TRANG THÔNG TIN LUẬN ÁN TIẾN SĨ

Tên đề tài luận án tiến sĩ:

NGHIÊN CỨU PHƯƠNG PHÁP XÁC ĐỊNH THỨ TỰ ƯU TIÊN CỦA THƯ ĐIỆN TỬ

Chuyên ngành: **Hệ thống thông tin**

Mã số: **9.48.01.04**

Họ và tên NCS: **Nguyễn Thanh Hà**

Người hướng dẫn khoa học:

1. PGS.TS. Trần Quang Anh

2. TS. Trần Hùng

Cơ sở đào tạo: Học viện Công nghệ Bưu chính Viễn thông

NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:

Luận án đề cập tổng quát về thư điện tử và xác định thứ tự ưu tiên của thư điện tử, các vấn đề còn tồn tại của bài toán xác định thứ tự ưu tiên của thư điện tử, từ đó đề xuất xây dựng tập dữ liệu thư điện tử tiếng Việt và một số phương pháp xác định thứ tự ưu tiên của thư điện tử tiếng Việt. Luận án có ý nghĩa trong thực tiễn dựa trên vai trò thiết yếu của thư điện tử và tác hại của hai vấn đề thư rác và quá tải thư điện tử trong đời sống. Luận án ý nghĩa về mặt khoa học dựa trên việc phân tích các vấn đề còn tồn tại trong các phương pháp cũ, từ đó đề xuất phương pháp mới để giải quyết các vấn đề đó. Hiệu quả của các đề xuất đối với việc giải quyết vấn đề thư rác và quá tải thư điện tử được đánh giá dựa trên thí nghiệm khoa học. Đóng góp mới của quá trình nghiên cứu thể hiện trong luận án như sau:

(1) Đề xuất phương pháp xây dựng tập luật lọc thư rác dành cho nền tảng SpamAssassin dựa trên mạng nơ-ron. Các phương pháp xây dựng tập luật SpamAssassin dựa trên học máy đều có hai bước lựa chọn đặc trưng và huấn luyện trọng số riêng biệt. Với cách làm này, hiệu quả của tập đặc trưng được chọn chưa được kiểm chứng trên dữ liệu. Mô hình mạng nơ-ron được đề xuất có khả năng hợp nhất hai khâu lựa chọn luật và gán điểm số, giúp nâng cao chất lượng của tập đặc trưng được chọn, từ đó nâng cao chất lượng của tập luật được xây dựng. Phương pháp có hiệu quả dự đoán ưu việt so với những phương pháp cũ.

(2) Đề xuất phương pháp sinh tập luật lọc thư rác cho SpamAssassin dựa trên phương pháp tối ưu đa mục tiêu. Trong phương pháp tối ưu hóa đơn mục tiêu, việc gán điểm số và tìm giá trị ngưỡng của tập luật được thực hiện tuần tự. Phương pháp này cho phép thực hiện hai tác vụ này đồng thời, từ đó giải quyết vấn đề quan trọng của bài toán sinh tập luật

SpamAssassin đó là sự cân bằng giữa hai tiêu chí đối nghịch là *recall* và FAR.

(3) Đề xuất phương pháp dự đoán hành động cho người dùng thư điện tử trên nền tảng SpamAssassin bằng mô hình phân loại đa lớp. Hệ thống SpamAssassin là hệ thống lọc thư rác phổ biến nhưng không có sẵn chức năng gợi ý hành động đối với thư điện tử. Luận án đã ứng dụng cách kỹ thuật chuyển đổi từ mô hình phân loại nhị phân thành mô hình phân loại đa lớp để bổ sung tính năng dự đoán hành động cho SpamAssassin. Phương pháp này có tính ứng dụng cao trên thực tế bởi vì sự phổ biến của hệ thống SpamAssassin và tốc độ xử lý nhanh của cơ chế luật có trọng số của SpamAssassin. Luận án cũng trình bày hai phương án nhằm cải thiện hiệu quả của mô hình dự đoán hành động nói trên dựa trên cải thiện hiệu quả của các máy phân loại nhị phân thành phần. Phương án thứ nhất ứng dụng luật ham cho tập luật SpamAssassin, giúp giảm tỷ lệ gợi ý nhầm đối với hành động xóa thư. Phương án thứ hai ứng dụng phương pháp sinh tập luật SpamAssassin dựa trên mạng nơ-ron từ nghiên cứu đã công bố số 2, giúp tăng độ chính xác chung của các gợi ý.

(4) Đề xuất phương pháp xếp hạng thư điện tử tiếng Việt theo hướng phân loại. Mô hình phân loại được đề xuất đã sử dụng các kỹ thuật học sâu, bao gồm cấu trúc mạng LSTM, bộ đặc trưng nội dung biểu diễn bằng kỹ thuật word2vec kết hợp với đặc trưng xã hội được trích xuất bằng một phương pháp mới. Phương pháp đề xuất có hiệu quả tốt hơn đáng kể so với phương pháp dựa trên máy phân loại SVM và bộ đặc trưng TF-IDF trước đó dành cho bài toán xếp hạng thư điện tử.

CÁC ỨNG DỤNG, KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HOẶC NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU:

Các kết quả nghiên cứu của luận án có khả năng ứng dụng trong các hệ thống máy chủ thư điện tử, đặc biệt là các hệ thống mà người sử dụng Việt Nam chiếm đa số. Có thể liệt kê ứng dụng thực tế của những đề xuất trong luận án như sau:

(1) Cải thiện độ chính xác của bộ lọc thư rác SpamAssassin trên các hệ thống máy chủ thư điện tử. Tập luật SpamAssassin được xây dựng bằng các đề xuất trong luận án có thể được sử dụng trực tiếp trên SpamAssassin.

(2) Bổ sung tính năng dự đoán hành động người dùng cho các hệ thống thư điện tử sử dụng bộ lọc thư rác SpamAssassin. Để triển khai đề xuất của luận án, cần phải điều chỉnh mã nguồn của hệ thống thư điện tử để thực thi các tập luật SpamAssassin theo thuật toán đề xuất thay vì thực thi SpamAssassin theo cách thông thường, đồng thời điều chỉnh ứng dụng hòm thư của người dùng để thể hiện kết quả dự đoán. Phương pháp phù hợp và dễ triển khai với các hệ thống thư điện tử mã nguồn mở.

(3) Bổ sung tính năng xếp hạng thư điện tử. Cần xây dựng ứng dụng độc lập cũng như điều chỉnh mã nguồn của ứng dụng hòm thư của người dùng để thực thi ứng dụng độc lập đó và thể hiện kết quả xếp hạng thư điện tử cho người dùng.

Xác nhận của đại diện tập thể **Người hướng dẫn khoa học**	**Nghiên cứu sinh**
PGS.TS. Trần Quang Anh	**Nguyễn Thanh Hà**