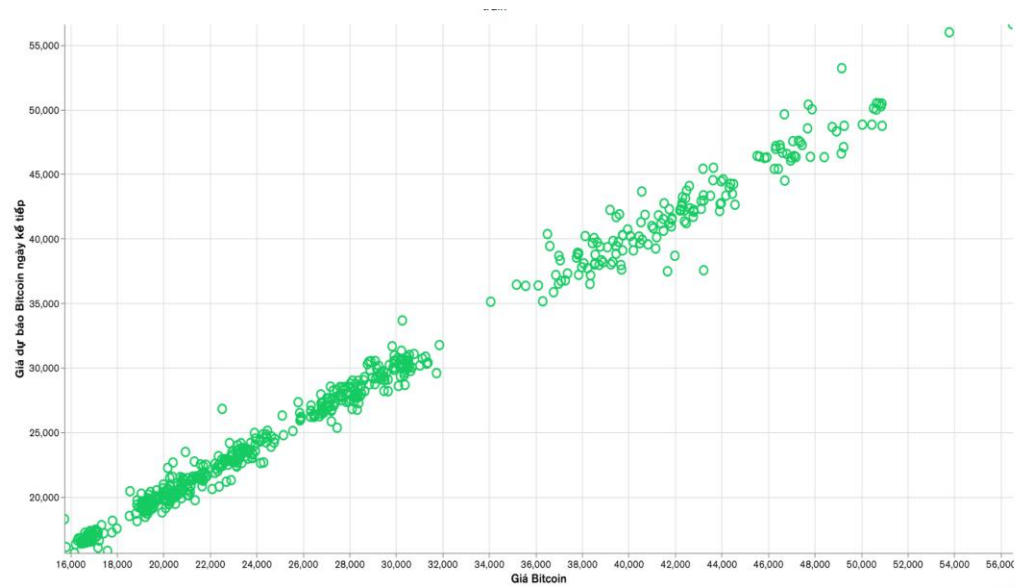
Kết quả của phương pháp hồi quy đa tuyến tính cho thấy kết quả dự báo Bitcoin trong ngày tiếp theo bám rất sát kết quả thực tế của Bitcoin trong tập huấn luyện Train.



Hình 3. Dữ liệu Bitcoin và kết quả dự báo cho tập Train trên MongoDB Chart.

3.1.2 Biểu đồ phân tán (Scatter plot):

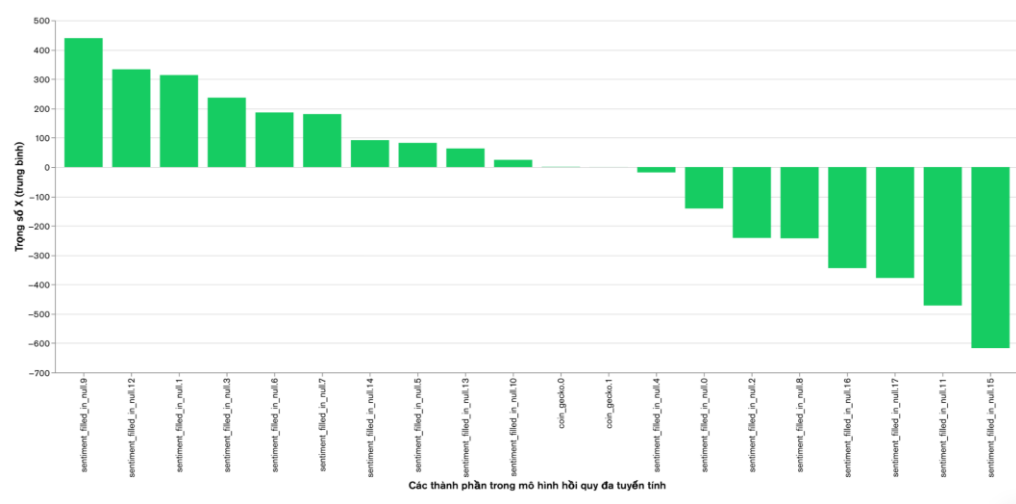
Kết quả của biểu đồ phân tán (càng trùng đường chéo 45 độ càng chính xác) cho thấy kết quả dự báo Bitcoin trong ngày tiếp theo bám khá sát kết quả thực tế của Bitcoin trong tập huấn luyện Train. Độ chính xác của giá dự báo nhìn vào đồ thị có thể ước lượng là hơn 90%.



Hình 4. Dữ liệu Bitcoin và kết quả dự báo theo biểu đồ phân tán trên MongoDB Chart.

3.1.3 Biểu đồ trọng số trong phương pháp hồi quy đa tuyến tính:

Kết quả của biểu đồ trọng số cho thấy các chỉ số ngôn ngữ cảm tính tích cực và tiêu cực trên các bài đăng tin tức có gây ảnh hưởng đến chất lượng dự báo giá Bitcoin. Hai trọng số cho giá Bitcoin và Ethereum trong lịch sử ngày trước đó (được dùng để dự báo giá Bitcoin cho ngày hôm sau) có giá trị nhỏ hơn nhiều trọng số của TextBloc khi phân tích cảm tính trong ngôn ngữ tin tức.



Hình 5. Trọng số trong mô hình hồi quy đa tuyến tính. Coin 0 là giá Bitcoin, Coin 1 là giá Ethereum.

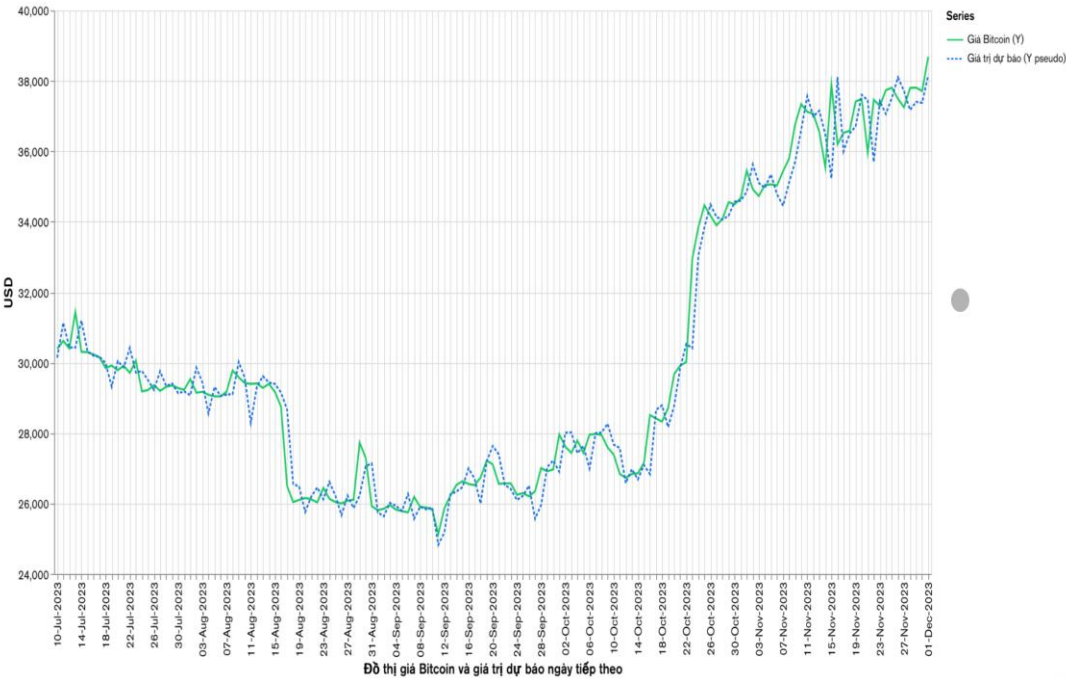
Bảng 1. Kết quả trọng số của mô hình hồi quy đa tuyến tính trong Hình 5.

Trọng số nguồn tin tức	altcoin	bitcoin	blockchain	defi	ethereum	nft
CryptoNews	-202	267	-215	270	-50	47
CryptoPotato	3	229	-170	384	75	-568
CoinTelegraph	318	94	47	-844	-357	-358

3.2 Xác định giá trị dự báo Bitcoin trên tập Test:

3.2.1 Phân tích hồi quy đa tuyến tính (Factor analysis) theo thời gian:

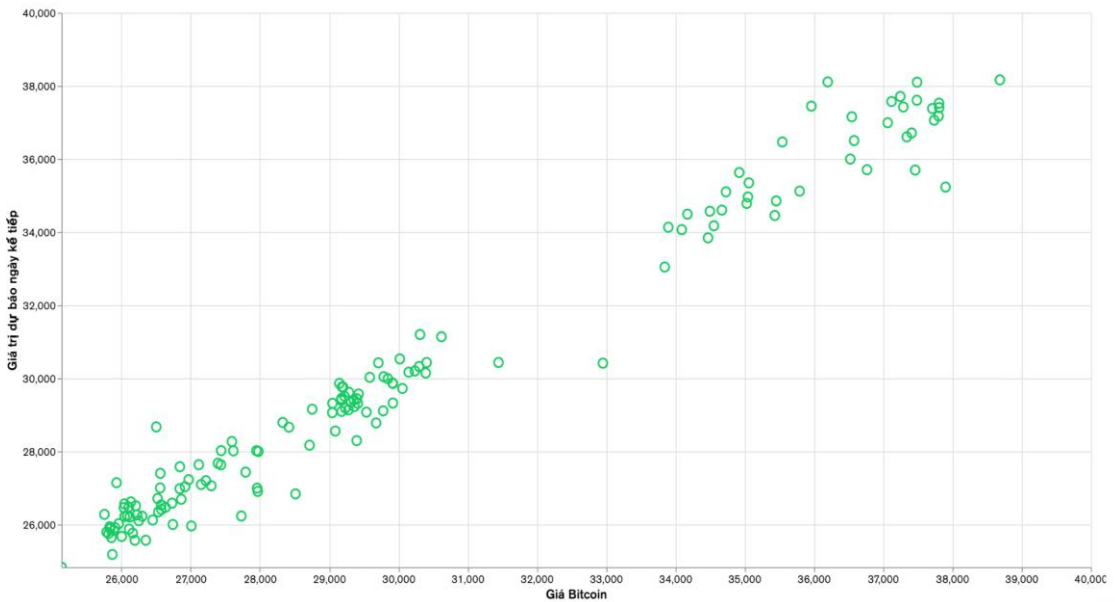
Có thể thấy chất lượng dự báo trên tập Test (sử dụng cùng trọng số trên tập Train) bị suy giảm nhưng vẫn đảm bảo được độ bám của dữ liệu Bitcoin thực tế.



Hình 6. Dữ liệu Bitcoin và kết quả dự báo cho tập Test trên MongoDB Chart.

3.2.2 Biểu đồ phân tán (Scatter plot):

Kết quả của biểu đồ phân tán (càng trùng đường chéo 45 độ càng chính xác) cho thấy kết quả dự báo Bitcoin trong ngày tiếp theo bám khá sát kết quả thực tế của Bitcoin trong tập kiểm thử Test. Độ chính xác của giá dự báo nhìn vào đồ thị có thể ước lượng là vẫn hơn 80%.



Hình 7. Dữ liệu Bitcoin và kết quả dự báo theo biểu đồ phân tán trên MongoDB Chart.

4. KẾT LUẬN

Nghiên cứu này đã chứng minh một cách định lượng về khả năng ứng dụng phân tích cảm xúc từ tin tức để dự đoán giá Bitcoin. Thông qua việc áp dụng phương pháp hồi quy đa tuyến tính (Factor Analysis) trên dữ liệu giá Bitcoin, Ethereum và điểm cảm xúc TextBlob từ ba nguồn tin tức lớn (CoinTelegraph, CryptoNews, CryptoPotato) trong giai đoạn 2021-2023, kết quả cho thấy mô hình dự báo giá Bitcoin cho ngày tiếp theo đạt độ chính xác cao trên tập huấn luyện (ước tính >90%) và duy trì độ bám khá tốt trên tập kiểm thử (ước tính >80%). Điều này khẳng định vai trò của cảm xúc thị trường, được phản ánh qua tin tức, như một yếu tố dự báo tiềm năng cho biến động giá Bitcoin.

Phân tích trọng số từ mô hình hồi quy đa tuyến tính đặc biệt nhấn mạnh tầm quan trọng của các chỉ số cảm xúc TextBlob. Các trọng số tương ứng với dữ liệu cảm xúc từ tin tức có giá trị lớn hơn đáng kể so với trọng số của giá Bitcoin và Ethereum lịch sử của ngày trước đó. Phát hiện này cho thấy, ít nhất trong khuôn khổ mô hình tuyến tính được sử dụng, thông tin cảm xúc được trích xuất từ tin tức hàng ngày đóng góp một phần quan trọng vào việc giải thích và dự báo biến động giá Bitcoin, vượt trội hơn so với việc chỉ dựa vào dữ liệu giá quá khứ gần nhất. Việc lưu trữ và xử lý dữ liệu này trên nền tảng MongoDB cũng cho thấy tính hiệu quả trong việc quản lý dữ liệu đa dạng và thay đổi nhanh của thị trường tiền điện tử.

Mặc dù kết quả khả quan, nghiên cứu vẫn có những hạn chế nhất định như việc chỉ sử dụng mô hình tuyến tính và công cụ TextBlob tương đối đơn giản. Hướng phát triển trong tương lai có thể bao gồm việc khám phá các mô hình phi tuyến phức tạp hơn (như mạng nơ-ron hồi quy), áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến hơn để nắm bắt sắc thái cảm xúc tinh tế hơn và mở rộng nguồn dữ liệu sang các nền tảng mạng xã hội. Bên cạnh đó, việc kiểm tra hiệu quả của việc sử dụng giá trị cảm xúc tích lũy hoặc dạng nhị phân, như gợi ý ban đầu, cũng là một hướng đi đáng giá để cải thiện độ chính xác và tính ổn định của mô hình dự báo giá Bitcoin.

TÀI LIỆU THAM KHẢO

- [1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," White Paper, 2009.
- [2] S. A. Sarkodie, M. Y. Ahmed, and T. Leirvik, "Trade volume affects Bitcoin energy consumption and carbon footprint," *Finance Research Letters*, vol. 48, 2022.
- [3] S. Park and J.-S. Yang, "Machine learning models based on bubble analysis for Bitcoin market crash prediction," *Engineering Applications of Artificial Intelligence*, vol. 135, p. 105678, 2024.
- [4] A. Bâra and S.-V. Oprea, "An ensemble learning method for Bitcoin price prediction based on volatility indicators and trend," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 105123, 2024.
- [5] S. Singh, A. Pise, and B. Yoon, "Prediction of Bitcoin stock price using feature subset optimization," *Heliyon*, vol. 10, no. 7, p. e12345, 2024.
- [6] A. C. Baroiu, V. Diaconita, and S. V. Oprea, "Bitcoin volatility in bull vs. bear market-insights from analyzing on-chain metrics and Twitter posts," *PeerJ Computer Science*, vol. 9, p. e1234, 2023.
- [7] S. Arslan, "Bitcoin price prediction using sentiment analysis and empirical mode decomposition," *Computational Economics*, 2024.
- [8] S. Sudalairajkumar, "Cryptocoins historical prices - CoinGecko," *Kaggle Dataset*, 2024. [Online]. Available: <https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrency-historical-prices-coingecko>
- [9] Olivier, "Crypto news dataset," *Kaggle Dataset*, 2024. [Online]. Available: <https://www.kaggle.com/datasets/oliviervha/crypto-news>
- [10] V. H. Tran, "Variational Bayes inference in digital receivers," Ph.D. dissertation, Trinity College Dublin, Ireland, 2014. [Online]. Available: <https://www.tara.tcd.ie/handle/2262/80528>
- [11] V. H. Tran and W. Wang, "Bayesian inference for PCA and MUSIC algorithms with unknown number of sources," *arXiv preprint arXiv:1809.10168*, 2018. [Online]. Available: <https://arxiv.org/abs/1809.10168>

ABSTRACT

FORECASTING BITCOIN MARKET PRICE VIA SENTIMENT ANALYSIS OF NEWS AND MONGODB

Tran Viet Hung^{1,*}, Ngo Ngoc Huynh Nhu²

¹ Ho Chi Minh City University of Industry and Trade

² Ho Chi Minh City University of Banking

*Email: hungtv@huit.edu.vn

Predicting Bitcoin prices presents a significant challenge in financial technology due to its high volatility. This study proposes a method combining Blockchain news sentiment analysis using TextBlob and Factor Analysis to identify linear correlations with Bitcoin prices. Data from three news sources (CoinTelegraph, CryptoNews, CryptoPotato) and historical Bitcoin prices were processed on the MongoDB platform. Results demonstrate that the model achieves high accuracy on both the training and testing datasets compared to price-only predictive methods. Regression weight analysis confirms that news sentiment, particularly regarding Bitcoin and Ethereum, has a stronger impact on price prediction than historical prices from the previous day. The study highlights MongoDB's effectiveness in managing diverse data formats and underscores the potential of sentiment analysis for forecasting cryptocurrency market trends.

Keywords: Bitcoin, Blockchain, sentiment analysis, MongoDB.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Trần Việt Hùng

Số điện thoại liên hệ: 0368670836

ĐÁNH GIÁ HIỆU SUẤT CỦA MÔ HÌNH SINH KHUẾCH TÁN THÔNG QUA LỰA CHỌN HÀM MẤT MẤT

Hoàng Thế Anh¹, Trần Đình Toàn^{1*}

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: toantd@huit.edu.vn

Ngày nhận bài 20/3/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Hiệu suất của mô hình sinh ảnh khuếch tán phụ thuộc mạnh vào lựa chọn hàm mất mát. Tuy nhiên, việc sử dụng phổ biến hàm MSE trong huấn luyện mô hình này vẫn còn nhiều hạn chế về mặt chất lượng ảnh sinh ra. Nghiên cứu này đánh giá tác động của việc thay đổi các hàm mất mát trong mô hình sinh khuếch tán dựa trên kiến trúc Unet. Thực nghiệm trên bộ dữ liệu MNIST cho thấy việc thay thế MSE bằng hàm CharbonnierLoss giúp cải thiện rõ rệt chất lượng ảnh, được phản ánh qua các chỉ số FID và KID. Kết quả cho thấy lựa chọn hàm mất mát phù hợp đóng vai trò quan trọng trong việc nâng cao hiệu quả sinh ảnh của mô hình.

Từ khóa: Diffusion Model, Image Generation, Loss Function, FID, KID.

1. MỞ ĐẦU

Trong những năm gần đây, lĩnh vực sinh ảnh bằng mô hình học sâu đã có những bước phát triển vượt bậc, đặc biệt là với sự xuất hiện của các mô hình như GANs [1, 2], VAEs và gần đây là mô hình khuếch tán (diffusion models) [1]. Các phương pháp này không chỉ góp phần cải thiện chất lượng ảnh sinh ra mà còn mở ra nhiều ứng dụng tiềm năng trong các lĩnh vực như tổng hợp dữ liệu, tăng cường dữ liệu huấn luyện, phục hồi ảnh và nhiều bài toán thị giác máy tính khác [3].

Trong số các phương pháp kể trên, mô hình khuếch tán, điển hình là Denoising Diffusion Probabilistic Models (DDPM), đang thu hút sự chú ý nhờ khả năng tạo ảnh sắc nét, tự nhiên và ổn định hơn so với GANs hoặc VAEs trong nhiều trường hợp [4, 5]. Với cơ chế khử nhiễu từng bước, DDPM cho phép tái tạo ảnh từ nhiễu Gaussian bằng cách học ngược lại quá trình khuếch tán, mang lại kết quả khả quan trong cả mô hình tổng quát và mô hình chuyên biệt theo miền dữ liệu [4], [6].

Tuy nhiên, hiệu quả của mô hình DDPM không chỉ phụ thuộc vào kiến trúc mạng như Unet mà còn bị chi phối bởi lựa chọn hàm mất mát được sử dụng trong quá trình huấn luyện [7]. Trong nhiều nghiên cứu trước đây, hàm mất mát Mean Squared Error (MSE) thường được sử dụng như một lựa chọn mặc định do tính đơn giản và khả năng hội tụ ổn định. Tuy nhiên, MSE lại dễ bị ảnh hưởng bởi nhiễu lớn và không phản ánh tốt các đặc trưng thị giác mà con người cảm nhận được. Điều này đặt ra câu hỏi: liệu các hàm mất mát khác có thể cải thiện chất lượng ảnh sinh ra khi tích hợp vào mô hình khuếch tán?

Xuất phát từ vấn đề trên, nghiên cứu này tập trung đánh giá hiệu suất của năm lựa chọn hàm mất mát khác nhau trong mô hình sinh khuếch tán DDPM sử dụng kiến trúc Unet: MSE, L1, Huber Loss, Log-Cosh Loss và Charbonnier Loss. Đây đều là các hàm mục tiêu có đặc tính khác nhau về độ nhạy với nhiễu, tính trơn, và khả năng phản ánh sai lệch giữa ảnh gốc và ảnh dự đoán [8, 9]. Thông qua việc huấn luyện các mô hình tương ứng trên tập dữ liệu MNIST một tập dữ liệu hình ảnh đơn giản nhưng tiêu chuẩn, nghiên cứu này sử dụng hai chỉ số đánh giá khách quan phổ biến là FID (Fréchet Inception Distance) và KID (Kernel Inception Distance) để đo lường chất lượng ảnh sinh ra [10, 11].

Mục tiêu cuối cùng là xác định loại hàm mất mát nào giúp mô hình sinh khuếch tán mang lại hiệu quả sinh ảnh tốt hơn trong bối cảnh hiện tại. Từ đó, nghiên cứu không chỉ cung cấp một góc nhìn rõ ràng hơn về ảnh hưởng của hàm mất mát đến hiệu suất mô hình, mà còn đề xuất hướng tối ưu hoá kiến trúc DDPM thông qua lựa chọn hàm mục tiêu sao cho phù hợp [12]. Đây có thể là một bước quan trọng trong việc nâng cao chất lượng đầu ra và mở rộng phạm vi ứng dụng thực tiễn của các mô hình sinh ảnh khuếch tán trong tương lai. Các kết quả được trình bày lần lượt ở phần 2 với phương pháp nghiên cứu, thực nghiệm kiểm chứng thể hiện ở phần 3, phần 4 trình bày một số kết luận và định hướng nghiên cứu tiếp theo.

2. PHƯƠNG PHÁP

2.1 Denoising Diffusion Probabilistic Models (DDPM)

Mô hình khuếch tán (Diffusion Model) là một phương pháp sinh ảnh dựa trên quá trình thêm nhiễu Gaussian dần dần vào dữ liệu gốc (quá trình forward) và sau đó học cách đảo ngược quá trình này để khử nhiễu [4], [6] và tái tạo lại ảnh (quá trình reverse).

Mô hình DDPM kết hợp với kiến trúc UNet hai chiều (2D UNet) [24], [26]. UNet là kiến trúc encoder-decoder có kết nối skip connection giữa các tầng encoder và decoder [13], cho phép mô hình giữ lại thông tin chi tiết ở nhiều cấp độ khi khử nhiễu từng bước thời gian. Trong quá trình sinh ảnh, mô hình UNet nhận vào ảnh nhiễu và thời điểm (timestep), sau đó dự đoán nhiễu tương ứng để dần tái tạo ảnh gốc. Thông tin về thời gian được đưa vào mô hình UNet bằng cách mã hóa vị trí dưới dạng các hàm sin – cos, sau đó được kết hợp vào các tầng mạng thông qua các khối ResNet. Mục tiêu của thực nghiệm là đánh giá hiệu quả sinh ảnh của các mô hình diffusion khi thay đổi hàm mất mát, trong cùng một cấu trúc mạng và điều kiện huấn luyện.

Trong quá trình khuếch tán tiến (Forward Process), ảnh $\mathbf{x}_0$ bị làm nhiễu dần dần qua T bước bằng cách thêm một lượng nhiễu có kiểm soát. Công thức tổng quát như (1) sau:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t I) \quad (1)$$

Trong đó:

$\mathbf{x}_t$ là mẫu tại bước thời gian t ,

$\mathbf{x}_{t-1}$ là mẫu tại bước thời gian $t-1$,

β_t là hệ số nhiễu tại bước t ,

I là ma trận đơn vị,

$\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ biểu thị phân phối chuẩn với trung bình $\boldsymbol{\mu}$ và hiệp phương sai $\boldsymbol{\Sigma}$.

Tổng quát sau t bước, phân phối của [1] $\mathbf{x}_t$ có thể được viết lại dưới dạng công thức (2):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)I) \quad (2)$$

Trong đó:

$$\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$$

Quá trình khử nhiễu là quá trình đảo ngược, trong đó mô hình học sâu $p_\theta(x_{t-1}|x_t)$ ước lượng nhiễu để từng bước tái tạo lại ảnh gốc, như trong công thức (3):

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (3)$$

Với $\mu_\theta(x_t, t)$ là giá trị trung bình của phân phối nhiễu được ước lượng.

Mô hình khuếch tán được sử dụng trong nghiên cứu này bao gồm hai giai đoạn chính: forward process (quá trình nhiễu hóa) và reverse process (quá trình sinh ảnh). Đây là hai bước nền tảng tạo nên cơ chế hoạt động của mô hình.

Ở giai đoạn forward, ảnh gốc x_0 được thêm nhiễu Gaussian dần dần qua từng bước thời gian $t = 1$ đến T . Quá trình này hoàn toàn không cần đến mạng học sâu, mà chỉ là thao tác thêm nhiễu có kiểm soát để mô phỏng sự phá hủy dần dần của dữ liệu ban đầu. Kết quả của quá trình này là một ảnh nhiễu hoàn toàn x_T , gần như không còn thông tin ban đầu, có thể được xem là một mẫu từ phân phối Gaussian chuẩn.

Giai đoạn reverse là phần quan trọng nhất của mô hình, nơi quá trình sinh ảnh diễn ra bằng cách khử nhiễu từng bước từ x_T về lại ảnh x_0 . Tại mỗi timestep t , mô hình học sâu UNet được sử dụng để dự đoán nhiễu [12, 13] đã được thêm vào ở bước forward tương ứng. Đầu vào của UNet là ảnh x_T hiện tại và timestep t , đầu ra là nhiễu ước lượng $\hat{\epsilon}_\theta(x_t, t)$.

Dựa vào dự đoán này, mô hình tính toán lại ảnh x_{t-1} , tiến dần về ảnh ban đầu.

Thông tin về thời gian t được đưa vào UNet thông qua một mảng nhúng (embedding), sau đó được tích hợp vào các tầng xử lý qua các khối ResNet để giúp mô hình điều kiện hóa theo từng bước thời gian [4]. Cấu trúc encoder-decoder cùng với các kết nối skip connections của UNet giúp mô hình tận dụng hiệu quả thông tin ở nhiều cấp độ trừu tượng khác nhau [12, 13], góp phần cải thiện chất lượng ảnh sinh ra.

Training	Sampling
1: Repeat 2: $x_0 \sim q(x_0)$ 3: $t \sim \text{Uniform}(\{1, \dots, T\})$ 4: $\epsilon \sim \mathcal{N}(0, I)$ 5: Take gradient descent step on $\nabla_\theta \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t) ^2$ 6: until converged	1: $x_T \sim \mathcal{N}(0, I)$ 2: for $t = T, \dots, 1$ do 3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$ 4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \cdot \text{UNet}(x_t, t) \right) + \sigma_t z$ 5: end for 6: return x_0

2.2 Hàm mất mát

2.2.1 Mean Squared Error (MSE)

MSE là một trong những hàm mất mát phổ biến và cơ bản nhất trong học sâu. Nó được tính bằng bình phương sai số giữa giá trị dự đoán và giá trị thực tế, như công thức (4):

$$\text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2 \quad (4)$$

MSE nhấn mạnh vào sai số lớn do bản chất bình phương, giúp mô hình hội tụ ổn định. Tuy nhiên, nó nhạy cảm với nhiễu và có thể dẫn đến ảnh sinh ra bị mờ (blurry), đặc biệt trong các bài toán sinh ảnh.

2.2.2 Mean Absolute Error (L1 Loss)

L1 Loss đo sai số tuyệt đối trung bình giữa đầu ra dự đoán và thực tế, như công thức (5):

$$\text{L1}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n |\mathbf{x}_i - \hat{\mathbf{x}}_i| \quad (5)$$

L1 ít nhạy cảm với nhiễu hơn MSE và có xu hướng bảo toàn biên tốt hơn trong ảnh. Tuy nhiên, nó có thể gây khó khăn cho quá trình tối ưu vì đạo hàm không liên tục tại 0.

2.2.3 Huber Loss

Huber Loss [8] kết hợp giữa MSE và L1 nhằm cân bằng giữa độ mượt và khả năng chống nhiễu, thể hiện như công thức (6):

$$\text{Huber}(\hat{\mathbf{x}}, \mathbf{x}) = \begin{cases} \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^2 & \text{nếu } |\mathbf{x} - \hat{\mathbf{x}}| \leq \delta \\ \delta(|\mathbf{x} - \hat{\mathbf{x}}| - \frac{1}{2}\delta) & \text{nếu } |\mathbf{x} - \hat{\mathbf{x}}| > \delta \end{cases} \quad (6)$$

Với tham số ngưỡng δ , Huber Loss hoạt động như MSE với sai số nhỏ và như L1 với sai số lớn, giúp ổn định hơn khi gặp nhiễu hoặc điểm ngoại lai.

2.2.4 Log-Cosh Loss

Log-Cosh là một biến thể mượt của L1, được định nghĩa như công thức (7) sau:

$$\text{LogCosh}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=1}^n \log(\cosh(\mathbf{x}_i - \hat{\mathbf{x}}_i)) \quad (7)$$

Do đạo hàm của nó tương đương với L1 cho sai số lớn và gần như MSE cho sai số nhỏ, Log-Cosh vừa ổn định trong tối ưu hóa vừa chống lại nhiễu tốt hơn MSE.

2.2.5 Charbonnier Loss

Charbonnier Loss [15] là một dạng làm mượt của L1 Loss, còn được gọi là "differentiable L1", có dạng như công thức (8):

$$\text{Charbonnier}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=1}^n \sqrt{(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2 + \epsilon^2} \quad (8)$$

Với ϵ là một hằng số nhỏ (ví dụ: 10^{-6}), Charbonnier giúp tránh đạo hàm không xác định như ở L1, đồng thời giữ được tính bền vững trước nhiễu, thường được dùng trong các tác vụ phục hồi ảnh.

3. THỰC NGHIỆM VÀ KẾT QUẢ

3.1 Dữ liệu

Tập dữ liệu được sử dụng là MNIST, bao gồm 70.000 ảnh số viết tay từ 0 đến 9, mỗi ảnh có kích thước 28×28 pixel và một kênh (grayscale). Đây là tập dữ liệu tiêu chuẩn thường dùng để kiểm tra các mô hình sinh ảnh vì đơn giản, rõ ràng và dễ dàng đánh giá bằng mắt thường..

Bài toán đặt ra là sinh ảnh số viết tay từ nhiều Gaussian thông qua cơ chế khử nhiễu của mô hình diffusion. Chất lượng ảnh sinh ra được đánh giá bằng chỉ số FID và KID nhằm phân ánh mức độ gần gũi giữa phân phối ảnh sinh và ảnh thực.

3.2 Thực nghiệm

Các thực nghiệm trong nghiên cứu này được thực hiện trên hệ thống máy tính Apple Mac mini M4 (phiên bản 2024), sử dụng vi xử lý Apple M4 tích hợp CPU 8 nhân (gồm 4 nhân hiệu năng cao và 4 nhân tiết kiệm điện), cùng với GPU 10 nhân hỗ trợ nền tảng tăng tốc Metal Performance Shaders (MPS). Máy được trang bị 16 GB bộ nhớ hợp nhất (unified memory), chia sẻ giữa CPU và GPU, chạy hệ điều hành macOS 15 Sequoia.

Mô hình được huấn luyện với các thông số như sau: số bước khuếch tán (diffusion steps) được đặt là 1000. Hàm mất mát Mean Squared Error (MSE) được sử dụng để đo sai số giữa ảnh dự đoán và ảnh gốc.

Learning rate là 0,0002 và batch size là 128. Tối ưu hóa được thực hiện bằng thuật toán Adam. Các mô hình lần lượt sử dụng các hàm mất mát L1, Huber Loss, Log-Cosh Loss và Charbonnier Loss được huấn luyện với cùng cấu hình và điều kiện, như trong bảng 1 sau:

Bảng 1.Cấu hình huấn luyện cho các mô hình DDPM-Unet với các hàm mất mát khác nhau

Tên mô hình	Hàm mất mát	Số bước khuếch tán	Batch size	Learning rate	Optimizer
DDPM-Unet (MSE)	Mean Squared Error	1000	128	0,0002	Adam
DDPM-Unet (L1)	Mean Absolute Error	1000	128	0,0002	Adam
DDPM-Unet (Huber)	Huber Loss	1000	128	0,0002	Adam
DDPM-Unet (Log-Cosh)	Log-Cosh Loss	1000	128	0,0002	Adam
DDPM-Unet (CharbonnierLoss)	CharbonnierLoss	1000	128	0,0002	Adam

3.3 Chỉ số đánh giá

3.3.1 Fréchet Inception Distance (FID)

FID đo khoảng cách giữa phân phối đặc trưng của ảnh thực và ảnh sinh ra, tính toán dựa trên đặc trưng trích xuất từ mạng InceptionV3 [9]. Công thức tính FID như công thức (9) sau:

$$FID = \left\| \mu_r - \mu_g \right\|^2 + Tr \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right) \tag{9}$$

Trong đó:

μ_r, Σ_r là trung bình và ma trận hiệp phương sai của ảnh thực.

μ_g, Σ_g là trung bình và ma trận hiệp phương sai của ảnh sinh ra.

3.3.2 Kernel Inception Distance (KID)

KID cũng đo khoảng cách [10] giữa phân phối của ảnh thực và ảnh sinh ra nhưng sử dụng thống kê kernel Maximum Mean Discrepancy (MMD). KID có ưu điểm là ước lượng không chệch (Unbiased Estimator), giúp đánh giá mô hình chính xác hơn trên tập dữ liệu nhỏ.

Cả hai chỉ số này đều được sử dụng để đánh giá chất lượng ảnh sinh ra từ mô hình DDPM, so sánh với các phương pháp sinh ảnh khác như GAN và VAE [1], [15-17].

3.4 Kết quả

3.4.1 Đánh giá

Hiệu suất của mô hình khuếch tán được đánh giá thông qua hai chỉ số phổ biến là FID (Fréchet Inception Distance) và KID (Kernel Inception Distance). Mô hình được huấn luyện trên tập dữ liệu MNIST và so sánh với một số phương pháp sinh ảnh khác. Kết quả cho thấy mô hình khuếch tán với kiến trúc UNet đạt chất lượng ảnh sinh ra cao, với chỉ số FID thấp và KID ổn định, phản ánh khả năng tái tạo ảnh gần với dữ liệu gốc và phân phối ảnh sinh có độ phân tán hợp lý.

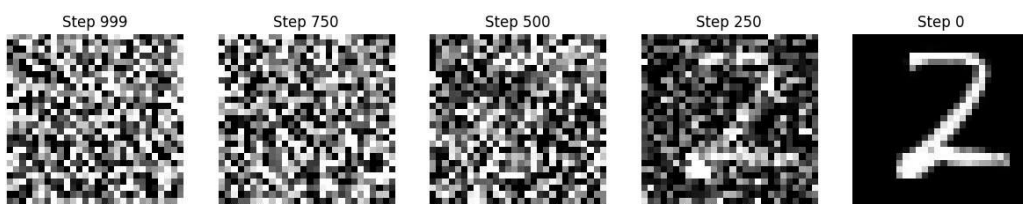
Bảng 2. So sánh kết quả hiệu suất của các mô hình sinh ảnh

Mô hình	FID ↓	KID ↓
DDPM Unet(CharbonnierLoss)	7,472	0,006
DDPM Unet (MSE)	10,79	0,009
DDPM Unet (HuberLoss)	13,78	0,159
DDPM Unet (L1)	17,085	0,014
DDPM Unet(LogCoshLoss)	33,14	0,4

Từ kết quả thực nghiệm ở bảng 2, cho thấy mô hình sinh khuếch tán sử dụng hàm mất mát CharbonnierLoss đạt hiệu suất sinh ảnh tốt nhất với chỉ số FID và KID thấp nhất trong số các mô hình đã khảo sát, phản ánh khả năng tái tạo ảnh gần với ảnh thật cao hơn so với kết hợp các hàm mất mát khác. Điều này cho thấy CharbonnierLoss là một lựa chọn tốt trong bối cảnh bài toán sinh ảnh bằng mô hình khuếch tán nói chung và trên tập dữ liệu MNIST nói riêng.

3.4.2 Ảnh minh họa

Hình 1 và hình 2 dưới đây minh họa quá trình sinh ảnh của mô hình khuếch tán. Ở hình 2, hàng đầu tiên là ảnh gốc từ tập dữ liệu MNIST, trong khi hàng thứ hai thể hiện các ảnh được sinh ra thông qua cơ chế khử nhiễu tuần tự đặc trưng của mô hình khuếch tán.



Hình 1. Quá trình khử nhiễu của mô hình khuếch tán



Hình 2. So sánh ảnh thật và ảnh được tạo ra

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu này đã tiến hành đánh giá hiệu suất của mô hình sinh ảnh khuếch tán (Diffusion Model) với nhiều hàm mất mát khác nhau trên tập dữ liệu MNIST. Kết quả thực nghiệm cho thấy hàm CharbonnierLoss là lựa chọn tối ưu nhất trong số các hàm khảo sát, cho chất lượng ảnh sinh ra tốt hơn rõ rệt theo cả hai chỉ số FID và KID.

Kết quả này có thể khẳng định việc thay đổi hàm mất mát (*một thành phần tương trưng chừng đơn giản*) có thể tạo ra khác biệt lớn trong hiệu quả sinh ảnh mà không cần sửa đổi nhiều kiến trúc mô hình. Điều này cũng mở ra tiềm năng cải tiến chất lượng ảnh sinh bằng các kỹ thuật tối ưu hóa hàm mục tiêu, đặc biệt trong các bài toán yêu cầu độ chính xác thị giác cao.

Trong tương lai, sẽ tiếp tục nghiên cứu tối ưu thời gian sinh ảnh bằng cách giảm số bước khuếch tán hoặc chuyển sang không gian tiềm ẩn có thể là hướng đi đáng cân nhắc. Ngoài ra, việc thử nghiệm trên các tập dữ liệu đa dạng và ứng dụng vào các bài toán thực tế như y tế hoặc thị giác máy tính sẽ giúp làm rõ hơn tính ứng dụng và độ linh hoạt của mô hình khuếch tán kết hợp với các hàm mất mát khác nhau.

5. TÀI LIỆU THAM KHẢO

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.
2. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 105–114. doi: 10.1109/CVPR.2017.19.
3. F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," 2022, doi: 10.48550/ARXIV.2209.04747.
4. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models." arXiv, 2020. doi: 10.48550/ARXIV.2006.11239.
5. M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs." arXiv, 2018. doi: 10.48550/ARXIV.1801.01401.
6. J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models." arXiv, 2020. doi: 10.48550/ARXIV.2010.02502.
7. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 586–595. doi: 10.1109/CVPR.2018.00068.

8. P. J. Huber, “Robust Estimation of a Location Parameter,” *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964, doi: 10.1214/aoms/1177703732.
9. H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss Functions for Neural Networks for Image Processing.” arXiv, 2015. doi: 10.48550/ARXIV.1511.08861.
10. M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs.” arXiv, 2018. doi: 10.48550/ARXIV.1801.01401.
11. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” 2017, doi: 10.48550/ARXIV.1706.08500.
12. C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.
13. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” arXiv, 2015. doi: 10.48550/ARXIV.1505.04597.
14. N. Le, V. S. Rathour, K. Yamazaki, K. Luu, and M. Savvides, “Deep reinforcement learning in computer vision: a comprehensive survey,” *Artif Intell Rev*, vol. 55, no. 4, pp. 2733–2819, Apr. 2022, doi: 10.1007/s10462-021-10061-9.
15. C. A. Bouman and M. Shapiro, “A multiscale random field model for Bayesian image segmentation,” *IEEE Trans. on Image Process.*, vol. 3, no. 2, pp. 162–177, Mar. 1994, doi: 10.1109/83.277898.
16. D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes.” arXiv, 2013. doi: 10.48550/ARXIV.1312.6114.
17. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” arXiv, 2015. doi: 10.48550/ARXIV.1512.03385.

ABSTRACT

EVALUATING THE PERFORMANCE OF DIFFUSION MODELS THROUGH LOSS FUNCTION SELECTION

Hoàng Thế Anh¹, Trần Đình Toàn^{1,*}

¹ *Ho Chi Minh City University of Industry and Trade*

*Email: toantd@huit.edu.vn

Date received: 20/3/2025; Date accepted: 20/5/2025

The performance of diffusion based image generation models is highly dependent on the choice of loss function. However, the widespread use of MSE loss in training these models still presents significant limitations regarding the quality of the generated images. This study evaluates the impact of substituting various loss functions in a diffusionbased image generation model built on the UNet architecture. Experiments on the MNIST dataset show that replacing MSE with CharbonnierLoss significantly improves image quality, as reflected by FID and KID metrics. The results demonstrate that selecting an appropriate loss function plays a crucial role in enhancing the model's image generation performance.

Keywords: Diffusion Model, Image Generation, Loss Function, FID, KID.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Hoàng Thế Anh

Số điện thoại liên hệ: 0366491109

NGHIÊN CỨU THUẬT TOÁN KHAI THÁC TẬP HỮU ÍCH CAO CHÉO (CROSS-LEVEL HUI) TRÊN CƠ SỞ DỮ LIỆU PHÂN CẤP

Vũ Văn Vinh^{1*}, Phạm Tấn Thuận¹, Khúc Hồng Nhung¹, Trần Ngọc Hiếu Thảo¹.

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: vinhvv@huit.edu.vn

Ngày nhận bài: 10/04/2025 ; Ngày chấp nhận đăng: 20/05/2025

TÓM TẮT

Trong thời đại công nghệ số ngày nay, dữ liệu đã trở thành nguồn tài nguyên quý giá không thể thiếu. Khai thác triệt để sức mạnh tiềm ẩn của dữ liệu vẫn là một thách thức đầy khó khăn, đặc biệt là với khối lượng dữ liệu khổng lồ và không ngừng gia tăng. Trong lĩnh vực khai phá dữ liệu, một trong những "đỉnh núi" phức tạp nhất cần phải chinh phục chính là khai thác Cross-Level High Utility Itemsets (CLHUIs), tìm kiếm các tập mục có lợi nhuận cao trong các danh mục sản phẩm khác nhau. Khai thác CLHUI nhằm phát hiện những mối liên hệ tiềm ẩn giữa sản phẩm và danh mục, đề xuất các kết hợp mang lại lợi nhuận tối đa cho doanh nghiệp, hỗ trợ đưa ra quyết định chiến lược trong việc phát triển, từ đó nâng cao khả năng cạnh tranh trên thị trường và trải nghiệm của người tiêu dùng. Nghiên cứu này tập trung vào việc biến đổi dữ liệu ban đầu về dạng chứa các mục trừ tượng, đề xuất cách xây dựng cấu trúc CUL và áp dụng thuật toán HMiner để cải tiến và tối ưu hóa hiệu suất khai phá tập CLHUIs một cách hiệu quả hơn cả về thời gian thực thi và bộ nhớ tiêu tốn.

Từ khóa: tập hữu ích cao chéo, tập phổ quát, giao dịch đóng

1. GIỚI THIỆU

Khai thác dữ liệu là quá trình sắp xếp thông qua các tập dữ liệu lớn để xác định các mẫu và mối quan hệ có thể giúp giải quyết các vấn đề kinh doanh thông qua phân tích dữ liệu. Các kỹ thuật và công cụ khai thác dữ liệu giúp doanh nghiệp dự đoán xu hướng trong tương lai và đưa ra quyết định kinh doanh sáng suốt hơn. Khai thác dữ liệu là một phần quan trọng của phân tích dữ liệu và là một trong những nguyên tắc cốt lõi của khoa học dữ liệu, sử dụng các kỹ thuật phân tích nâng cao để tìm thông tin hữu ích trong tập dữ liệu. Ở cấp độ chi tiết hơn, khai thác dữ liệu là một bước trong quy trình khám phá kiến thức trong cơ sở dữ liệu (KDD), một phương pháp khoa học dữ liệu để thu thập, xử lý và phân tích dữ liệu. Khai thác dữ liệu và KDD đôi khi được gọi thay thế cho nhau, nhưng chúng thường được coi là những thứ riêng biệt hơn.

Quá trình khai thác dữ liệu phụ thuộc vào việc thực hiện hiệu quả việc thu thập, lưu trữ và xử lý dữ liệu. Khai thác dữ liệu có thể được sử dụng để mô tả tập dữ liệu mục tiêu, dự đoán kết quả, phát hiện các vấn đề gian lận hoặc bảo mật, tìm hiểu thêm về cơ sở người dùng hoặc phát hiện các tắc nghẽn và phụ thuộc. Khai thác dữ liệu cũng có thể được thực hiện tự động hoặc bán tự động. Khai thác dữ liệu ngày nay hữu ích hơn do sự phát triển của dữ liệu lớn và kho dữ liệu. Các chuyên gia dữ liệu sử dụng khai thác dữ liệu phải có kinh nghiệm về ngôn ngữ lập trình và mã hóa, cũng như kiến thức thống kê để làm sạch, xử lý và giải thích dữ liệu.

Phát hiện các thông tin có giá trị từ các cơ sở dữ liệu khác nhau đã nhận được sự quan tâm đáng kể, dẫn đến sự gia tăng mạnh mẽ các chủ đề nghiên cứu liên quan. Trong số các chủ đề, bài toán khai thác tập mục có lợi ích cao là một trong những chủ đề quan trọng nhất, xuất phát từ bài toán khai thác tập mục phổ biến nổi tiếng.

Khai phá tập phổ biến là xác định các tập mục xuất hiện thường xuyên trong các giao dịch trong cơ sở dữ liệu. Tần suất của một tập mục được đo bằng sự hỗ trợ của tập mục đó, tức là số lượng giao dịch có chứa tập mục đó. Nếu độ hỗ trợ của một tập mục vượt quá ngưỡng hỗ trợ tối thiểu do người dùng chỉ định thì tập mục đó được coi là thường xuyên.

Việc khai thác các tập phổ biến chỉ tính đến sự hiện diện và vắng mặt của các mục. Các thông tin khác về các hạng mục không được xem xét, chẳng hạn như tiện ích độc lập của một hạng mục và tiện ích ngữ cảnh của một hạng mục trong giao dịch. Thông thường, trong cơ sở dữ liệu siêu thị, mỗi mặt hàng có giá/lợi nhuận riêng biệt và mỗi mặt hàng trong giao dịch được liên kết với một số lượng riêng biệt, nghĩa là số lượng mặt hàng được mua. Để tính toán độ hỗ trợ, thuật toán chỉ sử dụng thông tin của hai cột đầu tiên trong bảng giao dịch, thông tin của cả bảng tiện ích và các cột khác trong bảng giao dịch đều bị loại bỏ. Tuy nhiên, một tập mục có độ hỗ trợ cao có thể có độ hữu dụng thấp hoặc ngược lại.

Trong một số ứng dụng, chẳng hạn như phân tích thị trường, người ta có thể quan tâm nhiều hơn đến tiện ích hơn là sự hỗ trợ của các tập mục. Các thuật toán khai thác tập mục thường xuyên truyền thống không thể đánh giá thông tin tiện ích về tập mục. Giống như các tập phổ biến, các tập mục có tiện ích không nhỏ hơn ngưỡng tiện ích tối thiểu do người dùng chỉ định thường có giá trị thú vị và chúng được gọi là tập mục có tiện ích cao (HUIs).

Mặc dù HUIM [1] có nhiều ứng dụng nhưng bài báo đã bỏ qua việc các mục thường được sắp xếp theo phân loại. HUIM truyền thống chỉ có thể tìm thấy các mẫu liên quan đến các mục ở mức độ trừu tượng thấp nhất. Đây là một vấn đề lớn vì mặc dù một số mặt hàng như gà rán và pepsi có thể không xuất hiện trong HUI nhưng mặt hàng đồ ăn nhanh đó có thể là một phần của một số HUI. Do đó, thông tin quan trọng có thể bị thuật toán HUIM truyền thống bỏ sót.

Một thuật toán có tên ML-HUI Miner [2] để khai thác HUI trong đó các mục có mức độ trừu tượng khác nhau. Tuy nhiên, thuật toán có hai hạn chế quan trọng. Đầu tiên, thuật toán chỉ có thể tìm thấy các HUI trong đó tất cả các mục có cùng mức độ trừu tượng. Thứ hai, thuật toán sử dụng các thuộc tính đơn giản để giảm không gian tìm kiếm. ML-HUI Miner khai thác các mức độ trừu tượng khác nhau một cách độc lập và không sử dụng mối quan hệ giữa các mức độ trừu tượng để giảm không gian tìm kiếm. Để giải quyết những hạn chế này, bài viết này xác định vấn đề tổng quát hơn trong việc khai thác tất cả các tập mục hữu ích cao chéo cấp và các thuộc tính của nó, đồng thời đề xuất một thuật toán có tên CLH-Miner [3] để tìm tất cả các tập mục này một cách hiệu quả.

2. CÁC CÔNG TRÌNH LIÊN QUAN

Khai thác tập hữu ích cao (High-Utility Itemset Mining - HUIM) đã được nghiên cứu sâu rộng như một kỹ thuật khai phá dữ liệu tiên tiến nhằm tìm ra các tập hợp mục có giá trị hữu ích cao (ví dụ: lợi nhuận, tầm quan trọng) thay vì chỉ tập trung vào các tập hợp thường xuyên như trong Khai thác tập thường xuyên truyền thống (Frequent Itemset Mining - FIM). Năm 2005, Liu và các cộng sự [1] giới thiệu với thuật toán Two-Phase, Tseng và các cộng sự [4] đề xuất thuật toán HUI-Miner vào năm 2012, đã thiết lập các phương pháp nền tảng bằng cách giới thiệu khái niệm hữu ích và các chiến lược cắt tỉa hiệu quả như TWU (Transaction-Weighted Utility) và cấu trúc danh sách hữu ích (utility-list). Tuy nhiên, các phương pháp này chủ yếu tập trung vào các tập hợp mục ở một cấp độ duy nhất, giả định rằng tất cả các mục thuộc cùng một mức độ chi tiết, điều này hạn chế khả năng áp dụng trong các kịch bản thực tế nơi các mục thường được tổ chức theo phân cấp (ví dụ: phân loại trong bán lẻ hoặc y tế).

Để giải quyết dữ liệu phân cấp, Khai thác tập hữu ích cao chéo cấp (Cross-Level High-Utility Itemset Mining - Cross-Level HUIM) đã xuất hiện như một phần mở rộng của HUIM, nhằm khám phá các mẫu hữu ích cao chứa các mục ở các cấp độ trừu tượng khác nhau. Một trong những nỗ lực tiên phong trong hướng đi này là công trình của Wu và các cộng sự [3], người đã đề xuất thuật toán CLH-Miner. CLH-Miner tích hợp một cây phân loại (taxonomy) vào quá trình khai thác, cho phép phát hiện các tập hợp mục hữu ích cao ở nhiều cấp độ (ví dụ: "Sữa → Sữa tách béo" hoặc "Đồ điện tử → Điện thoại thông minh"). Phương pháp của CLH-Miner tận dụng chiến lược theo từng cấp độ kết hợp với cắt tia dựa trên hữu ích để giảm không gian tìm kiếm. Các phương pháp Cross-Level HUIM hiện tại vẫn đối mặt với thách thức về khả năng mở rộng và tính linh hoạt. Chẳng hạn, CLH-Miner gặp khó khăn với các phân loại lớn do việc liệt kê toàn bộ các kết hợp chéo cấp độ một cách đầy đủ, trong khi CHUI-Mine, dù hiệu quả hơn, lại phụ thuộc nhiều vào các phân cấp được xác định trước, khiến nó kém thích nghi với các cấu trúc phân loại động hoặc không đầy đủ. Các nghiên cứu gần đây, như của Krishnamoorthy [5], Zida [6],... đã khám phá các kỹ thuật dựa trên chiếu (projection-based) để cải thiện hiệu suất, nhưng vẫn thiếu các cơ chế mạnh mẽ để xử lý nhiễu hoặc dữ liệu bị thiếu trong các tập dữ liệu thực tế.

Song song với Cross-Level HUIM, Khai thác tập hữu ích cao đa cấp (Multilevel High-Utility Itemset Mining - Multilevel HUIM) đã thu hút sự chú ý như một cách tiếp cận chuyên biệt để khám phá các tập hợp mục hữu ích cao trong một cấu trúc phân cấp, tập trung vào các mẫu ở các cấp độ trừu tượng cụ thể thay vì xuyên suốt chúng. Một trong những đóng góp sớm nhất trong lĩnh vực này là công trình của Han và các cộng sự [6], người đã mở rộng thuật toán Apriori để khai thác các tập hợp mục thường xuyên đa cấp, sau đó được Yao và các cộng sự điều chỉnh cho khai thác hữu ích với thuật toán ML-HUI. ML-HUI áp dụng chiến lược từ trên xuống, khai thác các tập hợp mục theo từng cấp độ trong một phân loại (ví dụ: từ "Đồ uống" đến "Sữa" đến "Sữa tách béo"), và sử dụng các ngưỡng hữu ích được điều chỉnh cho từng cấp độ để lọc các mẫu quan trọng. Phương pháp này đảm bảo rằng các tập hợp mục hữu ích cao được xác định ở các mức độ chi tiết khác nhau, cung cấp cái nhìn sâu sắc vào cả xu hướng chung và cụ thể. Những tiến bộ tiếp theo bao gồm công trình của Shen và các cộng sự [8], người đã đề xuất thuật toán MHUI-Tree, sử dụng một cấu trúc cây hữu ích đa cấp để lưu trữ và truy xuất hiệu quả các tập hợp mục qua các cấp độ khác nhau. MHUI-Tree giảm thiểu các phép tính dư thừa bằng cách duy trì một biểu diễn nén của các giá trị hữu ích phân cấp, đạt được khả năng mở rộng tốt hơn cho các tập dữ liệu lớn. Tuy nhiên, các phương pháp Multilevel HUIM thường giả định một phân cấp cố định và các ngưỡng hữu ích đồng nhất, điều này có thể không nắm bắt đầy đủ các mối quan hệ động giữa các mục qua các cấp độ. Ví dụ, một tập hợp mục hữu ích cao ở cấp độ thấp hơn (như "Sữa tách béo") có thể bị bỏ qua nếu cấp độ cha (như "Sữa") không đáp ứng ngưỡng. Các nỗ lực gần đây, như của Zhang và các cộng sự, đã giới thiệu các kỹ thuật ngưỡng linh hoạt để giải quyết vấn đề này, nhưng vẫn còn thách thức trong việc cân bằng giữa hiệu quả tính toán và tính đầy đủ của các mẫu được khai thác. Những hạn chế này trong cả Cross-Level HUIM và Multilevel HUIM nhấn mạnh nhu cầu về một cách tiếp cận tích hợp mà nghiên cứu của chúng tôi hướng tới giải quyết.

3. CƠ SỞ LÝ THUYẾT

3.1 Một số khái niệm.

3.1.1 Định nghĩa 1 (Cơ sở dữ liệu giao dịch)

Cho $I = \{i_1, i_2, \dots, i_m\}$ là một tập hợp các itemset. Một cơ sở dữ liệu giao dịch là một tập hợp nhiều giao dịch $D = \{T_1, T_2, \dots, T_n\}$ sao cho với mỗi giao dịch T_c , $T_c \in I$, và T_c có một định danh duy nhất c được gọi là T_{id} .

Với mỗi mục $i \in I$, một số dương $p(i)$ được gọi là hữu ích ngoài của i .

Với mỗi mục $i \in T_c$, số dương $q(i, T_c)$ được gọi là lợi ích nội của i trong giao dịch T_c .

Bảng 1. Cơ sở dữ liệu giao dịch D

t_{id}	Transaction
T_1	(a, 1); (c, 2); (g, 3)
T_2	(c, 2); (e, 3); (f, 2)
T_3	(a, 2); (b, 1); (c, 3); (d, 2); (e, 1); (f, 3)
T_4	(b, 2); (c, 1); (f, 3)
T_5	(a, 3); (b, 2); (f, 1)
T_6	(c, 2); (d, 5); (f, 2); (g, 2)
T_7	(a, 4); (c, 1); (d, 2); (e, 2)

Bảng 2. Bảng giá trị hữu ích ngoại của CSDL giao dịch D

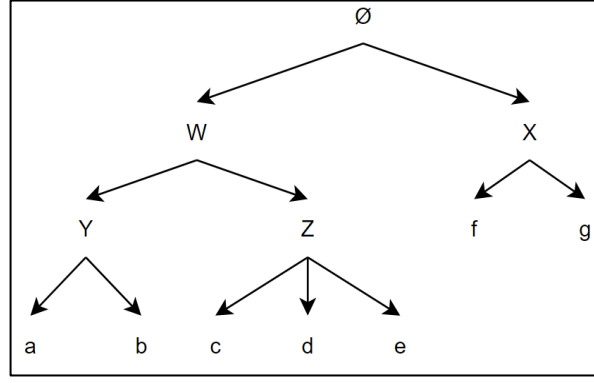
Item	a	b	c	d	e	f	g
EU	4	3	5	2	2	1	3

Dựa vào bảng 1 và 2 sẽ được sử dụng cho tất cả ví dụ trong bài báo này.

Ví dụ: $p(a) = 4$; $p(c) = 5$; $q(a, T_1) = 41$; $q(c, T_1) = 52$

3.1.2 Định nghĩa 2 (Taxonomy)

Phân loại τ là một biểu đồ tuần hoàn có hướng (tree) được xác định cho cơ sở dữ liệu D. Mỗi lá là một mục $i \in I$. Một nút bên trong biểu thị một danh mục trừu tượng tổng hợp tất cả các nút lá con cháu (item) hoặc các danh mục con cháu thành một hạng cấp cao hơn. Cạnh con-cha giữa hai mục (tổng quát) i, j trong τ biểu thị mối quan hệ is-a. Các nút bên trong được gọi là các mục tổng quát (GI) hoặc các mục trừu tượng (AI). Tập tất cả các mục tổng quát được ký hiệu là GI, và tập hợp tất cả các mục tổng quát hoặc mục lá được ký hiệu là $AI = GI \cup I$. Ngoài ra, giả sử tồn tại một quan hệ $LR \subseteq GI \times I$ sao cho $(g, i) \in LR$ nếu có đường đi từ g đến i . Và, giả sử tồn tại một quan hệ $GR \subseteq AI \times AI$ sao cho $(d, f) \in GR$ nếu có đường đi từ d đến f . Tập mục X là tập các mục sao cho $X \subseteq AI$ và $\exists i, j \in X | i \in Desc(j, \tau)$. Tập mục X là tập mục tổng quát nếu $\exists g \in X$ sao cho $g \in GI$. Có nghĩa là tập mục X sẽ chứa nhiều mục con nhưng trong đó không có mục con nào có quan hệ GR với nhau.



Hình 1. Cây phân loại τ

$$AI = GI \cup I$$

$$LR \subseteq GI \times I \quad (W, a) \in LR; (w, f) \notin LR$$

$$GR \subseteq AI \times AI \quad (W, a) \in GR \text{ j } (W, X) \notin GR$$

3.1.3 Định nghĩa 3 (Descendant/Specialization/Sibling)

Các mục lá của mục tổng quát g trong phân loại τ là tất cả các lá có thể đạt được bằng cách đi theo các đường dẫn bắt đầu từ g. Tập hợp này được định nghĩa chính thức là $\text{Leaf}(g, \tau) = \{i | (g, i) \in LR\}$. Các phần tử con (Descendant) của một phần tử (GI) d là tập $\text{Desc}(d, \tau) = \{f | (d, f) \in GR\}$. Một tập mục X được gọi là sự chuyên biệt hóa (Specialization) của tập mục Y nếu $|X| = |Y|$ và $\forall f \in X, f \in Y \vee \exists d \in Y | f \in \text{Desc}(d, \tau)$. Hai phần tử a và b được gọi là anh em ruột (Sibling) nếu chúng có cùng cha mẹ trong τ . Hơn nữa, $\text{level}(d)$ biểu thị số cạnh cần đi qua để đạt được mục d bắt đầu từ gốc của τ .

Ví dụ:

$$\text{Left}(W) = \{a, b, c, d, e\}, \quad \text{Left}(X) = \{f, g\}$$

$$\text{Desc}(W, \tau) = \{Y, Z, a, b, c, d, e\}$$

W_c được gọi là tập mục con (Specialization) của WZ

Y và Z được gọi là anh em ruột (Sibling)

$$\text{Level}(W) = 1, \text{Level}(Y) = 2$$

3.1.4 Định nghĩa 4 (Utility of an item/itemset)

Giá trị hữu ích của một mục i trong một giao dịch T_c được ký hiệu là $u(i, T_c)$ và được định nghĩa là $p(i) \times q(i, T_c)$. Giá trị hữu ích của một tập hợp các mục X (một nhóm các mục $X \subseteq I$) trong một giao dịch T_c được ký hiệu là $u(X, T_c)$ và được định nghĩa là $u(X, T_c) = \sum_{i \in X} u(i, T_c)$. Giá trị hữu ích của một tập hợp các mục X (trong cơ sở dữ liệu) được ký hiệu là $u(X)$ và được định nghĩa là $u(X) = \sum_{T_c \in g(X)} u(X, T_c)$ trong đó $g(X)$ là tập hợp các giao dịch chứa X.

3.1.5 Định nghĩa 5 (Utility of a generalized item/itemset)

Giá trị hữu ích của một mục P tổng quát g trong một giao dịch T_c được ký hiệu và định nghĩa là $u(g, T_c) = \sum_{i \in \text{Leaf}(g, \tau)} p(i) \times q(i, T_c)$. Giá trị hữu ích của một tập hợp mục tổng quát GX trong một giao dịch T_c được định nghĩa là $u(GX, T_c) = \sum_{d \in GX} u(d, T_c)$. Giá trị hữu ích

của một tập hợp mục tổng quát GX (trong cơ sở dữ liệu) được ký hiệu và định nghĩa là $u(GX) = \sum_{T_c \in g(GX)} u(GX, T_c)$, trong đó $g(GX) = \{T_c \in D | \exists X \subseteq T_c \wedge X \text{ là một phần tử hạ dẫn của } GX\}$

3.1.6 Định nghĩa 6 (Cross-level high utility mining)

Bài toán khai thác tập hữu ích cao liên mức được định nghĩa là tìm tất cả các tập hữu ích cao liên mức (CLHUIs - cross-level high utility itemsets). Một tập mục (tổng quát) X là một CLHUI nếu $u(X) \geq \text{minutil}$ với ngưỡng minutil được người dùng chỉ định.

3.1.7 Định nghĩa 7 (Extension)

Giả sử X là một tập mục. Các mở rộng dựa trên phép nối (join-based extensions) của X là các tập mục thu được bằng cách thêm một mục y vào X sao cho $y \in AI$, $y \succ i, \forall i \in X, y \notin \text{Desc}(i, \tau)$. Các mở rộng dựa trên phép thuế (tax-based extensions) của X là các tập mục thu được bằng cách thay thế mục cuối cùng y của X theo $\succ$ một mục hạ dẫn của y.

3.1.8 Định nghĩa 8 (Tax-Utility-list)

Danh sách tiện ích thuế $tuList(X)$ của X chứa một bộ dữ liệu (tid, iutil, rutil) cho mỗi giao dịch Ttid chứa X. Phần tử iutil của bộ dữ liệu là tiện ích của X trong Ttid, tức là $u(X, T_{tid})$. Phần tử rutil của một bộ dữ liệu được định nghĩa là $\sum_{i \in T_{tid} \wedge i <_x \forall x \in X} u(X, T_{tid})$. Ngoài ra, $tuList(X)$ chứa các mục con, một tập hợp các con trở tới danh sách tiện ích thuế của các mục con của X trong phân loại.

3.1.9 Định nghĩa 9 (GWU)

Tiện ích giao dịch (TU) của giao dịch T_c là tổng hữu dụng của tất cả các mục trong T_c . tức là $TU(T_c) = \sum_{x \in T_c} u(x, T_c)$. Việc sử dụng trọng số tổng quát (GWU) của tập mục (tổng quát) $X \subseteq AI$ được định nghĩa là tổng tiện ích giao dịch của các giao dịch chứa X, tức là $GWU(X) = \sum_{T_c \in g(X)} TU(T_c)$.

3.2 Một số tính chất

3.2.1 Tính chất 1 (Tính đơn điệu GWU):

Giả sử có hai tập mục X và Y sao cho Y là một chuyên biệt hóa của siêu tập hợp X. GWU của X không nhỏ hơn tiện ích của X ($GWU(X) \geq u(X)$). Hơn nữa, GWU của X không kém gì tiện ích của các siêu tập hợp của X ($GWU(X) \geq u(Y) \forall Y \supset X$). Bằng chứng. Nếu $Y \subset X$ thì $u(Y) \leq GWU(Y) \leq GWU(X) < \text{minutil}$. Nếu Y là một chuyên biệt hóa của X thì $u(Y) \leq u(X) \leq GWU(X) < \text{minutil}$.

3.2.2 Tính chất 2 (Cắt tỉa không gian tìm kiếm bằng GWU):

Đối với bất kỳ tập mục X nào, nếu $GWU(X) < \text{minutil}$, thì X là tập mục có tiện ích thấp cũng như tất cả các phần mở rộng dựa trên phép nối và dựa trên thuế của X.

3.2.3 Tính chất 3 (Cắt tỉa không gian tìm kiếm bằng tiện ích còn lại):

Với tổng thứ tự $\prec$ trên một phân loại, giới hạn trên tiện ích còn lại của tập mục X là tổng của các giá trị iutil và rutil trong tax-utility-list của X. Về mặt hình thức, giới hạn trên này được định nghĩa là $\text{reu}(X) = \sum_{e \in tuList(X)} (e.iutil + e.rutil)$. Nếu $\text{reu}(X) < \text{minutil}$ thì X và tất cả các tiện ích mở rộng dựa trên thuế của X không phải là CLHUI

4. THUẬT TOÁN ĐỀ XUẤT

4.1 Ý tưởng

Trong thuật toán CLH-Miner sử dụng cấu trúc Tax Utility List và chiến lược cắt tia GWU và RU để giảm bớt không gian tìm kiếm CLHUIs. Điều này khá tương đồng với cấu trúc Utility List và chiến lược cắt tia TWU, RU của thuật toán HUIMiner để khai thác HUIs. Ở hướng khai thác HUIs có một thuật toán là HMiner đã sử dụng cấu trúc CUL và EUCS (của thuật toán FHM) kết hợp cùng với nhiều chiến lược cắt tia để giảm bớt không gian tìm kiếm rất hiệu quả. Trong hướng khai thác MLHUIs, N.T.Tung cùng các cộng sự đã áp dụng cấu trúc này để khai thác MLHUIs một cách hiệu quả và tối ưu, đó là thuật toán MLHMiner. Như vậy, chúng ta cũng có thể áp dụng cấu trúc CUL và EUCS cùng các chiến lược cắt tia trong khai phá CLHUIs. Ngoài chiến lược cắt tia GWU và RU của thuật toán CLH-Miner thì MLHMiner còn sử dụng thêm các chiến lược cắt tia khác: EUCS để xem xét 2 tập mục khi kết hợp với nhau nếu nhỏ hơn minUtil thì loại bỏ các ứng viên ngay mà không cần phải tính toán thêm, chiến lược cắt tia LA-prune và C-prune để loại bỏ ngay các tập kết hợp trong quá trình tính toán.

Bước đầu tiên, thuật toán xây dựng lại CSDL giao dịch mới từ CSDL giao dịch cũ bổ sung thêm những tập mục GI trực tiếp vào giao dịch để có cái nhìn trực quan hơn. Trong đó, những tập mục lá được biểu diễn bằng kí tự thường, còn những mục GI được biểu diễn bằng chữ in hoa. Mỗi tập mục ta biểu diễn bằng tên và lợi nhuận trong giao dịch hiện tại thay vì tên và số lượng.

Bảng 3. Cơ sở dữ liệu sau khi biến đổi

T _{id}	Transaction
T ₁	(X, 9); (W, 8); (Y, 4); (Z, 10); (a, 4); (c, 10)
T ₂	(X, 2); (W, 16); (f, 2); (Z, 16); (c, 10);
T ₃	(X, 3); (W, 32); (f, 3); (Y, 11); (Z, 21); (d, 4); (a, 8); (c, 15);
T ₄	(X, 3); (W, 11); (f, 3); (Y, 6); (Z, 5); (c, 5);
T ₅	(X, 1); (W, 18); (f, 1); (Y, 18); (a, 12);
T ₆	(X, 8); (W, 20); (f, 2); (Z, 20); (d, 10); (c, 10);
T ₇	(W, 29); (Y, 16); (Z, 13); (d, 4); (a, 16); (c, 5);

4.2 Một số khái niệm

4.2.1 Định nghĩa 10 (Prefix utility):

Cho một tập mục X và một mục mở rộng $y \in I$, tiện ích tiền tố của tập mục Xy trong giao dịch T_j được định nghĩa: $PU(Xy, T_j) = U(X, T_j)$. Nếu $X = \emptyset$ thì $PU(y, T_j) = 0$

Giả sử ta có thứ tự sắp xếp: $X < W < g < f < Y < Z < b < e < d < a < c$, ta được:

$$PU(Y, T_1) = 0; \quad PU(c, T_1) = 0$$

$$PU(XY, T_1) = U(X, T_1) = U(a, T_1) = 4.$$

4.2.2 Định nghĩa 11 (Closed prefix utility)

Tiền ích tiền tố đóng của tập mục $X = \{x_1, x_2 \dots x_k\}$ trong giao dịch T_j , ký hiệu là $CPU(X, T_j)$, được định nghĩa là:

$$CPU(X, T_c) = \begin{cases} PU(X, T_c), & \text{if } |X| > 1, |E(X - x_k)| = |T_c/X - x_k| \\ 0 & \end{cases}$$

Trong đó,

$|E(X - x_k)|$ là số lượng mục phía sau của X và có thể mở rộng trong cơ sở dữ liệu.

$|T_c/X - x_k|$ là số lượng mục phía sau của X và có thể mở rộng trong giao dịch T_c .

Hữu ích tiền tố đóng của tập mục X có số lượng mục nhỏ hơn hoặc bằng 2 trong cơ sở dữ liệu được định nghĩa là $CPU(X) = \sum_{T_c \in G(X)} CPU(X, T_c)$.

Hữu ích tiền tố đóng của tập mục X có số lượng mục lớn hơn 2 trong cơ sở dữ liệu được ký hiệu là $CPU(X)$ và định nghĩa là $CPU(Xyz) = CU(Xy)$.

4.2.3 Định nghĩa 12 (Hữu ích đóng)

Hữu ích đóng của tập mục X trong giao dịch T_c được ký hiệu là $CU(X, T_c)$ và tính toán như sau:

$$CU(X, T_c) = \begin{cases} U(X, T_c), & \text{if } |X| > 1, |E(X - x_k)| = |T_c/X - x_k| \\ 0 & \end{cases}$$

Hữu ích đóng của tập mục X có số lượng mục nhỏ hơn hoặc bằng 2 trong cơ sở dữ liệu được định nghĩa là $CU(X) = \sum_{T_c \in G(X)} CU(X, T_c)$.

Hữu ích đóng của tập mục X có số lượng lớn hơn 2 trong cơ sở dữ liệu được định nghĩa là $CU(Xyz) = CU(Xy) + CU(Xz) - CPU(Xy) + \sum_{T_c \in G(Xyz)} CU(Xyz, T_c)$.

4.2.4 Định nghĩa 13

Hữu ích còn lại đóng của tập mục X trong giao dịch T_c được ký hiệu là $CRU(X, T_c)$ được định nghĩa và tính toán như sau:

$$CRU(X, T_c) = \begin{cases} RU(X, T_c), & \text{if } |X| > 1, |E(X - x_k)| = |T_c/X - x_k| \\ 0 & \end{cases}$$

Hữu ích đóng của tập mục X có số lượng mục nhỏ hơn hoặc bằng 2 trong cơ sở dữ liệu được định nghĩa là $CRU(X) = \sum_{T_c \in G(X)} RU(X, T_c)$.

Hữu ích đóng của tập mục X có số lượng mục lớn hơn 2 trong cơ sở dữ liệu được định nghĩa là $CRU(Xyz) = CRU(Xz)$.

4.2.5 Định nghĩa 14

Các giá trị hữu ích không đóng của tập mục X có số lượng mục nhỏ hơn hoặc bằng 2 được định nghĩa và tính toán như sau:

$$NU(X) = U(X) - CU(X)$$

$$NRU(X) = RU(X) - CRU(X)$$

$$NPU(X) = PU(X) - CPU(X)$$

Các giá trị hữu ích không đóng của tập mục X có số lượng mục lớn hơn 2 được định nghĩa và tính toán như sau:

$$NU(Xyz) = NU(Xy) + NU(Xz) - NPU(Xy)$$

$$NRU(Xyz) = NRU(Xz)$$

$$NPU(Xyz) = NPU(Xy)$$

4.3 Cấu trúc dữ liệu

4.3.1 Cấu trúc EUCS

Vận dụng ý tưởng của ECUS được đề xuất trong thuật toán FHM, nghiên cứu sử dụng cấu trúc EUCS là một ma trận chứa GWU của 2 tập mục khi kết hợp với nhau. GWU của một tập mục là lợi nhuận tối đa mà tập mục này có thể mang lại trong CSDL. Tuy nhiên, để giảm bộ nhớ lưu trữ, cấu trúc ECUS sẽ được lưu theo cấu trúc gồm 3 danh sách như sau:

Bảng 4. Cấu trúc lưu trữ EUCS cải tiến

X	X	X	X	X	X	W	f	f	f	f	f	Y	Y	Y	Z	d	d	a
W	Y	Z	d	a	c	f	Y	Z	d	a	c	Z	d	c	a	a	c	c
137	91	118	63	77	118	114	68	95	63	54	95	101	64	101	87	64	92	87

Cấu trúc lưu trữ gồm 3 danh sách giúp tối ưu bộ nhớ để lưu trữ EUCS, đặc biệt với các cơ sở dữ liệu thưa.

4.3.2 Cấu trúc CUL

Cấu trúc CUL(X) là cấu trúc chứa NU(X), NRU(X), CU(X), CRU(X), CPU(X) và danh sách các bộ dữ liệu TidList chứa thông tin về Tid, NU(X, T_c), NRU(X, T_c), PU(X, T_c) và PPOS(X, T_c). PPOS(X, T_j) được xác định như sau:

$$PPOS(X, T_{c-i}) = \begin{cases} prev(xk) \neq -1, & |CUL[prev(xk, T_c)].tidList| \\ -1 & \end{cases}$$

Itemset X				
	NU	NRU	CU/CRU/CPU	
T _j	NU	NRU	PU	PPOS

Hình 2. Cấu trúc CUL

Phương pháp xây dựng 1-CUL sẽ được thực hiện như trong khai thác HUI. Tuy nhiên trong quá trình tìm các CLHUI ta cần lưu ý: Trước khi thực hiện kết hợp các tập mục lại với nhau, ta xét xem mục hiện tại có lợi nhuận có thỏa ngưỡng hay không? Nếu thỏa điều kiện ta đưa tập đang xét vào CLHUIS. Tiếp tục thuật toán kiểm tra tổng lợi nhuận và lợi nhuận của các phần tử phía sau của X có thỏa ngưỡng hay không? Nếu tổng lớn hơn, thuật toán thực hiện kết hợp các k-CUL để tạo ra (k+1)- CUL

Không giống khai phá như HUIs và MLHUIs, thuật toán chỉ cần kết hợp CUL đang xét với tất cả những tập mục phía sau nó. Khi khai phá CLHUIs, thuật toán cần xác định phần mở rộng của từng tập mục trước khi thực hiện kết hợp, vì cần phải loại bỏ các mục có quan hệ cha con. Đối với mục X, ta xác định được phần mở rộng là W, Y, Z, d, a, c. Đầu tiên, ta xây dựng sẵn danh sách exCTUs, pruneLA, pruneC, eyTid để trống. Tiếp theo, chúng ta duyệt qua phần mở rộng và kiểm tra EUCS xem khi X kết hợp với tập mục đang duyệt có nhỏ hơn minUtil không? Trường hợp nhỏ hơn, ta gán null cho exCTUs ngay lập tức (không kết hợp). Đối với danh sách pruneLA ta cộng toàn bộ lợi nhuận và lợi nhuận phía sau X. Thuật toán tiếp tục duyệt qua từng giao dịch của CUL(X) và áp dụng chiến lược tỉa LA-prune và C-prune để giảm bớt không gian tìm kiếm.

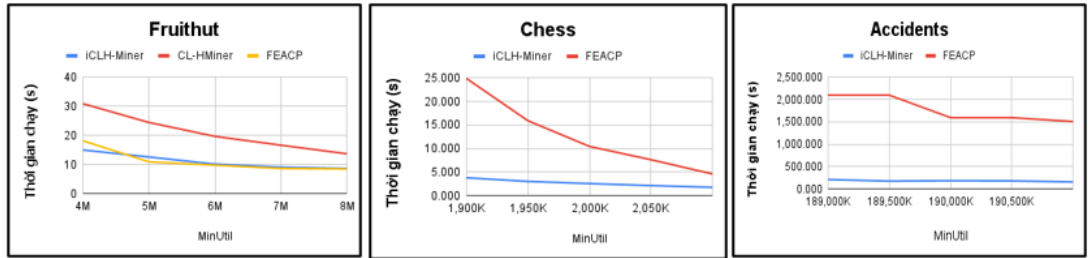
5. THỰC NGHIỆM

Để đánh giá hiệu suất của thuật toán CLH-Miner cải tiến, các thử nghiệm đã được thực hiện trên máy tính có bộ xử lý Intel Core-i5 và RAM 16 GB, chạy hệ điều hành Windows 11. Chúng em đã so sánh hiệu suất của CLH-Miner đã cải tiến về thời gian thực hiện và mức sử dụng bộ nhớ cao nhất để khai thác CLHUI với CLH-Miner cũ và thuật toán FEACP. Tuy nhiên, vì thời gian khai thác CLHUI của thuật toán CLH-Miner quá lớn trong các CSDL dày nên thực nghiệm chỉ so sánh được trên bộ dữ liệu Fruithut.

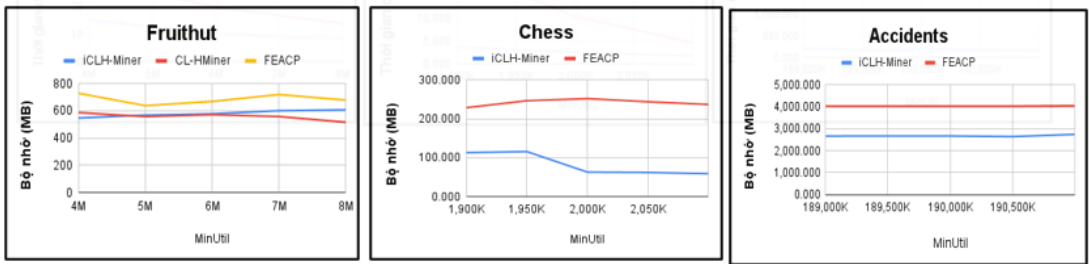
Bảng 5. Đặc điểm của các CSDL thực nghiệm

CSDL	Số giao dịch	Số mục	GI	Level	Tmax	Tavg	Mật độ
Fruithut	181,970	1256	43	4	36	3.58	Thưa
Chess	3,196	75	30	3	37	37.00	Dày
Accidents	340,183	468	216	6	51	33.80	Dày

Kết quả thực nghiệm được trình bày trong biểu đồ sau



Hình 3. Biểu đồ so sánh về thời gian thực thi



Hình 4. Biểu đồ so sánh bộ nhớ sử dụng

Kết quả cho thấy rằng trên những cơ sở dữ liệu mà mỗi giao dịch có số lượng sản phẩm lớn thì thuật toán cải tiến đạt hiệu quả cao hơn rất nhiều so với các thuật toán trước như CLH-Miner và FEACP. Tuy nhiên, trên những cơ sở dữ liệu có ít sản phẩm trên một giao dịch như fruithut thì thuật toán cải tiến cũng chưa đạt được hiệu quả cao hơn so với các thuật toán trước đó. Vậy có thể thấy rằng thuật toán cải tiến đạt hiệu quả cao đối với những cơ sở dữ liệu dày hơn.

6. KẾT LUẬN

Nghiên cứu đã tìm hiểu được các khái niệm liên quan tới khai thác tập hữu ích cao chéo và các khái niệm liên quan. Nghiên cứu đã tìm hiểu phương pháp khai thác tập hữu ích cao chéo và đề xuất được phương pháp để nâng cao hiệu quả khai thác CLHUIs hiệu quả hơn cả

về mặt thời gian và bộ nhớ, đặc biệt là với CSDL thưa. Trong thời gian tới, chúng tôi sẽ tập trung vào việc cải tiến thuật toán để việc khai thác CLHUIs có kết quả tốt hơn với cơ sở dữ liệu dày và cũng xem xét việc áp dụng khai thác CLHUIs trên cơ sở dữ liệu mà các mục có lợi ích thay đổi hoặc cơ sở dữ liệu tồn tại các mục chưa lợi ích âm.

TÀI LIỆU THAM KHẢO

1. Y. Liu, W. Liao, A. Choudhary, "A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets," in PAKDD 2005, HaNoi, VietNam, 2005.
2. Trinh D. D. Nguyen, Loan T. T. Nguyen, Adrianna Kozierekiewicz, Thiet Pham, Bay Vo , "An Efficient Approach for Mining High-Utility Itemsets from Multiple Abstraction Levels," Intelligent Information and Database Systems, vol. 12672, p. 92–103, 2021.
3. Fournier-Viger, P., Wang, Y., Lin, J.CW., Luna, J.M., Ventura, S, "Mining Cross-Level High Utility Itemsets," Trends in Artificial Intelligence Theory and Applications, vol. 12144, p. 858–871, 2020.
4. M. Liu, J. Qu, "Mining high utility itemsets without candidate generation," in CIKM, 2012.
5. S. Krishnamoorthy, "HMiner: Efficiently mining high utility itemsets," Expert Systems with Applications, vol. 90, pp. 168-183, 2017.
6. S. Zida, P. Fournier-Viger, J.C. Lin, C. Wu, V.S. Tseng, "EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining," Knowledge and Information Systems, vol. 51, no. 2, pp. 595-625 , 2017.

ABSTRACT

IMPROVE THE CROSS-LEVEL HIGH-UTILITY ITEMSET MINING ALGORITHMS ON HIERARCHICAL DATABASE

Vũ Văn Vinh^{1*}, Phạm Tấn Thuận¹, Khúc Hồng Nhung¹, Trần Ngọc Hiếu Thảo¹.

¹*Ho Chi Minh City University of Industry and Trade*

*Email: vinhvv@huit.edu.vn

ABSTRACT

Today, data has become an invaluable and indispensable resource. Fully harnessing the hidden power of data remains a significant challenge, especially given the ever-growing volume of information. In the field of data mining, one of the most complex ‘peak’ to conquer is the mining of Cross-Level High Utility Itemsets (CLHUIs), which involves identifying high-utility itemsets across different product categories. CLHUI mining aims to uncover hidden relationships between products and categories, suggesting combinations that maximize business profits and support strategic decision-making for development. This, in turn,

enhances market competitiveness and improves consumer experience. This study focuses on transforming the original data into an generated item-based format, proposes the construction of the CUL structure, and applies the HMiner algorithm to improve and optimize the efficiency of CLHUI mining in terms of both execution time and memory consumption

Keywords: cross-level high utility itemsets, generated itemset, completed transaction,

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Vũ Văn Vinh

Số điện thoại liên hệ: 0909 443 886

GIẢI PHÁP BẢO MẬT CHO MẠNG XÃ HỘI THU NHỎ

**Phạm Đình Văn¹, Nhiêu Hồng Phúc¹, Nguyễn Như Quang Tùng¹,
Trần Thị Vân Anh¹, Đinh Thị Mận^{1*}**

¹*Trường Đại học Công Thương Thành phố Hồ Chí Minh*

*Email: mandt@huit.edu.vn

Ngày nhận bài: 28/3/2025; Ngày chấp nhận đăng: 20/5/2025

TÓM TẮT

Việc sử dụng mạng xã hội đối với người dùng hiện nay còn có nhiều rủi ro về nguy cơ mất an toàn vì chưa quan tâm đến các chính sách bảo mật, tài khoản dễ bị tấn công, hồ sơ người dùng quản lý chưa được tối ưu hóa. Trong bài báo này, nhóm tác giả trình bày phương pháp xây dựng ứng dụng mạng xã hội thu nhỏ trên đa nền tảng (Web, Mobile) bên cạnh những tính năng thông dụng của một mạng xã hội như đăng bài viết, bình luận, nhắn tin,... ứng dụng được tích hợp các tính năng bảo mật gồm thiết lập các chế độ xác thực người dùng khi đăng nhập: sử dụng mạng học sâu nhận diện và xác thực, đăng nhập thông qua mã OTP, mã PIN, cảnh báo khi có đăng nhập từ hai thiết bị trở lên, quản lý tài khoản người dùng tùy chọn riêng tư hoặc công khai giúp người dùng trao đổi thông tin trên mạng xã hội an toàn. Ngoài ra, việc quản lý hồ sơ người dùng được tối ưu thông qua chức năng tạo Portfolio cá nhân với các tùy chọn giao diện khác nhau. Kết quả nghiên cứu góp phần tạo môi trường giao tiếp trên mạng xã hội an toàn, thông tin được bảo mật và quản lý hồ sơ hiệu quả.

Từ khóa: Mạng xã hội thu nhỏ, ResNet, T3 Stack, Portfolio.

1. GIỚI THIỆU

Các nhà phát triển ứng dụng thường gặp nhiều thách thức trong việc đồng bộ hóa giữa giao diện người dùng (frontend), logic xử lý phía máy chủ (backend) và cơ sở dữ liệu, đặc biệt đối với các hệ thống phức tạp. Vì vậy, lựa chọn công nghệ phù hợp để xây dựng ứng dụng đáp ứng các yêu cầu quản lý, tăng hiệu suất, có khả năng mở rộng và bảo trì là một số yếu tố quan trọng. T3 Stack là một tập hợp các công nghệ hiện đại, được sử dụng để xây dựng các ứng dụng web đạt hiệu suất cao, khả năng bảo trì tốt, phát triển nhanh chóng, an toàn và hiệu quả với cơ sở dữ liệu [1]. Do đó, trong bài báo này nhóm tác giả lựa chọn T3 Stack để phát triển ứng dụng.

Nhận dạng khuôn mặt nhằm xác minh danh tính của người dùng được ứng dụng trong nhiều lĩnh vực [2-4]. Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) được sử dụng rộng rãi trong các bài toán phân loại hình ảnh, nhận diện khuôn mặt, [5-7]... nhiều mô hình phân loại dựa trên CNN đã được phát triển như: AlexNet, VGGNet, GoogLeNet, ResNet, MobileNet và EfficientNet [5]. Trong các mạng CNN, vấn đề vanishing gradient hoặc exploding gradient thường xuất hiện trong quá trình huấn luyện. Hiện tượng này khiến cho gradient của các lớp trước đó giảm dần về 0 (đối với vanishing gradient) hoặc tăng lên quá lớn (đối với exploding gradient) khi quá trình huấn luyện diễn ra, gây khó khăn cho việc cập nhật

trọng số và dẫn đến hiệu suất mô hình kém. Để giải quyết vấn đề này, kiến trúc ResNet (Residual Network) sử dụng các khối residual và kết nối tắt (shortcut connections). Các kết nối tắt cho phép gradient truyền trực tiếp qua các lớp, bỏ qua các lớp trung gian nên huấn luyện các mạng nơ-ron sâu hơn và hiệu quả hơn và thường được ứng dụng trong các bài toán nhận diện khuôn mặt [5]. Vì vậy, trong bài báo này nhóm tác giả sử dụng mạng ResNet nhận diện khuôn mặt để xác thực người dùng giúp tăng tính bảo mật của ứng dụng.

Đóng góp chính của bài báo gồm:

- Sử dụng công nghệ T3 Stack xây dựng mạng xã hội thu nhỏ.
- Sử dụng mạng ResNet nhận diện khuôn mặt giúp xác thực người dùng.
- Tích hợp các tính năng bảo mật mật giúp người dùng trao đổi thông tin an toàn.
- Sử dụng Portfolio trong ứng dụng để quản lý hồ sơ người dùng hiệu quả.
- Phát triển ứng dụng mạng xã hội trên đa nền hỗ trợ người dùng tương tác, trao đổi thông tin, quản lý hồ sơ và bảo mật dữ liệu cá nhân, với khả năng triển khai trên cả web và mobile để nâng cao tính tiện dụng.

Phần còn lại của bài báo được trình bày gồm các phần như sau: phần 2 trình bày về cơ sở lý thuyết, phần 3 trình bày phương pháp xây dựng mạng xã hội thu nhỏ, phần 4 cài đặt thực nghiệm, phần 5 kết luận.

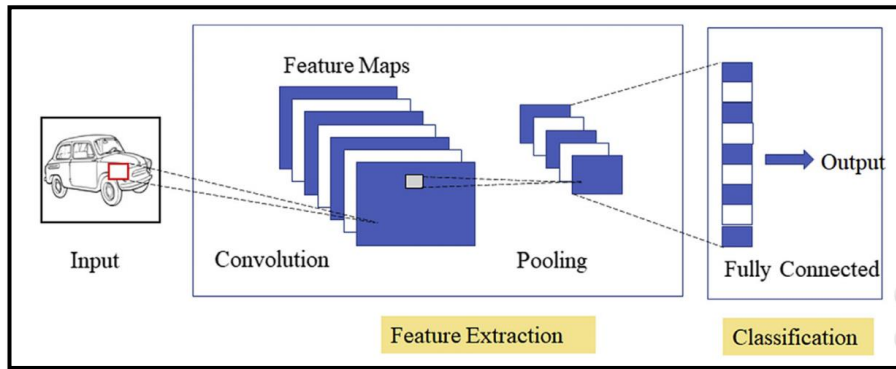
2. CƠ SỞ LÝ THUYẾT

2.1. T3 Stack

Trong T3 Stack, TypeScript, Next.js, Tailwind CSS, tRPC, Prisma, Supabase là các công nghệ và thư viện được kết hợp để xây dựng ứng dụng web full-stack nhanh chóng và hiệu quả. JavaScript kết hợp Next.js 13 với App Router cho phép tối ưu hóa hiệu suất và SEO (Search Engine Optimization), TypeScript để đảm bảo tính an toàn khi lập trình, Tailwind CSS (Cascading Style Sheets) giúp thiết kế giao diện nhanh chóng, Prisma hỗ trợ truy vấn dữ liệu, tRPC xây dựng API an toàn và hiệu quả, Supabase đóng vai trò như một backend-as-a-service. Sự kết hợp này giúp đơn giản quy trình phát triển, giảm thiểu lỗi và nâng cao hiệu suất làm việc [8-11].

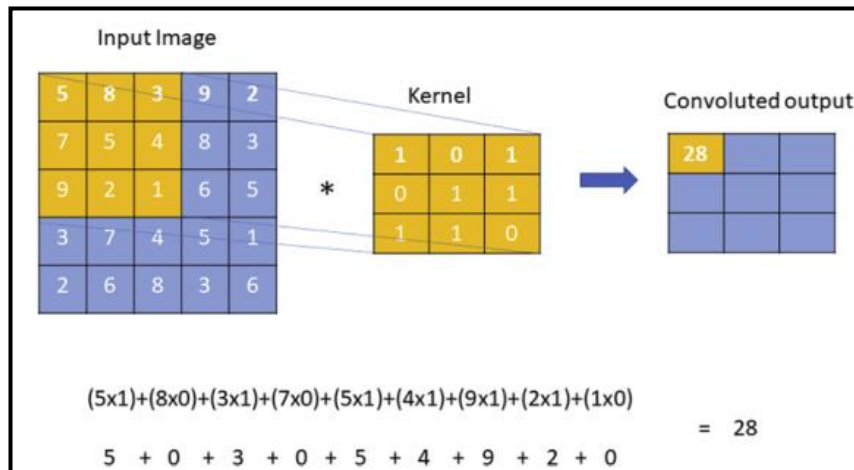
2.2. Mạng học sâu

Mạng CNN được chia thành hai phần chính: phần rút trích đặc trưng (feature extraction) mạng sẽ tiến hành tính toán hàng loạt phép tích chập và phép hợp nhất (pooling) để phát hiện các đặc trưng; phần phân lớp (classification) một layer với các liên kết đầy đủ sẽ đóng vai trò như một bộ phân lớp các đặc trưng đã rút trích được trước đó [12].



Hình 1. Kiến trúc mạng CNN [12]

Tầng Convolutional: là tầng quan trọng nhất và cũng là layer đầu tiên của mạng CNN. Các tham số của tầng Convolutional bao gồm một tập hợp bộ lọc có thể học được (kernel). Các bộ lọc đều nhỏ về mặt không gian nhưng mở rộng toàn bộ độ sâu của khối lượng đầu vào.



Hình 2. Tính toán tích chập của Convolution layer mạng CNN [12]

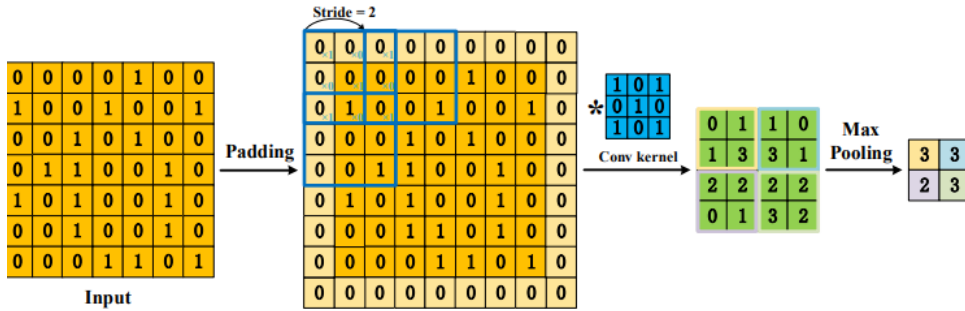
Kết quả đầu ra trong tầng Convolutional được gọi là feature map. Feature map phụ thuộc vào ba tham số:

Depth: đại diện cho số lượng bộ lọc được sử dụng trong tính toán tích chập.

Stride: số bước trượt bộ lọc trên đầu vào.

Padding: cho phép kiểm soát kích thước đầu ra. Việc áp dụng tích chập cho một đầu vào nếu giảm kích thước đầu ra dễ dẫn đến mất thông tin.

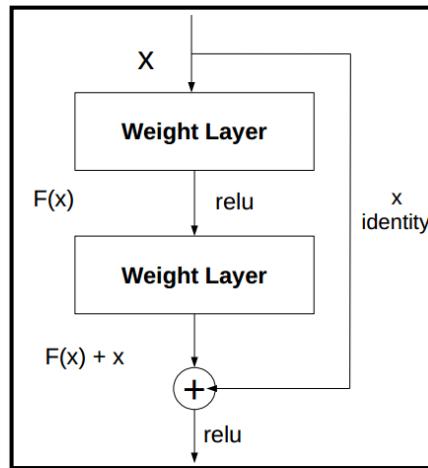
Tầng Pooling: thực hiện việc giảm bớt số lượng tham số khi hình ảnh quá lớn. công việc này giảm kích thước của mỗi features map ở *Convolutional Layer* và cũng để chọn thông tin quan trọng nhất của *features*. Các *pooling* có nhiều loại khác nhau: *max pooling* (lấy phần tử lớn nhất), *average pooling* (lấy tổng trung bình), *sum pooling* (lấy tổng tất cả các phần tử).



Hình 3. Max Pooling - Pooling layer [12]

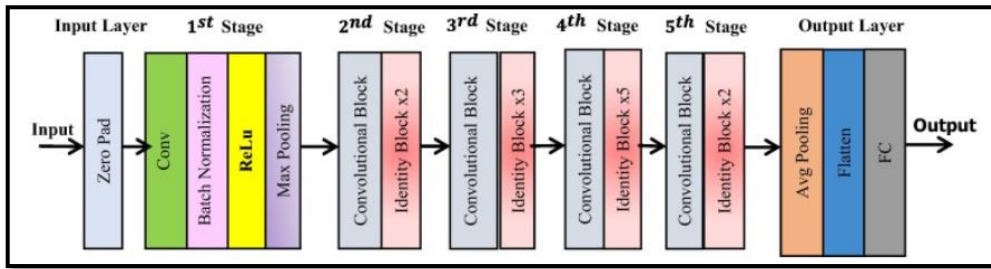
Tầng Fully Connected: là tầng cuối trong mạng CNN. Tầng này có nhiệm vụ đưa ra kết quả đầu ra (output) sau khi đã nhận được ảnh truyền từ các layer trước đó (convolutional layer, pooling layer).

ResNet (Residual Neural Network): ResNet thiết kế khắc phục được vấn đề vanishing gradient hay exploding gradient thường gặp trong các mạng nơ-ron sâu [13]. Trong kiến trúc mạng ResNet, Residual Block là một khối được thiết kế để giải quyết vấn đề vanishing gradient trong các mạng nơ-ron sâu, cho phép thông tin được truyền trực tiếp từ đầu vào đến đầu ra của khối mà không cần qua các lớp trung gian. Điều này được thực hiện thông qua một kết nối tắt (skip connection)



Hình 4. Residual Block [13]

Trong mô hình ResNet [13], các tác giả đã đề xuất nhiều thiết lập ResNet với 18, 34, 50, 101 và 152 lớp, trong đó mỗi lớp của ResNet bao gồm một số khối. Trong các cấu hình này, các khối Identity và Convolutional được kết hợp để tạo ra một mô hình ResNet. Trong bài báo này, nhóm tác giả sử dụng ResNet 50 để phân đoạn vì ResNet 50 có các lớp sâu hơn ResNet 34 và ít tham số hơn ResNet 101, 152 dẫn đến thời gian huấn luyện nhanh hơn [13].

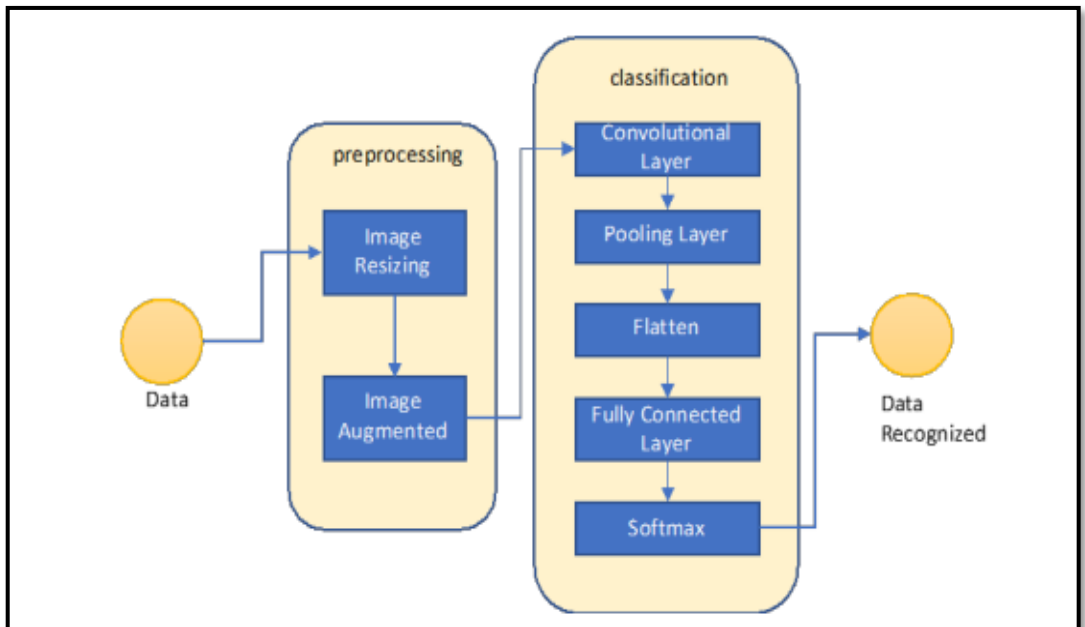


Hình 5. Kiến trúc mô hình ResNet 50 [14]

ResNet50 có tổng cộng 50 lớp, bao gồm các tầng convolutional, tầng pooling và tầng fully connected. Mạng được chia thành 5 stages, mỗi stages gồm một hoặc nhiều khối residual. Khối residual là thành phần cốt lõi của ResNet, sử dụng các shortcut connections để truyền trực tiếp thông tin từ các lớp trước đến các lớp sau [14].

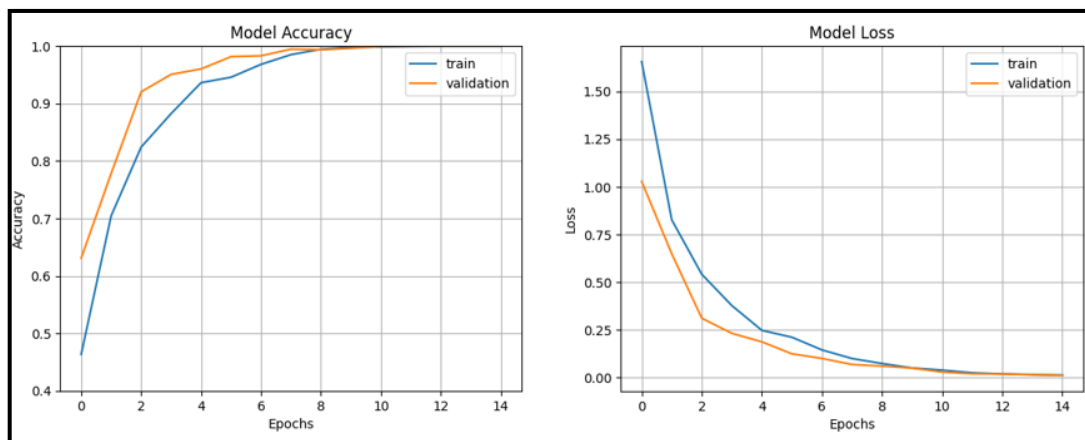
2.3. Ứng dụng ResNet trong nhận diện khuôn mặt

Trong mô hình này, ResNet50 được sử dụng như một bộ trích xuất đặc trưng tiền huấn luyện với trọng số từ tập ImageNet. Các lớp fully connected phía sau được thiết kế lại để phù hợp với số lớp tương ứng với các đối tượng khuôn mặt cụ thể trong tập dữ liệu huấn luyện. Quá trình huấn luyện được thực hiện trên tập ảnh được tiền xử lý và phân chia thành hai phần: huấn luyện và kiểm thử giúp mô hình học được đặc trưng khuôn mặt từng lớp và đánh giá được hiệu suất tổng thể.



Hình 6. Mô hình phân loại hình ảnh [14]

Trọng số của các lớp trong ResNet50 được đóng băng để giữ lại các đặc trưng học được từ ImageNet, trong khi các lớp phía sau được huấn luyện với tốc độ học tối ưu để nhận diện chính xác từng khuôn mặt.



Hình 7. Kết quả huấn luyện

Kết quả huấn luyện trong Hình 7 biểu diễn độ chính xác và hàm mất mát. Kết quả cho thấy mô hình đạt hiệu năng cao, với độ chính xác vượt trội sau nhiều epoch. Mô hình được sử dụng để dự đoán ảnh thực tế và lưu trữ dưới dạng file .h5, sẵn sàng cho việc tích hợp vào các hệ thống xác thực khuôn mặt.

Mô hình nổi bật trong nhận diện khuôn mặt nhờ khả năng trích xuất đặc trưng sâu với cấu trúc mạng có các kết nối tắt, giúp giảm thiểu vấn đề mất mát gradient so với AlexNet hay VGGNet, vốn có kiến trúc đơn giản hơn nhưng dễ gặp hiện tượng quá khớp khi xử lý tập dữ liệu lớn. So với GoogLeNet, hiệu suất vượt trội hơn trong việc học các đặc trưng phức tạp nhờ độ sâu mạng lớn mà vẫn duy trì được hiệu quả tính toán, trong khi GoogLeNet tập trung vào giảm tham số qua các Inception module. MobileNet và EfficientNet dù tối ưu cho thiết bị di động với chi phí tính toán thấp, thường kém hơn về độ chính xác trong các tác vụ nhận diện khuôn mặt phức tạp, đặc biệt khi cần phân biệt nhiều đối tượng với độ tương đồng cao. Trong ứng dụng cụ thể, mô hình ResNet được huấn luyện với trọng số ImageNet và tinh chỉnh các lớp fully connected, đạt độ chính xác cao và ổn định, phù hợp cho các hệ thống xác thực khuôn mặt yêu cầu hiệu năng cao.

3. PHƯƠNG PHÁP XÂY DỰNG MẠNG XÃ HỘI THU NHỎ

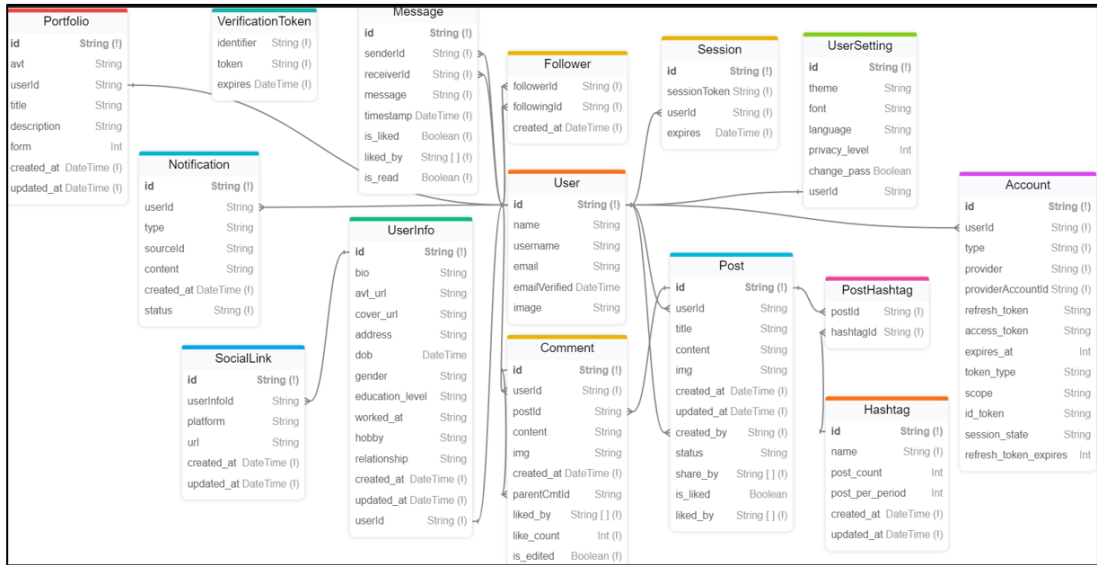
Ứng dụng mạng xã hội thu nhỏ được xây dựng trên nền tảng web application và mobile app. Phía web application, ứng dụng được phát triển bằng T3 Stack với các công nghệ chính như Next.js, TypeScript, Tailwind CSS, Prisma và tRPC đảm bảo hiệu năng cao, dễ mở rộng và an toàn. Ứng dụng mobile được xây dựng bằng React Native giúp hệ thống tích hợp tính năng portfolio cho phép người dùng tạo và lưu trữ hồ sơ cá nhân một cách trực quan. Các kỹ thuật bảo mật được áp dụng toàn diện trong quá trình xác thực, truyền dữ liệu và lưu trữ thông tin. Mạng học sâu được sử dụng để nhận diện khuôn mặt giúp xác thực người dùng.

3.1. Thiết kế cơ sở dữ liệu

Ứng dụng mạng xã hội thu nhỏ được xây dựng trên hệ quản trị cơ sở dữ liệu PostgreSQL, được quản lý thông qua Prisma ORM. Hệ thống được thiết kế với kiến trúc quan hệ – đối tượng, giúp tối ưu hóa tương tác mạng xã hội, quản lý xác thực người dùng, cá nhân hóa hồ sơ, tích hợp tính năng nhắn tin, bình luận, thông báo.

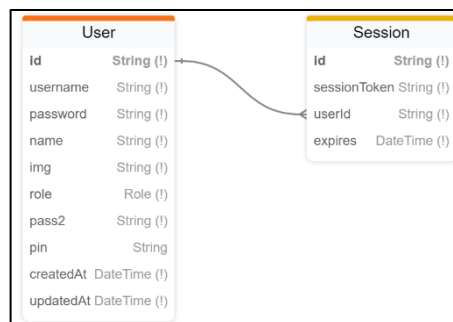
Hệ thống cơ sở dữ liệu trong dự án được thiết kế và quản lý bằng Prisma, cho phép tổ chức chặt chẽ các thông tin cốt lõi của mạng xã hội. Dữ liệu người dùng được lưu trữ chi tiết

qua các bảng như User, UserInfo, UserSetting, trong đó quản lý thông tin cá nhân, cài đặt riêng tư, ảnh đại diện, thông tin tiểu sử và các liên kết mạng xã hội. Các hoạt động tương tác như bài đăng, bình luận, thả tim, hashtag và thông báo được xử lý thông qua các bảng Post, Comment, Hashtag, Notification, cùng các mối liên kết chặt chẽ để đảm bảo hiệu quả và tính toàn vẹn dữ liệu. Ngoài ra, cơ sở dữ liệu cũng hỗ trợ các tính năng nâng cao như lưu trữ portfolio cá nhân qua bảng Portfolio, hệ thống theo dõi bạn bè với bảng Follower, nhắn tin trực tiếp qua bảng Message, xác thực bảo mật thông qua tích hợp với NextAuth qua các bảng như Account, Session, VerificationToken.

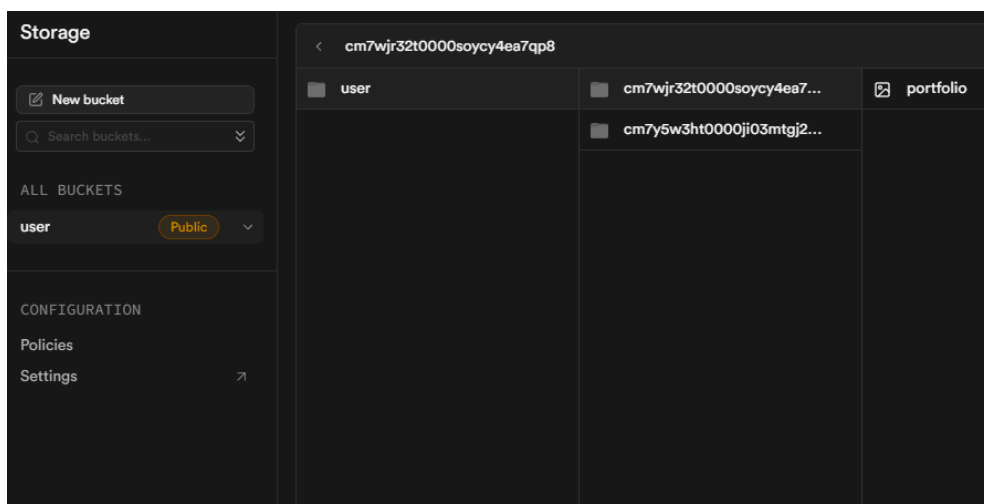


Hình 8. Cơ sở dữ liệu mạng xã hội thu nhỏ - phân hệ User

Cơ sở dữ liệu của ứng dụng admin được thiết kế riêng biệt và tối ưu cho việc quản lý người dùng có quyền truy cập hệ thống ở các cấp độ khác nhau. Dựa trên Prisma, mô hình dữ liệu bao gồm hai bảng chính là User và Session, cho phép quản lý tài khoản đăng nhập, phiên làm việc và phân quyền một cách linh hoạt. Bảng User chứa các thông tin định danh như tên người dùng, mật khẩu, ảnh đại diện, cùng các thông tin bảo mật nâng cao như mật khẩu cấp hai (pass2) và mã PIN. Vai trò người dùng được phân loại rõ ràng qua enum Role, bao gồm ba cấp: SUPERADMIN, ADMIN và EMPLOYEE, giúp kiểm soát quyền truy cập theo nhiệm vụ. Trong khi đó, bảng Session theo dõi thông tin đăng nhập và thời gian hiệu lực của phiên làm việc, đảm bảo quá trình xác thực và bảo mật được kiểm soát chặt chẽ. Việc tách riêng cơ sở dữ liệu cho admin giúp tăng cường tính bảo mật và dễ dàng mở rộng hệ thống trong môi trường đa ứng dụng.



Hình 9. Cơ sở dữ liệu mạng xã hội thu nhỏ - phân hệ Admin



Hình 10. Cơ sở dữ liệu và storage đã triển khai trên Supbase

3.2. Kỹ thuật bảo mật

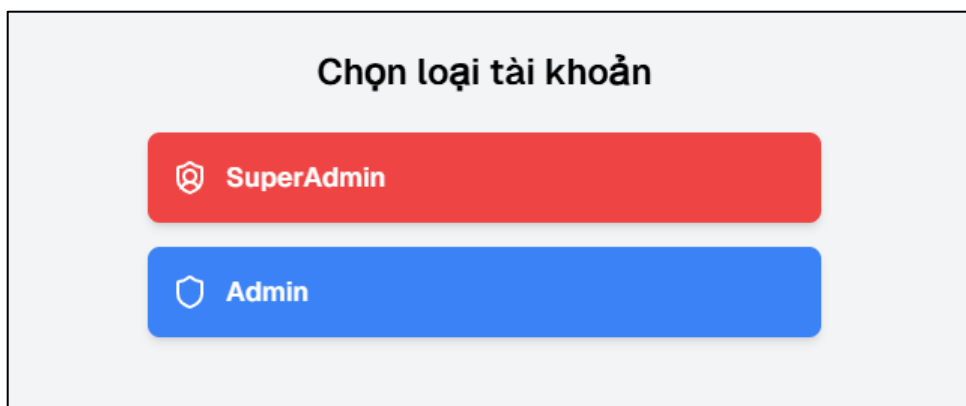
Người dùng đăng ký, đăng nhập trên ứng dụng bằng các tài khoản từ nhà cung cấp thứ 3 như Google, Github, Discord thông qua NextAuth từ đó tận dụng khả năng bảo mật có sẵn của các nền tảng này. Đồng thời, ứng dụng có thêm cơ chế Session để giữ thông tin phiên khi đăng nhập, người dùng không cần đăng nhập nhiều lần khi truy cập trang ứng dụng và có thể bảo mật khi đăng nhập trên nhiều thiết bị.

Bảo vệ biến môi trường chứa các thông tin đường dẫn database, khóa bảo mật của các dịch vụ,... bằng cách ẩn file chứa các biến này khi triển khai dự án lên đường dẫn công khai.

Mỗi người dùng được thiết lập một đường dẫn trang cá nhân và kho lưu trữ ảnh riêng trong database, chế độ công khai hoặc riêng tư tùy thuộc theo cài đặt cá nhân của người dùng. Áp dụng mật khẩu cấp 2 khi người dùng có các thao tác chỉnh sửa cài đặt cá nhân liên quan tới bảo mật thông tin, người dùng có thể bật/tắt nếu muốn.

Mã hóa các trường thông tin quan trọng của người dùng trong database.

Sử dụng mạng học sâu nhận diện khuôn mặt để xác thực người dùng khi đăng nhập vào hệ thống.



Hình 11. Giao diện chọn vai trò đăng nhập cho quản trị viên

Sau khi lựa chọn vai trò, quản trị viên sẽ được chuyển hướng tới trang đăng nhập. Quản trị viên đăng nhập bằng 2 lựa chọn gồm: tài khoản, mật khẩu kèm mật khẩu cấp 2 hoặc đăng nhập bằng khuôn mặt.

Đăng nhập bằng tài khoản, mật khẩu, mật khẩu cấp 2: người dùng nhập đầy đủ thông tin, chọn đăng nhập để chuyển hướng tới trang quản trị theo vai trò

Hình 12. Giao diện đăng nhập cho quản trị viên

Đăng nhập bằng xác thực khuôn mặt: đăng nhập bằng xác thực khuôn mặt được thiết kế để đảm bảo an toàn và hiệu quả cho quản trị viên. Khi bắt đầu, hệ thống chuyển hướng quản trị viên tới trang nhận diện khuôn mặt. Camera tự động chụp 10 khung hình và gửi dữ liệu về server. Server xử lý, định dạng hình ảnh và chuyển chúng dưới dạng input tới Endpoint của mô hình AI. Sau quá trình xử lý, server nhận phản hồi từ Endpoint để chuyển hướng quản trị viên tới trang quản trị tương ứng với vai trò của họ hoặc hiển thị thông báo lỗi nếu xác thực khuôn mặt không thành công.

Đánh giá hiệu quả bảo mật:

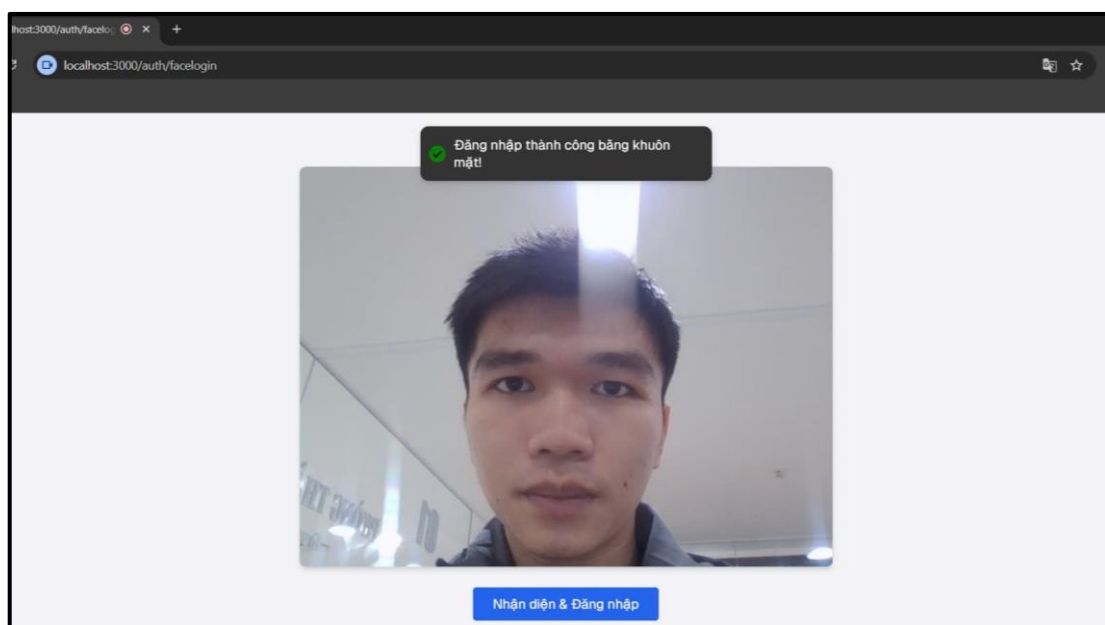
Độ chính xác: Mô hình AI đạt độ chính xác từ 70% đến 80%, tùy thuộc vào điều kiện ánh sáng và góc chụp.

Thời gian xử lý: Toàn bộ quy trình, từ chụp ảnh đến nhận phản hồi từ mô hình AI, hoàn tất trong vòng dưới 10 giây.

Độ trễ hệ thống: Độ trễ trung bình từ khi gửi yêu cầu đến khi nhận phản hồi từ Endpoint là khoảng 2-3 giây, đảm bảo trải nghiệm mượt mà.

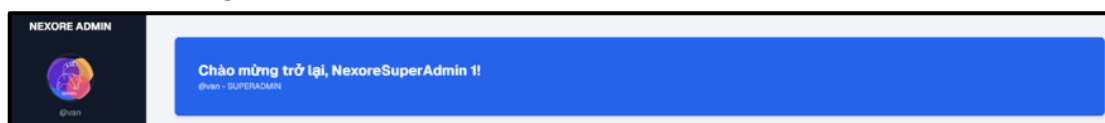
Tỷ lệ lỗi: Tỷ lệ xác thực sai (false negative) dưới 10% trong điều kiện tối ưu, góp phần tăng cường độ tin cậy của hệ thống.

Các chỉ số trên cho thấy hệ thống không chỉ đảm bảo tính bảo mật mà còn tối ưu về hiệu suất, đáp ứng yêu cầu quản trị nhanh chóng và an toàn.



Hình 13. Giao diện xác thực khuôn mặt khi đăng nhập ứng dụng

Hệ thống xác thực khuôn mặt đã được tối ưu hóa để không yêu cầu chụp ảnh thủ công. Khi người dùng truy cập, hệ thống tự động nhận diện khuôn mặt thông qua camera tích hợp. Nếu xác thực thành công, hệ thống sẽ tự động trả về thông tin tài khoản tương ứng với khuôn mặt đã được ghi nhận mà không cần thêm thao tác chụp ảnh. Giao diện hiện tại hiển thị trạng thái xác thực thành công.



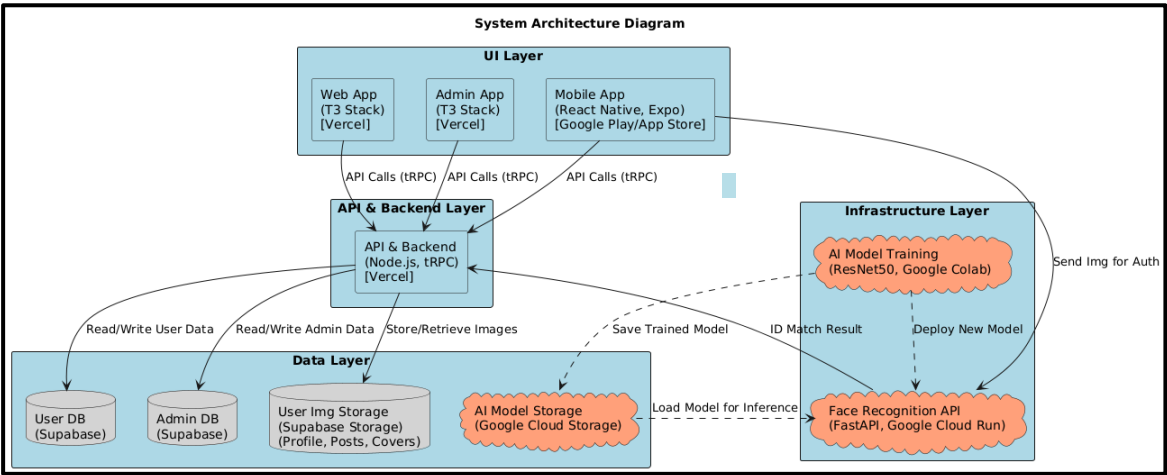
Hình 14. Giao diện sau khi xác thực khuôn mặt thành công

4. THỰC NGHIỆM

4.1. Kiến trúc ứng dụng

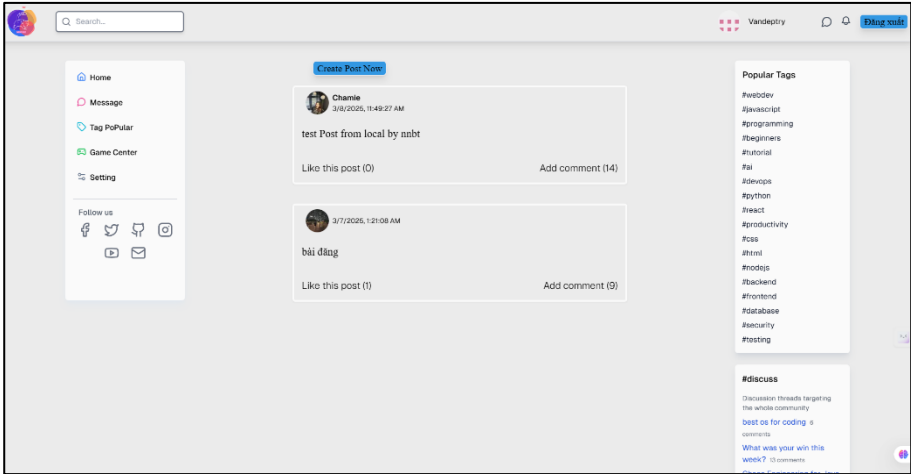
Hệ thống được thiết kế để phục vụ ba nhóm người dùng: Ứng dụng web dành cho người dùng truy cập thông qua trình duyệt; ứng dụng di động dành cho người dùng truy cập qua điện thoại; Ứng dụng web dành cho quản trị viên để quản lý người dùng.

Bên cạnh đó, hệ thống sử dụng dịch vụ nhận diện khuôn mặt dựa trên AI Model được triển khai trên Google Cloud để xác thực danh tính người dùng. Hệ thống được chia thành bốn lớp chính, bao gồm: Lớp giao diện người dùng (UI Layer): Chịu trách nhiệm hiển thị thông tin và nhận thao tác từ người dùng; Lớp dịch vụ API & Backend (API & Backend Layer): Xử lý logic nghiệp vụ và giao tiếp với cơ sở dữ liệu; Lớp lưu trữ dữ liệu (Data Layer): Lưu trữ thông tin người dùng, nội dung và mô hình AI; Lớp hạ tầng (Infrastructure Layer): Bao gồm các dịch vụ triển khai AI, lưu trữ mô hình và huấn luyện AI.



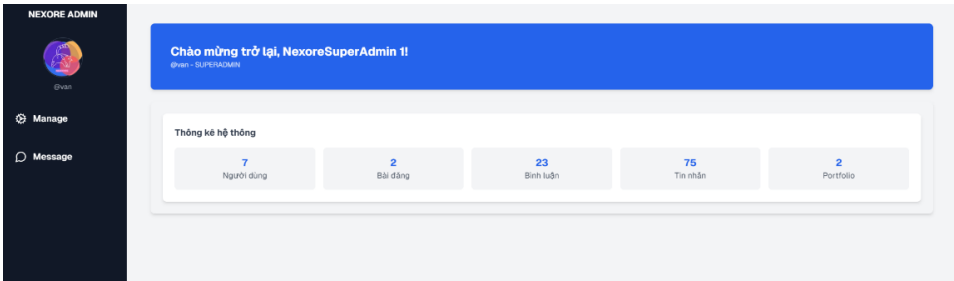
Hình 15. Kiến trúc hệ thống ứng dụng mạng xã hội thu nhỏ

4.2. Kết quả cài đặt

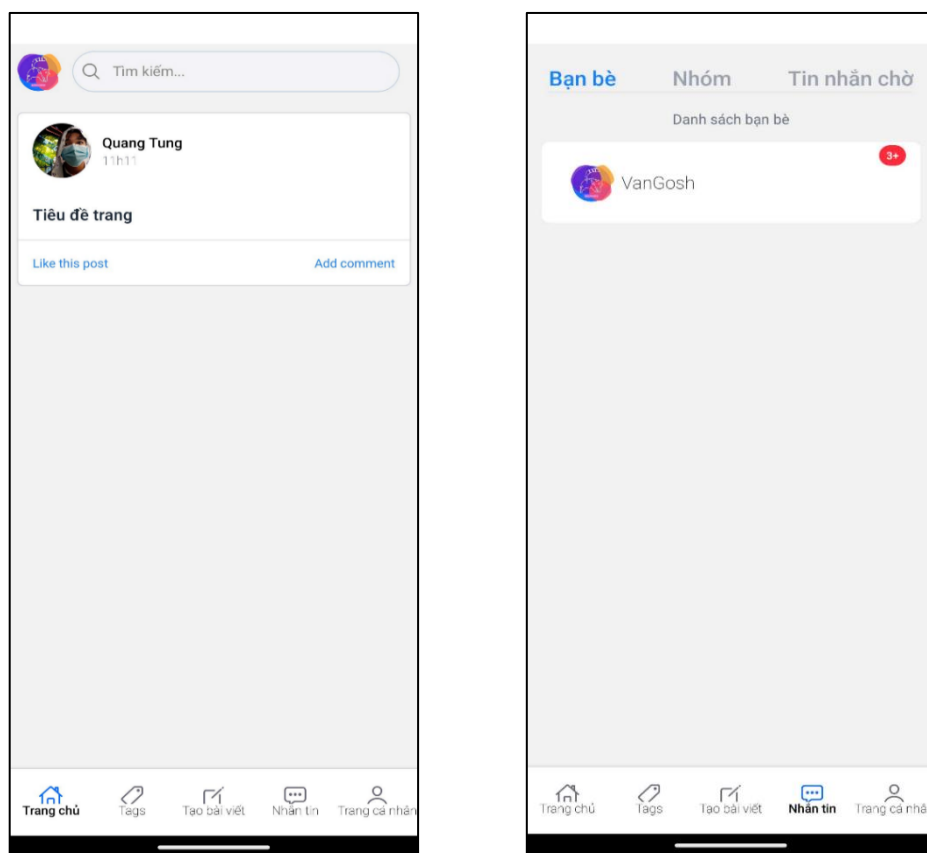


Hình 16. Giao diện mạng xã hội thu nhỏ trên nền tảng web cho user

Hệ thống đã được triển khai thử nghiệm với quy mô nhỏ, từ 10 đến 20 người dùng tham gia test thực tế. Trong giai đoạn này, tổng số bài đăng ghi nhận được dao động từ 5 đến dưới 10 bài, tất cả đều được xử lý theo giờ gần thực, phản ánh mức độ tương tác ban đầu của người dùng trên hệ thống.



Hình 17. Giao diện ứng dụng quản lý trên nền tảng web



Hình 18. Giao diện mạng xã hội thu nhỏ trên nền tảng Mobile

5. KẾT LUẬN

Với ý tưởng xây dựng một mạng xã hội thu nhỏ tích hợp đa nền tảng, nhóm tác giả đã phát triển thành công một hệ thống cho phép người dùng tương tác thông qua các chức năng như đăng bài, bình luận, nhắn tin, tạo portfolio cá nhân và quản lý tài khoản với các chế độ riêng tư hoặc công khai. Hệ thống được phát triển dựa trên T3 Stack và React Native, mang lại hiệu suất cao, khả năng mở rộng và đồng bộ giữa website và mobile app. Bên cạnh đó, hệ thống tích hợp các tính năng bảo mật như xác thực bằng khuôn mặt sử dụng mạng ResNet50, xác thực OTP và mã PIN, bảo vệ truy cập từ nhiều thiết bị, cùng với cơ chế mã hóa dữ liệu và mật khẩu cấp hai giúp người dùng trao đổi thông tin trên không gian mạng được an toàn. Hướng phát triển tiếp theo trong tương lai, nhóm tác giả nghiên cứu các phương pháp nhận diện khuôn mặt bảo mật như công nghệ quét mống mắt, đồng thời tích hợp thêm AI phát hiện và ngăn chặn bình luận tiêu cực và tối ưu hóa đề xuất bài viết.

TÀI LIỆU THAM KHẢO

1. Panta, B. (2024). Full stack web app development using T3 stack.
2. Van Truong, P., & Thao, T. T. (2020). Nhận dạng khuôn mặt trực tiếp và ứng dụng trong mở cửa tự động. *Journal of Measurement, Control, and Automation*, 1(1).

3. Karnila, S., Irianto, S., & Kurniawan, R. (2019). Face recognition using content based image retrieval for intelligent security. *International Journal of Advanced Engineering Research and Science*, 6(1), 91-98.
4. Nguyễn, T. H., Trịnh, T. T. L., Trần, B. T., Phan, K. Y. N., Trần, T. Đ., & Nguyễn, T. N. (2020). Giải pháp điểm danh sinh viên bằng nhận diện gương mặt với đặc trưng Haar-Like kết hợp thuật toán rừng ngẫu nhiên.
5. Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173, 24-49.
6. Wieczorek, M., Siłka, J., Woźniak, M., Garg, S., & Hassan, M. M. (2021). Lightweight convolutional neural network model for human face detection in risk situations. *IEEE Transactions on Industrial Informatics*, 18(7), 4820-4829.
7. Vu, H. N., Nguyen, M. H., & Pham, C. (2022). Masked face recognition with convolutional neural networks and local binary patterns. *Applied Intelligence*, 52(5), 5497-5512.
8. Ikram, H., & Ardiansyah, M. (2024). Development of Online NAS Application User Guide Website using NextJS 13 and REST API to Enhance User Understanding. *Jurnal Inotera*, 9(1), 53-58.
9. Jones, R. (Ed.). (2014). ECOOP 2014--Object-Oriented Programming: 28th European Conference, Uppsala, Sweden, July 28--August 1, 2014, Proceedings (Vol. 8586). Springer.
10. Goldberg, J. (2022). Learning TypeScript. " O'Reilly Media, Inc."
11. Rifandi, F., Adriansyah, T. V., & Kurniawati, R. (2022). Website Gallery Development Using Tailwind CSS Framework. *Jurnal E-Komtek (Elektro-Komputer-Teknik)*, 6(2), 205-214.
12. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12), 6999-7019.
13. Thoma, M. (2017). Analysis and optimization of convolutional neural network architectures. *arXiv preprint arXiv:1707.09725*.
14. Shehab, L. H., Fahmy, O. M., Gasser, S. M., & El-Mahallawy, M. S. (2021). An efficient brain tumor image segmentation based on deep residual networks (ResNets). *Journal of King Saud University-Engineering Sciences*, 33(6), 404-412.

ABSTRACT

SERURITY METHOD FOR MINI SOCIAL NETWORK

Pham Dinh Van¹, Nhieuhong Phuc¹, Nguyen Nhu Quang Tung¹

Tran Thi Van Anh¹, Dinh Thi Man^{1*}

¹*Ho Chi Minh City University of Industry and Trade*

*Email: *mandt@huit.edu.vn*

The use of social networks for users today still has many risks of insecurity because of not paying attention to security policies, accounts are vulnerable to attacks, and user profiles have not been optimized. In this article, the authors present a method of building a miniature social network application on multiple platforms (Web, Mobile) in addition to the common features of a social network such as posting articles, commenting, messaging, etc. The application is integrated with security features including setting up user authentication modes when logging in: using deep learning networks to identify and authenticate, logging in via OTP code, and PIN code; warning when logging in from two or more devices; managing user accounts with private or public options to help users exchange information on social networks safely. In addition, user profile management is optimized through the function of creating

personal Portfolios with different interface options. The research results contribute to creating a safe social network communication environment, secure information, and effective profile management.

Keywords: Mini social network, ResNet, T3 Stack, Portfolio

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Phạm Đình Văn

Số điện thoại liên hệ: 0376699063

HỆ THỐNG QUẢN LÝ DINH DƯỠNG THÔNG MINH

Võ Ngọc Nguyên Chương¹, Trần Trí Cường¹, Phạm Vũ Lương Bằng¹,
Trần Vinh Hiền¹, Đỗ Quang Vinh¹, Trần Văn Thọ^{1*}

¹Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: thotv@huit.edu.vn

Ngày nhận bài: 20/04/2025; Ngày chấp nhận đăng: 12/05/2025

TÓM TẮT

Bối cảnh các bệnh liên quan đến dinh dưỡng ngày càng gia tăng đòi hỏi các giải pháp quản lý chế độ ăn uống hiệu quả và cá nhân hóa. Nghiên cứu này trình bày việc phát triển và đánh giá hệ thống quản lý dinh dưỡng thông minh SmartMeal, một ứng dụng di động đa nền tảng nhằm hỗ trợ người dùng xây dựng và duy trì lối sống lành mạnh. Hệ thống tích hợp các công nghệ Trí tuệ Nhân tạo tiên tiến: Học tăng cường (RL) với thuật toán Proximal Policy Optimization (PPO) kết hợp Neural Collaborative Filtering (NCF) để đưa ra gợi ý thực đơn đa bước, cá nhân hóa và thích nghi động; các mô hình học sâu (ResNet, YOLO, Segment Anything Model - SAM) cho phép nhận diện món ăn, ước lượng calo từ hình ảnh với độ chính xác cao; cùng một Chatbot tư vấn dinh dưỡng thông minh dựa trên kỹ thuật Retrieval-Augmented Generation (RAG) và tích hợp Function Calling để cung cấp các thông tin đáng tin cậy có ngữ cảnh giúp truy xuất dữ liệu cá nhân và thông tin dinh dưỡng chi tiết. Hệ thống được triển khai trên nền tảng Flutter, mang lại trải nghiệm người dùng trực quan và liền mạch. Kết quả thực nghiệm cho thấy hệ thống đạt hiệu quả cao trong việc gợi ý ($NDCG@3 \approx 0.9-1.0$), nhận diện thực phẩm (độ chính xác ResNet 85%), ước lượng calo (sai số trung bình $\approx 20\%$) và trả lời câu hỏi dinh dưỡng (độ chính xác RAG 85%). Mặc dù còn một số hạn chế về tối ưu phần thưởng trong mô hình học tăng cường và sự phụ thuộc vào chất lượng dữ liệu, SmartMeal chứng minh tiềm năng lớn trong việc cải thiện sức khỏe cộng đồng thông qua quản lý dinh dưỡng cá nhân hóa và thông minh.

Từ khóa: Quản lý dinh dưỡng, Học tăng cường, PPO, Nhận diện hình ảnh, Ước lượng calo, Chatbot RAG, Cá nhân hóa.

1. MỞ ĐẦU

Trong kỷ nguyên số, sức khỏe và dinh dưỡng trở thành mối quan tâm hàng đầu của xã hội. Tỷ lệ mắc các bệnh mãn tính không lây liên quan đến chế độ ăn uống như béo phì, tiểu đường, tim mạch đang gia tăng đáng báo động trên toàn cầu và tại Việt Nam [1]. Việc xây dựng một chế độ ăn uống khoa học, phù hợp với thể trạng, mục tiêu sức khỏe và sở thích cá nhân là cực kỳ quan trọng nhưng cũng đầy thách thức đối với nhiều người do thiếu kiến thức, thời gian hoặc công cụ hỗ trợ hiệu quả.

Các hệ thống gợi ý thực phẩm hiện có thường dựa trên các phương pháp lọc cộng tác (CF) hoặc dựa trên nội dung truyền thống [2]. Tuy nhiên, những phương pháp này thường mang tính tĩnh, khó thích ứng với sự thay đổi động về nhu cầu dinh dưỡng, mục tiêu sức khỏe (giảm cân, tăng cơ), tình trạng dị ứng hay sở thích của người dùng theo thời gian. So với các ứng dụng hiện có như MyFitnessPal, Yazio hay FatSecret – vốn chủ yếu dựa trên nhập liệu

thủ công và các công cụ phân tích tĩnh – SmartMeal nổi bật nhờ khả năng cá nhân hóa động, tích hợp đa mô hình AI và tự động phân tích bữa ăn từ hình ảnh. Ví dụ, MyFitnessPal cung cấp cơ sở dữ liệu thực phẩm phong phú, nhưng thiếu khả năng gợi ý thực đơn dựa trên mục tiêu sức khỏe hoặc lịch sử tương tác. Yazio tập trung vào theo dõi lượng calo nhưng không hỗ trợ phân tích hình ảnh hoặc đối thoại theo ngữ cảnh như Chatbot SmartMeal. Sự kết hợp các công nghệ học tăng cường và Chatbot RAG của SmartMeal giúp vượt xa những giới hạn truyền thống của các ứng dụng này. SmartMeal mang lại những ưu điểm nổi bật hơn như các ứng dụng truyền thống ở trên chủ yếu tập trung vào ghi nhật ký dinh dưỡng thủ công (dễ sai sót và tốn thời gian), cung cấp cơ sở dữ liệu thực phẩm lớn nhưng khả năng gợi ý thực đơn cá nhân hóa sâu sắc, thích ứng động theo thời gian thực và khả năng phân tích bữa ăn từ hình ảnh còn hạn chế hoặc không có. Chúng cũng thường thiếu khả năng đánh giá chất lượng dinh dưỡng của thực phẩm và đặc biệt là thiếu khả năng cung cấp tư vấn dinh dưỡng đáng tin cậy, có ngữ cảnh mà thường chỉ dựa trên các câu trả lời cố định hoặc không có khả năng đối thoại tự nhiên.

Để giải quyết những hạn chế này, nghiên cứu này đề xuất xây dựng hệ thống quản lý dinh dưỡng thông minh mang tên SmartMeal. Mục tiêu của hệ thống là cung cấp một giải pháp toàn diện, tích hợp: (1) Hệ thống gợi ý thực đơn cá nhân hóa sử dụng học tăng cường (PPO) và lọc cộng tác no-ron (NCF); (2) Khả năng nhận diện thực phẩm, ước lượng calo chính xác từ hình ảnh bằng các mô hình Học sâu (ResNet, YOLO, SAM); (3) Chatbot tư vấn dinh dưỡng thông minh dựa trên Retrieval-Augmented Generation (RAG) sử dụng mô hình Gemini và tích hợp Function Calling để trả lời các câu hỏi một cách đáng tin cậy, có ngữ cảnh, truy cập dữ liệu lịch sử ăn uống và thông tin sức khỏe cá nhân, cũng như cung cấp thông tin dinh dưỡng chi tiết thông qua truy vấn cơ sở dữ liệu; (4) Triển khai các mô hình trên ứng dụng di động Flutter. Nghiên cứu này đóng góp vào việc ứng dụng các kỹ thuật AI tiên tiến để tạo ra một công cụ hỗ trợ dinh dưỡng hiệu quả, góp phần nâng cao sức khỏe cá nhân và cộng đồng.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Hệ thống SmartMeal được thiết kế với kiến trúc client-server, trong đó ứng dụng di động Flutter đóng vai trò giao diện người dùng (client) và các mô hình AI được xử lý tại backend Flask.

2.1. Mô hình gợi ý thực phẩm cá nhân hóa (PPO + NCF)

Quá trình gợi ý được mô hình hóa như một Quy trình Quyết định Markov (MDP) và tối ưu hóa bằng học tăng cường (RL) với thuật toán Proximal Policy Optimization (PPO) [3, 4], một phương pháp Actor-Critic hiệu quả và ổn định.

Trạng thái (State): Biểu diễn bằng các vector đa chiều kết hợp nhiều thông tin như thông tin tương tác người dùng - món ăn (lấy từ NCF), sở thích động (mục tiêu calo, protein, bữa ăn hiện tại), và lịch sử gợi ý (sử dụng cơ chế Cross-Attention [5] để nắm bắt sự đa dạng).

Hành động (Action): Chọn một món ăn từ danh sách ứng viên (đã được lọc sơ bộ để giảm không gian hành động).

Phần thưởng (Reward): Hàm phần thưởng được thiết kế để cân bằng nhiều yếu tố:

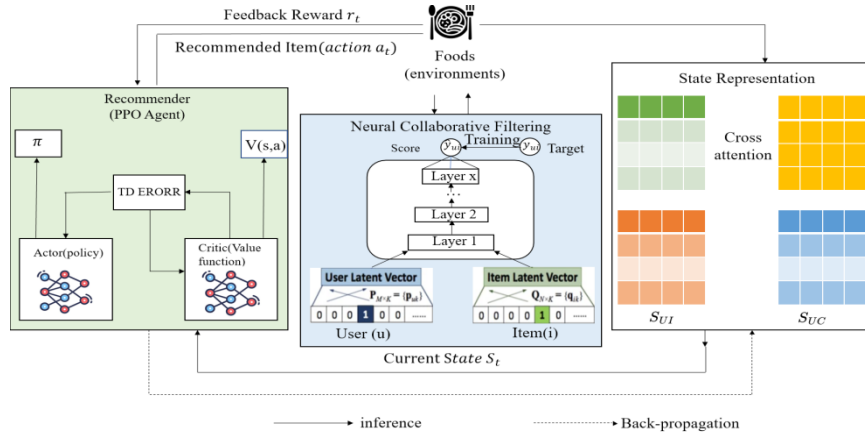
- + PreferenceScore: Mức độ phù hợp với sở thích người dùng, dự đoán bởi mô hình Neural Collaborative Filtering (NCF) [6] được huấn luyện trên dữ liệu lịch sử đánh giá. NCF sử dụng kiến trúc kết hợp GMF và MLP để học các tương tác phi tuyến.

- + NutritionScore: Mức độ đáp ứng mục tiêu dinh dưỡng (calo, macro-nutrients).

- + DiversityScore: Khuyến khích sự đa dạng, phạt nếu gợi ý món ăn lặp lại gần đây.

- + UserFeedback (RLHF): Trọng số phần thưởng được điều chỉnh dựa trên đánh giá trực tiếp (1-5 sao) của người dùng đối với gợi ý [7].

Chính sách (Policy) và Hàm giá trị (Value Function): Được xấp xỉ bởi các mạng nơ-ron sâu (Deep RL) và cập nhật bằng thuật toán PPO với cơ chế clipping để đảm bảo huấn luyện ổn định.



Hình 1. Kiến trúc tổng quan của hệ thống gợi ý thực phẩm

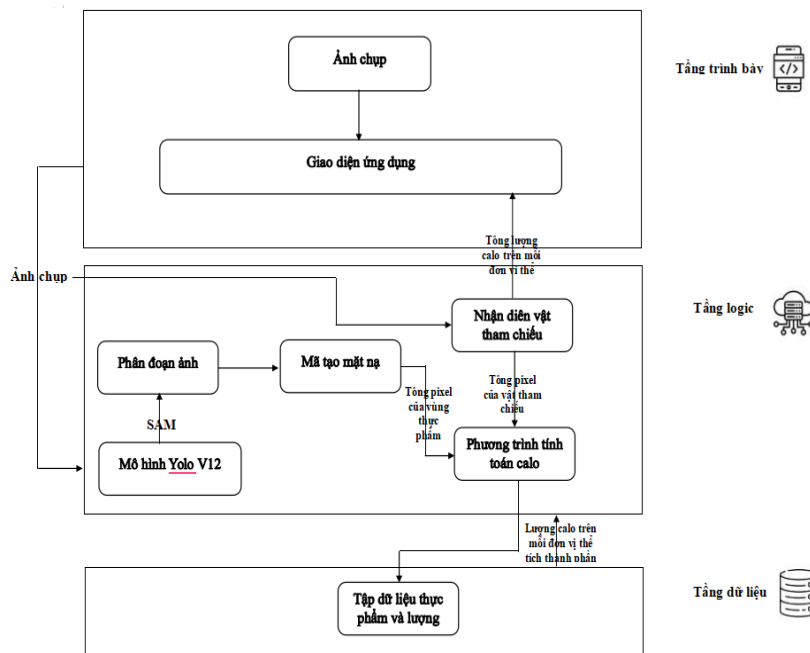
2.2. Mô hình phân tích hình ảnh thực phẩm (ResNet, YOLO, SAM)

Để hỗ trợ theo dõi dinh dưỡng tự động và đảm bảo chất lượng thực phẩm:

Nhận diện món ăn: Sử dụng mô hình ResNet-50 [8] được huấn luyện trước trên ImageNet và tinh chỉnh (fine-tuned) trên tập dữ liệu hình ảnh thực phẩm (bao gồm món ăn Việt Nam và các dữ liệu công khai [9, 10]) để phân loại món ăn hoặc các nguyên liệu thực phẩm dựa trên các đặc trưng hình ảnh như màu sắc, kết cấu của món ăn đó.

Phát hiện và Phân đoạn: Mô hình YOLOv12 [11, 12] được dùng để phát hiện vị trí các đối tượng thực phẩm trong ảnh. Sau đó, Segment Anything Model (SAM) [13] thực hiện phân đoạn chính xác từng đối tượng dựa trên bounding box từ YOLO.

Ước lượng Calo: Diện tích pixel của từng phần thực phẩm (từ SAM) được chuyển đổi sang diện tích thực tế dựa trên một vật thể tham chiếu có kích thước chuẩn (ví dụ: thẻ CMND) trong ảnh. Trọng lượng được ước tính dựa trên diện tích và dữ liệu tham chiếu trong cơ sở dữ liệu món ăn. Calo và các chất dinh dưỡng khác được tính toán dựa trên trọng lượng ước tính và thông tin dinh dưỡng/100g từ cơ sở dữ liệu.



Hình 2. Sơ đồ nhận dạng hình ảnh và ước lượng calo của món ăn

2.3. Chatbot tư vấn dinh dưỡng (RAG)

Để cung cấp thông tin dinh dưỡng đáng tin cậy và trả lời câu hỏi người dùng:

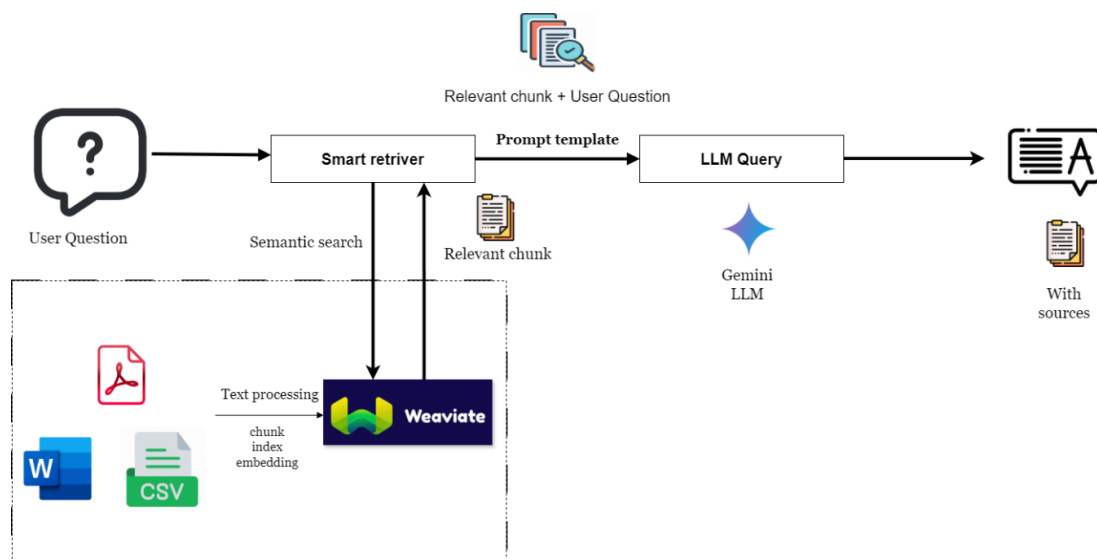
Kiến trúc RAG [14]: Kết hợp mô hình truy xuất (Retriever) và mô hình sinh văn bản (Generator) kết hợp với tính năng Function Calling [15]. Điều này cho phép mô hình ngôn ngữ không chỉ sinh văn bản dựa trên ngữ cảnh được truy xuất mà còn gọi các hàm (functions) bên ngoài để thực hiện các tác vụ cụ thể hoặc truy cập dữ liệu động.

Cơ sở tri thức (Knowledge Base): Bao gồm: (1) Dữ liệu hỏi - đáp (Q&A) dinh dưỡng; (2) Tài liệu chuyên sâu (file .docx giáo trình về dinh dưỡng [16]); (3) Cơ sở dữ liệu món ăn (thành phần, dinh dưỡng trích xuất từ file .pdf về thành phần dinh dưỡng của các món ăn [17]).

Xử lý dữ liệu: Văn bản từ cơ sở tri thức được chia nhỏ (chunking) thành các đoạn ngắn có nghĩa. Các đoạn văn bản này được vector hóa bằng mô hình embedding tiếng Việt chuyên biệt (sử dụng Sentence Transformer [Vietnamese Embedding]) để tạo ra biểu diễn vector. Các vector này được lưu trữ trong cơ sở dữ liệu vector (sử dụng Weaviate) với chỉ mục HNSW (Hierarchical Navigable Small Worlds) để truy vấn nhanh.

Truy xuất (Retrieval): Khi người dùng đặt câu hỏi, câu hỏi đó cũng được vector hóa. Hệ thống thực hiện tìm kiếm các đoạn văn bản có biểu diễn vector tương đồng nhất với câu hỏi trong cơ sở dữ liệu vector (sử dụng Hybrid Search kết hợp tìm kiếm ngữ nghĩa và từ khóa) để lấy ra các đoạn ngữ cảnh (context) liên quan nhất.

Sinh văn bản (Generation): Các đoạn ngữ cảnh sau khi truy xuất sẽ được kết hợp với câu hỏi gốc của người dùng và một prompt hướng dẫn vai trò (ví dụ: "Bạn là một chuyên gia dinh dưỡng. Hãy trả lời câu hỏi sau dựa trên thông tin được cung cấp...") được đưa vào một mô hình ngôn ngữ lớn (LLM). Trong nghiên cứu này, nhóm sử dụng mô hình Gemini Pro (Google) làm thành phần sinh văn bản của kiến trúc RAG. Việc lựa chọn Gemini được dựa trên các yếu tố: (1) khả năng hiểu và sinh ngôn ngữ tiếng Việt vượt trội so với GPT-3.5 hay Mistral, (2) tốc độ phản hồi thấp khi tích hợp API, và (3) khả năng tương thích tốt với chức năng Function Calling và truy vấn dữ liệu động trong hệ thống thực tế. LLM sẽ sinh ra câu trả lời bằng tiếng Việt, bám sát thông tin đã truy xuất, đảm bảo tính chính xác và ngữ cảnh.



Hình 3. Tổng quan hệ thống Chatbot dinh dưỡng với kỹ thuật RAG

3. KẾT QUẢ VÀ THẢO LUẬN

Để đánh giá hiệu quả của hệ thống SmartMeal, nhóm nghiên cứu tiến hành các thử nghiệm thực nghiệm dựa trên bốn tiêu chí chính: (1) hiệu suất gợi ý thực đơn; (2) độ chính xác của mô hình nhận diện và phân tích hình ảnh thực phẩm; (3) khả năng tư vấn dinh dưỡng của Chatbot RAG; và (4) đánh giá trải nghiệm người dùng thông qua giao diện ứng dụng. Các tiêu chí đánh giá này được lựa chọn dựa trên các chuẩn mực phổ biến trong các lĩnh vực hệ thống gợi ý (NDCG, Mean Reward), thị giác máy tính (độ chính xác phân loại, sai số ước lượng calo, độ chính xác phát hiện/phân đoạn), và xử lý ngôn ngữ tự nhiên (độ chính xác trích xuất các thông tin liên quan tới câu hỏi, độ chính xác của câu trả lời, thời gian phản hồi).

3.1. Hiệu suất gợi ý thực đơn cá nhân hóa (PPO + NCF)

Hệ thống gợi ý được đánh giá thông qua chỉ số Normalized Discounted Cumulative Gain [18] (NDCG@3), chỉ số này sẽ cho thấy kết quả trong việc xếp hạng các món ăn phù hợp với từng loại người dùng và món ăn ở ngữ cảnh tĩnh. Cụ thể:

NDCG@3 dao động trong khoảng 0.896 đến 1.000, chứng tỏ chất lượng gợi ý thực phẩm cho các ngữ cảnh tĩnh rất cao. NDCG@K là một chỉ số đo lường chất lượng xếp hạng của một danh sách gợi ý, có tính đến vị trí của các mục phù hợp trong danh sách (mục phù hợp ở vị trí cao hơn có trọng số lớn hơn).

Mean Reward thu được sau huấn luyện PPO đạt mức 0.50. Mean Reward là giá trị trung bình của hàm phần thưởng được tính toán sau mỗi bước tương tác trong quá trình huấn luyện RL, phản ánh mức độ tối ưu hóa trung bình mà thuật toán PPO đạt được trong việc cân bằng các yếu tố của hàm phần thưởng (sở thích, dinh dưỡng, đa dạng, phản hồi người dùng). Kết quả này phản ánh sự cải thiện đáng kể trong khả năng cân bằng giữa các yếu tố như sở thích người dùng, giá trị dinh dưỡng và sự đa dạng thực đơn.

Một thách thức còn tồn tại là độ lệch trong việc đáp ứng chính xác lượng calo mong muốn, với sai lệch trung bình vào khoảng 421 kcal, chủ yếu đến từ sự ước lượng chưa hoàn hảo trong module phân tích hình ảnh.

3.2. Nhận diện và phân tích hình ảnh thực phẩm

Khả năng nhận diện và phân tích dinh dưỡng từ hình ảnh thực phẩm được kiểm tra bằng các mô hình học sâu và đánh giá qua độ chính xác:

ResNet-50 đạt độ chính xác 85% trong nhận diện tên món ăn từ ảnh. Độ chính xác này được tính toán trên một tập dữ liệu thử nghiệm riêng gồm các hình ảnh món ăn đã được gán nhãn chính xác.

YOLOv12 phát hiện vị trí thực phẩm đạt độ chính xác 70% trong các bức ảnh thực tế, bao gồm nhiều nền và ánh sáng khác nhau. Độ chính xác phát hiện (Precision) được tính dựa trên tỷ lệ các bounding box được phát hiện đúng so với tổng số bounding box được mô hình dự đoán trên tập dữ liệu thử nghiệm ảnh thực tế.

Segment Anything Model (SAM) kết hợp cùng YOLO cho phép phân đoạn chính xác đối tượng, hỗ trợ tính diện tích tương đối để ước lượng trọng lượng và calo.

Sai số trung bình trong ước lượng calo đạt khoảng $\pm 20\%$, ở mức có thể chấp nhận được cho các ứng dụng không chuyên sâu y tế.

3.3. Chatbot tư vấn dinh dưỡng (RAG)

Đánh giá được thực hiện bằng RAGAS, kết hợp các chỉ số như tính trung thực, mức độ liên quan của câu trả lời và ngữ cảnh, cùng với đánh giá thủ công bởi chuyên gia dinh dưỡng.

Độ chính xác tổng thể đạt 85% trên các metric chính của RAGAS trong các câu hỏi liên quan đến nhu cầu ăn uống, tương tác dinh dưỡng và tình trạng sức khỏe cá nhân. Kết quả này phản ánh khả năng của Chatbot trong việc truy xuất thông tin phù hợp từ cơ sở tri thức và sinh ra câu trả lời đúng đắn dựa trên ngữ cảnh đó.

Chatbot vẫn gặp hạn chế khi xử lý các câu hỏi định lượng phức tạp như so sánh chi tiết dinh dưỡng hoặc tính toán nhiều thông số, do mô hình ngôn ngữ không có khả năng suy luận số liệu chính xác nếu không được hỗ trợ bởi các hàm Function Calling chuyên biệt.

Chatbot hoạt động hiệu quả với truy vấn mô tả thông thường, nhưng còn hạn chế khi gặp câu hỏi yêu cầu tính toán định lượng hoặc suy luận số liệu chính xác do giới hạn của dữ liệu văn bản và khả năng prompt LLM. Thời gian phản hồi trung bình dưới 2 giây, phù hợp với yêu cầu thời gian thực.

3.4. Đánh giá giao diện và trải nghiệm người dùng

Ứng dụng SmartMeal được triển khai bằng Flutter, mang lại trải nghiệm giao diện mượt mà, thân thiện và tương thích đa nền tảng (Android, iOS). Đánh giá sơ bộ từ nhóm người dùng thử ($n = 30$) ghi nhận:

Tỷ lệ hài lòng tổng thể đạt 90% về tính dễ sử dụng, giao diện rõ ràng và tốc độ phản hồi.

Đa số người dùng bày tỏ mong muốn sử dụng lâu dài nếu hệ thống có thêm tính năng theo dõi cân nặng hoặc đồng bộ dữ liệu từ thiết bị đeo sức khỏe. Trong quá trình thử nghiệm, nhóm người dùng này bao gồm chủ yếu là sinh viên và nhân viên văn phòng có quan tâm đến việc theo dõi và cải thiện chế độ ăn uống. Nhóm người dùng được hướng dẫn cách sử dụng các tính năng chính như tạo hồ sơ dinh dưỡng, đặt mục tiêu, nhập liệu bữa ăn (thủ công hoặc qua ảnh), và tương tác với Chatbot (bao gồm cả việc hỏi về lịch sử ăn uống của bản thân). Dữ liệu đầu vào cho hệ thống bao gồm: (1) Hồ sơ người dùng: tên, tuổi, giới tính, cân nặng, chiều cao, mục tiêu sức khỏe, thói quen vận động, món ăn dị ứng, chế độ ăn. (2) Hình ảnh món ăn: chụp từ camera điện thoại, yêu cầu có vật thể chuẩn hóa (thẻ căn cước, thìa, đĩa tiêu chuẩn). (3) Câu hỏi dinh dưỡng: nhập dưới dạng tự nhiên (ví dụ ‘Tôi ăn 1 bát phở thì bao nhiêu calo?’). Người dùng có thể tương tác với Chatbot hoặc sử dụng chức năng gợi ý thực đơn theo từng bữa. Lịch sử tương tác/đánh giá món ăn (được thu thập trong quá trình thử nghiệm cho NCF), hình ảnh món ăn (do người dùng chụp và tải lên cho module phân tích ảnh), và các câu hỏi về

đinh dưỡng (do người dùng đặt cho Chatbot, bao gồm các câu hỏi truy vấn lịch sử cá nhân). Hệ thống SmartMeal được thiết kế để hỗ trợ nhiều đối tượng người dùng khác nhau với các mục tiêu khác nhau (người muốn giảm cân, tăng cân, ăn uống lành mạnh nói chung) bằng cách cho phép tùy chỉnh mục tiêu dinh dưỡng trong hồ sơ cá nhân. Tuy nhiên, phần thử nghiệm ban đầu tập trung vào việc đánh giá tính năng cốt lõi. Người dùng thử cũng đã đưa ra góp ý về việc làm rõ hơn quy trình nhập liệu qua ảnh, bổ sung hướng dẫn sử dụng cho người dùng có nhu cầu dinh dưỡng đặc biệt (ví dụ: phụ nữ mang thai, người ăn chay, người bị tiểu đường), và làm rõ nguồn dữ liệu/cơ sở khoa học cho các khuyến nghị dinh dưỡng.

Các chỉ số trong Bảng 1 được lựa chọn dựa trên thông lệ đánh giá trong từng lĩnh vực: (1) NDCG@3: đánh giá độ phù hợp của top 3 thực đơn gợi ý với sở thích người dùng. (2) Sai số ước lượng calo: đo mức chênh lệch so với giá trị chuẩn. (3) RAGAS: đánh giá độ chính xác của Chatbot trong truy xuất ngữ cảnh và sinh câu trả lời. (4) Tỷ lệ hài lòng người dùng: đánh giá cảm nhận qua khảo sát định tính từ người dùng thử.

Dưới đây là bảng tổng hợp kết quả thực nghiệm các module trong hệ thống SmartMeal, bao gồm các chỉ số đánh giá đã được thảo luận chi tiết ở trên:

Bảng 1. Tổng hợp kết quả thực nghiệm các module trong hệ thống SmartMeal

Module	Chỉ số đánh giá	Kết quả đạt được	Ghi chú
Gợi ý thực đơn (PPO + NCF)	NDCG@3	0.896 – 1.000	Chất lượng gợi ý cao trong top 3
	Mean Reward sau huấn luyện PPO	0.50	Phản ánh khả năng cân bằng mục tiêu
	Sai lệch calo trung bình (kcal)	~421 kcal	Cần tối ưu thêm trong ước lượng ảnh
Nhận diện món ăn (ResNet-50)	Tỷ lệ chính xác phân loại món ăn	85%	Đánh giá trên tập dữ liệu thực phẩm
Phát hiện thực phẩm (YOLOv12)	Tỷ lệ chính xác phát hiện bounding box	70%	Đánh giá trên ảnh thực tế đa dạng điều kiện
Phân đoạn thực phẩm (SAM)	Khả năng phân đoạn hình ảnh chính xác tương đối	Cao	Đánh giá dựa trên mask phân đoạn, hỗ trợ ước lượng diện tích
Ước lượng calo (SAM + YOLO)	Sai số trung bình ước lượng calo	±20%	So sánh với giá trị thực tế, sử dụng vật thể tham chiếu
Chatbot tư vấn dinh dưỡng (RAG + Functions)	Tỷ lệ chính xác RAGAS	85%	Đánh giá bằng framework RAGAS. Giới hạn với câu hỏi định lượng phức tạp.
	Thời gian phản hồi trung bình	< 2 giây	Đáp ứng yêu cầu thời gian thực (có thể tăng nhẹ khi gọi hàm)

Module	Chỉ số đánh giá	Kết quả đạt được	Ghi chú
Trải nghiệm người dùng (Flutter)	Tỷ lệ hài lòng tổng thể	90%	Đánh giá từ 30 người dùng thử nghiệm về tính dễ sử dụng, giao diện, tốc độ.

4. KẾT LUẬN

Nghiên cứu này đã phát triển thành công một hệ thống quản lý dinh dưỡng thông minh tích hợp nhiều công nghệ trí tuệ nhân tạo hiện đại, bao gồm học tăng cường (Proximal Policy Optimization - PPO), lọc cộng tác thần kinh (Neural Collaborative Filtering - NCF), các mô hình học sâu cho phân tích hình ảnh (ResNet, YOLO, Segment Anything Model - SAM) và Chatbot tư vấn dinh dưỡng dựa trên Retrieval-Augmented Generation (RAG). Thông qua việc thiết kế hệ thống gợi ý đa bước và tương tác, ứng dụng SmartMeal không chỉ cung cấp thực đơn cá nhân hóa phù hợp với sở thích, nhu cầu và mục tiêu sức khỏe của người dùng, mà còn hỗ trợ nhận diện thực phẩm, ước lượng calo chính xác từ hình ảnh thực tế và đưa ra tư vấn dinh dưỡng theo ngữ cảnh một cách linh hoạt.

Các kết quả thực nghiệm đã minh chứng hiệu quả của hệ thống trên nhiều khía cạnh: độ chính xác cao trong gợi ý ($NDCG@3 \approx 0.9-1.0$), khả năng nhận diện món ăn đạt 85% (ResNet), và độ chính xác khi trả lời của Chatbot đạt 85%. Quá trình phát triển cũng gặp một số thách thức kỹ thuật thực tế, bao gồm việc thu thập và gán nhãn dữ liệu đa dạng và số lượng lớn cho các hình ảnh (đặc biệt là các ảnh món ăn và các dữ liệu dinh dưỡng của món ăn Việt Nam), tối ưu hàm phần thưởng phức tạp trong RL để cân bằng các yếu tố mục tiêu một cách hiệu quả, và tinh chỉnh mô hình ngôn ngữ cho Chatbot sao cho xử lý tốt hơn các câu hỏi định lượng và duy trì ngữ cảnh dài.

Những kết quả và kinh nghiệm thu được khẳng định tiềm năng của SmartMeal trong việc góp phần cải thiện sức khỏe cộng đồng thông qua cá nhân hóa chế độ ăn uống bằng công nghệ AI và thể hiện ưu thế vượt trội trong các tính năng dựa trên AI (gợi ý động, phân tích ảnh, tư vấn ngữ cảnh/truy xuất dữ liệu cá nhân) so với các ứng dụng quản lý dinh dưỡng truyền thống chủ yếu dựa vào nhập liệu thủ công và cơ sở dữ liệu tĩnh.

Một số thách thức kỹ thuật khi triển khai hệ thống bao gồm: (1) Việc thu thập dữ liệu món ăn Việt Nam với nhãn đầy đủ là khó khăn do thiếu tập dữ liệu công khai, đòi hỏi xử lý thủ công. (2) Sai số trong việc ước lượng calo từ ảnh chịu ảnh hưởng bởi chất lượng ảnh, góc chụp, ánh sáng, và vật thể tham chiếu. (3) Chatbot RAG gặp hạn chế trong các truy vấn cần suy luận định lượng sâu (ví dụ: so sánh chi tiết vi chất dinh dưỡng), do các mô hình ngôn ngữ hiện tại chưa thể xử lý phép tính số học phức tạp nếu không có hỗ trợ từ các hàm ngoại.

Trong tương lai, nghiên cứu sẽ tập trung vào ba hướng phát triển chính nhằm nâng cao hiệu quả và khả năng mở rộng của hệ thống:

Tối ưu hóa kỹ thuật: Nâng cao chất lượng phần thưởng trong mô hình PPO thông qua kỹ thuật Inverse Reinforcement Learning (IRL) để học phần thưởng từ hành vi lý tưởng thay vì thiết kế thủ công. Áp dụng các mô hình tiên tiến như Vision Transformer (ViT) cho nhận diện hình ảnh và Soft Actor-Critic (SAC) để cải thiện độ ổn định trong học tăng cường. Nghiên cứu huấn luyện embedding chuyên biệt để cải thiện độ chính xác truy xuất và sinh văn bản trong Chatbot RAG.

Mở rộng dữ liệu và phạm vi ứng dụng: Thu thập và mở rộng tập dữ liệu món ăn Việt Nam, kết hợp cả món ăn đường phố và thực đơn đặc biệt theo vùng miền. Tích hợp dữ liệu từ các thiết bị đeo tay (wearables) để theo dõi lượng tiêu hao calo và điều chỉnh thực đơn phù

hợp. Phát triển phiên bản đa ngôn ngữ (tiếng Anh, Hàn, Nhật...) để mở rộng phạm vi người dùng toàn cầu.

Tăng cường trải nghiệm và giá trị người dùng: Tích hợp tính năng gợi ý theo ngân sách tài chính, thời gian chế biến, hoặc tình trạng nguyên liệu sẵn có trong nhà bếp. Thêm khả năng tự học (self-learning) của Chatbot từ phản hồi người dùng, qua đó cải thiện trải nghiệm hội thoại và độ tin cậy theo thời gian. Kết nối hệ thống với cộng đồng dinh dưỡng, chuyên gia sức khỏe để cung cấp tư vấn cá nhân hóa sâu hơn.

Trong các phiên bản mở rộng, hệ thống có thể bổ sung các chế độ ăn đặc biệt như: chế độ ăn cho phụ nữ mang thai, người ăn chay, hoặc bệnh nhân tiểu đường. Mỗi chế độ sẽ được cá nhân hóa dựa trên hồ sơ người dùng và khuyến nghị từ chuyên gia, giúp tăng độ chính xác và ứng dụng thực tiễn của hệ thống.

TÀI LIỆU THAM KHẢO

- [1] Báo VnExpress, "Số người Việt béo phì tăng nhanh nhất Đông Nam Á," [Online]. Available: <https://vnexpress.net/so-nguoi-viet-beo-phi-tang-nhanh-nhat-dong-nam-a-4635765.html>, August. 2023.
- [2] GeeksforGeeks, "What are Recommender Systems?". Available: <https://www.geeksforgeeks.org/what-are-recommender-systems/>, January. 2024.
- [3] C. P. Hughes, "Understanding PPO: A Game-Changer in AI Decision-Making Explained for RL Newcomers" Medium, August. 2023. Available: <https://medium.com/@chris.p.hughes10/understanding-ppo-a-game-changer-in-ai-decision-making-explained-for-rl-newcomers-913a0bc98d2b>, June. 2024.
- [4] Y. Zhang, Q. Liu, E. Chen, Z. Huang, Q. Liu, and Y. Su, "An interactive food recommendation system using reinforcement learning" Expert Syst. Appl., vol. 253, p. 124313, October. 2024. doi: 10.1016/j.eswa.2024.124313.
- [5] GeeksforGeeks, "Cross-Attention Mechanism in Transformers," GeeksforGeeks, February. 2024. Available: <https://www.geeksforgeeks.org/cross-attention-mechanism-in-transformers/>, April. 2025.
- [6] B. Tjahjono, P. D. Ball, V. I. Vitanov, Y. V. довиков, and W. Sutopo, "Learn supply chain in the Industry 4.0 era," Int. J. Lean Six Sigma, vol. 12, no. 1, pp. 59-82, 2020.
- [7] J. Shulman, "Reinforcement Learning from Human Feedback (RLHF)," Hugging Face, July. 2023. Available: <https://huggingface.co/blog/rlhf>, April. 2025.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.
- [9] R. Potdar, "101 Food Images Dataset," Kaggle, 2018. Available: <https://www.kaggle.com/datasets/dansbecker/food-101/>, March. 2025.
- [10] C. K. Tran, "Vietnamese Food Classifier using fast.ai," chriskhanhtran.github.io, August. 2021. Available: <https://chriskhanhtran.github.io/posts/vn-food-classifier/>, March. 2025.
- [11] Wikipedia contributors, "You Only Look Once," in Wikipedia, 2024. Available: https://en.wikipedia.org/wiki/You_Only_Look_Once, March. 2025.
- [12] Ultralytics, "Các mô hình YOLOv12," Ultralytics Docs, November. 2023. Available: <https://docs.ultralytics.com/vi/models/yolo12/>, March. 2025.
- [13] Meta AI, "Segment Anything Model (SAM)," V7 Labs, April. 2023. Available: <https://www.v7labs.com/blog/segment-anything-model-sam>, March. 2025.
- [14] M. Tuan, "ChatGPT Series #5 - Tìm hiểu về Retrieval-Augmented Generation (RAG)," Viblo, December. 2023. Available: <https://viblo.asia/p/chatgpt-series-5-tim-hieu-ve-retrieval-augmented-generation-rag-Ny0VGRd7LPA>, March. 2025.
- [15] P. Tran, "Xây dựng hệ thống RAG tích hợp Function Calling (with source code)," Viblo, June. 2024. Available: <https://viblo.asia/p/quy-trinh-xay-dung-he-thong-rag-tich-hop-function-calling-with-source-code-vlZL98GZJQK>, April. 2025.

- [16] P. D. Tường, "Dinh dưỡng và một số vấn đề sức khỏe cộng đồng.", September. 2024. Available: <https://pharma360.vn/huong-nghep/pdf-dinh-nuong-va-mot-so-van-de-suc-khoe-cong-dong-pham-duy-tuong-25987/>, October. 2024.
- [17] Bộ Y Tế Viện Dinh Dưỡng Quốc Gia, "Bảng thành phần dinh dưỡng thực phẩm Việt Nam". Available: https://www.fao.org/fileadmin/templates/food_composition/documents/pdf/VTN_FCT_2007.pdf, 2007.
- [18] GeeksforGeeks, "Normalized Discounted Cumulative Gain – Multilabel Ranking Metrics | ML," GeeksforGeeks, July. 2021. Available: <https://www.geeksforgeeks.org/cross-attention-mechanism-in-transformers/>, April. 2025.

ABSTRACT

SMART NUTRITION MANAGEMENT SYSTEM

Authors: Vo Ngoc Nguyen Chuong¹, Tran Tri Cuong¹, Pham Vu Luong Bang¹,
Tran Vinh Hien¹, Do Quang Vinh¹, Tran Van Tho^{1*}

¹*Ho Chi Minh City University of Industry and Trade; 140 Le Trong Tan, Tay Thanh Ward,
Tan Phu District, Ho Chi Minh City*

*Email: thotv@huit.edu.vn

The increasing prevalence of nutrition-related diseases demands effective and personalized dietary management solutions. This study presents the development and evaluation of SmartMeal, an intelligent nutrition management system - a cross-platform mobile application designed to support users in building and maintaining a healthy lifestyle. The system integrates advanced Artificial Intelligence technologies: Reinforcement Learning (RL) using the Proximal Policy Optimization (PPO) algorithm combined with Neural Collaborative Filtering (NCF) to provide multi-step, personalized, and dynamically adaptive meal recommendations; Deep Learning models (ResNet, YOLO, Segment Anything Model - SAM) to accurately recognize dishes and estimate calories from images; and an intelligent nutrition advisory Chatbot based on Retrieval-Augmented Generation (RAG) technology to deliver reliable information. The system is implemented on the Flutter platform, offering an intuitive and seamless user experience. Experimental results demonstrate high performance in recommendation ($NDCG@3 \approx 0.9-1.0$), food recognition (ResNet accuracy 85%), calorie estimation (mean error $\approx 20\%$), and nutritional question answering (RAG accuracy 85%). Despite some limitations in RL reward optimization and data quality dependency, SmartMeal shows great potential in enhancing public health through intelligent and personalized nutrition management.

Keywords: Nutrition management, Reinforcement Learning, PPO, Image recognition, Calorie estimation, RAG Chatbot, Personalization.

Tôi xin cam kết bài báo này chưa được đăng và gửi đăng ở bất kỳ tạp chí nào khác.

Tác giả:

Họ và tên: Trần Văn Thọ

Số điện thoại liên hệ: 0989282536