

TRÍCH YẾU LUẬN ÁN TIẾN SĨ

Tên luận án: “KHAI PHÁ QUAN ĐIỂM VỚI CÁC PHƯƠNG PHÁP QUẦN THỂ”

Tác giả: Trần Quang Phúc

Ngành: Khoa học máy tính

Mã số: 9480101

Cơ sở đào tạo: Trường Đại học Bách khoa – Đại học Đà Nẵng

1. Mục tiêu và đối tượng nghiên cứu của luận án

1.1. Mục tiêu của luận án

Mục tiêu tổng quát của luận án là nghiên cứu và đề xuất các mô hình khai phá quan điểm dựa trên các phương pháp học quần thể nhằm khắc phục những hạn chế của các mô hình khai phá quan điểm truyền thống. Các mục tiêu cụ thể bao gồm:

- Khảo sát tổng quan các nghiên cứu và hệ thống khai phá quan điểm, làm rõ những thành tựu và hạn chế hiện nay.
- Nghiên cứu mô hình quan điểm có cấu trúc dưới dạng bộ não (thực thể, khía cạnh, tình cảm, người đưa ra quan điểm, thời gian) và các kỹ thuật trích xuất bộ não từ dữ liệu văn bản.
- Đề xuất mô hình khai phá quan điểm theo ảnh hưởng kết hợp nhằm nâng cao độ chính xác, khả năng khái quát hóa và khả năng diễn giải.
- Xây dựng mô hình khai phá quan điểm dựa trên rừng phân loại, rừng mật độ và rừng đa tạp, phù hợp với đặc trưng dữ liệu quan điểm có cấu trúc.
- Đánh giá và so sánh các mô hình đề xuất với các phương pháp khai phá quan điểm hiện có trên các tập dữ liệu thực tế.

1.2. Đối tượng nghiên cứu

Đối tượng nghiên cứu của luận án bao gồm:

- Dữ liệu văn bản chứa quan điểm của người dùng.
- Quan điểm ở các mức độ khác nhau, trong đó tập trung vào mức độ khía cạnh.
- Các kỹ thuật trích xuất và biểu diễn dữ liệu quan điểm dưới dạng bộ não.
- Các phương pháp học quần thể, đặc biệt là các mô hình rừng, áp dụng cho khai phá quan điểm.

2. Các phương pháp nghiên cứu đã sử dụng

- Phương pháp nghiên cứu tài liệu: phân tích, tổng hợp các công trình nghiên cứu liên quan đến khai phá quan điểm, học quần thể và xử lý ngôn ngữ tự nhiên.

- Phương pháp thực nghiệm: xây dựng, cài đặt và đánh giá các mô hình đề xuất trên dữ liệu quan điểm thực tế (bằng ngôn ngữ R và Python) nhằm kiểm chứng hiệu quả của phương pháp.

3. Những đóng góp chính

3.1. Đóng góp khoa học

- Đề xuất mô hình Rừng phân loại quan điểm (F_{CF}) dựa trên biểu diễn ma trận bộ năm quan điểm và entropy Shannon có giám sát, bảo toàn cấu trúc ngữ nghĩa khi xây dựng cây phân loại, tích hợp giải thích qua tầm quan trọng đặc trưng và biểu đồ phụ thuộc từng phần (PDP).
- Đề xuất mô hình Rừng mật độ quan điểm (F_{DF}) tích hợp DBSCAN, entropy Gaussian và SHAP nhằm xử lý dữ liệu mất cân bằng theo mật độ phân bố và diễn giải tầm quan trọng, tương tác giữa các thành phần bộ năm.
- Đề xuất mô hình Rừng đa tập quan điểm (F_{MF}) học trực tiếp cấu trúc đa tập từ tính đồng xuất hiện tại nút lá của rừng ngẫu nhiên không giám sát (ma trận quan hệ W , khoảng cách đa tập $dist_M$), không cần giải bài toán trị riêng toàn cục như ISOMAP/UMAP, phân cụm DBSCAN trên không gian đa tập và tổng hợp kết quả qua Stacking.

3.2. Đóng góp thực tiễn

- Xây dựng quy trình chuyển đổi bốn bộ dữ liệu đánh giá gốc (hai bộ tiếng Anh yelpHotelData, yelpResData - dữ liệu công bố của Mukherjee và cộng sự; hai bộ tiếng Việt VLSP-2018 Restaurant, VLSP-2018 Hotel – dữ liệu của cộng đồng VLSP-2018) sang cấu trúc bộ năm quan điểm chuẩn hóa, phục vụ huấn luyện và đánh giá mô hình; mô hình trích xuất bộ năm quan điểm từ văn bản đạt $F1 = 0,78$ trên tập Twitter US Airline Sentiment.
- Cung cấp cơ sở thực nghiệm so sánh ba mô hình rừng đề xuất với các mô hình học máy, học sâu và Transformer hiện có (Logistic Regression, SVM, Random Forest, Gradient Boosting, LSTM, BiLSTM-Attention, BERT, RoBERTa, PhoBERT), hỗ trợ lựa chọn mô hình phù hợp cho các hệ thống khai phá quan điểm có tài nguyên tính toán hạn chế và yêu cầu khả năng diễn giải cao.

4. Kết quả chính

- Mô hình Rừng phân loại quan điểm (F_{CF}) đạt độ chính xác 0,77 trên yelpHotelData và yelpResData (số liệu đã công bố, chia ngẫu nhiên theo dòng bộ năm), diễn giải tốt qua tầm quan trọng đặc trưng và PDP; kiểm chứng thêm trên Amazon Food Reviews và TripAdvisor Hotel Reviews đạt Accuracy 0,80–0,82 và 0,91 (kèm hạn chế Recall 0,50 trên TripAdvisor do mất cân bằng nhãn).

- Mô hình Rừng mật độ quan điểm (F_{DF}) đạt độ chính xác trung bình 0,78 (số liệu đã công bố) mà không cần GPU; giá trị SHAP xác định thực thể (Entity) là đặc trưng chủ đạo (trung bình $|SHAP| \approx 0,087$).
- Mô hình Rừng đa tạp quan điểm (F_{MF}) đạt độ chính xác trung bình cao nhất trong ba mô hình đề xuất (0,80, số liệu đã công bố), chạy trên CPU và diễn giải được qua ma trận quan hệ W và độ lợi thông tin trên đa tạp $IG_{\mathcal{M}}$.
- Xây dựng thành công quy trình chuyển đổi dữ liệu đánh giá gốc sang cấu trúc bộ năm quan điểm (e, a, s, h, t) trên bốn bộ dữ liệu (yelpHotelData, yelpResData, VLSP-2018 Restaurant, VLSP-2018 Hotel); trên VLSP-2018, F_{DF} đạt Balanced Accuracy 0,724 (Hotel) và 0,660 (Restaurant), trong khi mốc tham chiếu PhoBERT đọc văn bản gốc đạt 0,832 và 0,791.
- Khi đánh giá lại theo quy trình chia theo đánh giá gốc (5 hạt giống $\times$ 10.000 đánh giá), Balanced Accuracy của mọi mô hình chỉ dùng bộ năm giảm về khoảng 0,51–0,57: F_{CF} và F_{DF} tương đương Random Forest, F_{MF} thấp hơn Random Forest khoảng 0,02–0,03; các mô hình đọc văn bản gốc (BERT, RoBERTa) đạt 0,76–0,82. Luận án định lượng mức rò rỉ dữ liệu khi chia theo dòng bộ năm (Accuracy 0,946 so với 0,611 với cùng Random Forest) và khuyến nghị chia theo đánh giá gốc.

Chuỗi tiến triển độ chính xác theo số liệu đã công bố từ Rừng phân loại quan điểm (0,77) đến Rừng mật độ quan điểm (0,78) và Rừng đa tạp quan điểm (0,80) phản ánh việc khai thác cấu trúc dữ liệu ngày càng tinh tế, đồng thời cả ba mô hình duy trì khả năng diễn giải qua cấu trúc cây, cụm và ma trận quan hệ. Như vậy, đóng góp của ba mô hình rừng nằm chủ yếu ở cách biểu diễn bộ năm có cấu trúc và khả năng diễn giải; độ chính xác bị giới hạn bởi lượng thông tin của bốn thành phần (e, a, h, t).

Đà Nẵng, ngày tháng năm 2026

Xác nhận của người hướng dẫn khoa học

Người hướng dẫn 1

Người hướng dẫn 2

Nghiên cứu sinh

GS.TS. Huỳnh Xuân Hiệp

TS. Lê Thị Mỹ Hạnh

Trần Quang Phúc