GIẢI BÀI TOÁN PHÂN LỚP ĐA NHÃN

Đề tài "A multi-label classification algorithm and its application to reputation management on Vietnamese hotel domain" (Một thuật toán phân lớp đa nhãn và ứng dụng của nó trong quản lý danh tiếng trong miền dữ liệu tiếng Việt về khách sạn) của nhóm sinh viên nghiên cứu khoa học tại Phòng thí nghiệm Khoa học Dữ liệu và Công nghệ Tri thức (Khoa Công nghệ thông tin, Trường Đại học Công nghệ) đã đoạt Giải Ba giải thưởng Sinh viên nghiên cứu khoa học cấp Bộ GD&ĐT và Giải Nhì cấp ĐHQGHN.

■ TUYẾT NGA

KẾ THỪA VÀ PHÁT HUY Ý TƯỞNG

Hiện nay, sự quan tâm dành cho việc khai thác tri thức từ dữ liệu lớn ngày càng tăng, đặc biệt là các giải thuật học sâu, nhưng hướng tiếp cận này đòi hỏi lượng dữ liệu đủ lớn, và khả năng tính toán cao để sinh ra mô hình. Công việc xử lý lượng dữ liệu có nhãn lớn đòi hỏi nhiều thời gian và công sức do việc gắn nhãn thường được tiến hành thủ công. Trước thực tế như vậy, nhóm sinh viên tham gia nghiên cứu khoa học tại Phòng thí nghiệm Khoa học Dữ liệu và Công nghệ Tri thức (DS&KTLab) gồm Nguyễn Văn Quang (K57CA), Trần Thị Minh Tươi (K59T) và Trần Thị Yến (K58T) đã chung tay thành lập nhóm nghiên cứu. Việc này được thầy Hà Quang Thụy và cô Phạm Thị Ngân ủng hộ và khuyến khích các thành viên làm việc theo nhóm.

Nguyễn Văn Quang cho biết, nhóm nghiên cứu đã bắt đầu triển khai đề tài này từ tháng 7 năm 2016 nhằm hướng đến một cách giải quyết khác cho bài toán phân lớp đa nhãn ở một góc nhìn mới hơn. Đó là xây dựng và nâng cấp một mô hình phân lớp đa nhãn dựa trên giải pháp bán giám sát vừa tận dụng được sự sẵn có của dữ liệu không nhãn cũng như khai thác được sự tương quan giữa các nhãn với nhau. Mỗi mẫu có thể có một hoặc nhiều nhãn. Ví dụ, một bức ảnh phong cảnh có thể gồm một hoặc nhiều đối tượng như cây cối, động vật, sông, hay con người. Nhiệm vụ của mô hình phân lớp đa nhãn là gán cho một mẫu quan sát chưa biết các nhãn được định nghĩa trước càng chính xác càng tốt. Phương pháp đề xuất này có thể ứng dụng trên nhiều miền dữ liệu khác nhau. Cụ thể ở đây, nhóm tích hợp vào bài toán quản lý danh tiếng khách sạn. Với các dữ liệu bình luận, nhận xét về khách sạn trên website hay các mạng xã hội mô hình sẽ phân loại xem chủ đề họ đang nói liên quan đến những khía cạnh nào về khách sạn như giá cả, vị trí, chất lượng phục vụ, đồ ăn... Từ đó, chúng ta xem xét và dự đoán quan điểm trong mỗi khía cạnh đấy là tích cực hay tiêu cực. Điều này giúp các khách sạn có thể nắm bắt được thông tin phản hồi của khách hàng để có thể nâng cao chất lượng dịch vụ.

Trước đó hai năm, cũng tại phòng thí nghiệm DS&KTLab đã có nhóm sinh viên nghiên cứu về bài toán phân lớp đa nhãn. Tuy nhiên, Quang khẳng định điểm nổi bật của đề tài hiện tại là đã khai thác được dữ liệu không nhãn, giải quyết được một số những vấn đề gặp phải trong bài toán phân lớp đa nhãn như không gian nhãn. Kết quả thực nghiệm cũng cho thấy độ chính xác của mô hình phân lớp tăng lên đáng kể. Những kết quả thu được từ đề tài này đều có đóng góp vào công bố khoa học tại hội nghị khoa học quốc tế về các hệ thống thông tin và dữ liệu thông minh 2017 và tạp chí Journal of Information and Telecomunication (Taylor Francis). Bên cạnh đó, phương pháp phân lớp đa nhãn được đề xuất cũng được nhóm đưa vào áp

dụng và thử nghiệm trên miền dữ liệu tiếng Việt về khách sạn với vai trò theo dõi và quản lý danh tiếng từ các nguồn bình luận nhận xét trên mạng. Hiện tại, đề tài đang được mở rộng nghiên cứu để công bố khoa học và tiếp tục triển khai trên nhiều miền dữ liệu.

CHÌA KHÓA THÀNH CÔNG

Khó khăn gặp phải trong quá trình nghiên cứu được Quang chia sẻ là mỗi thành viên trong nhóm chưa có nhiều kinh nghiệm về nghiên cứu khoa học, vốn kiến thức nền tảng còn hạn chế. Để vượt qua những trở ngại về kiến thức nhóm đã tích cực củng cố và trang bị thêm nhiều kiến thức và kĩ năng. Sự tư vấn và hướng dẫn của thầy Hà Quang Thụy là những định hướng rất quý báu với các thành viên trong thời gian đầu tham gia nghiên cứu khoa học. Hàng tuần, nhóm đều tổ chức seminar do từng thành viên trong nhóm tự trình bày cho các thành viên khác nghe về chủ đề hay bài toán nào đó được phân công tìm hiểu. Sự tích cực học tập, nghiên cứu từ mỗi cá nhân và tập thể chính là yếu tố quyết định cho những kết quả đạt được.

Hầu hết các thành viên trong nhóm đều là sinh viên năm ba và năm cuối. Khối lượng bài tập và các đồ án trên lớp tương đối nhiều nên việc quan trọng nhất của nhóm là phân bố thời gian tham gia nghiên cứu. Để duy trì tham gia nghiên cứu khoa học hiệu quả, Quang chia sẻ, nhóm chủ động phối hợp công việc, họp nhóm và báo cáo kết quả nghiên cứu hàng tuần dưới sự hướng dẫn trực tiếp của thầy/cô để tiến độ công việc được cập nhật liên tục. Khi nhóm gặp phải vấn đề khó khăn trong việc triển khai ý tưởng thì thầy/cô hướng dẫn là người cố vấn và phân tích những hạn chế của hệ thống, đồng thời là người đưa ra lời khuyên, định hướng cụ thể. Ngoài những kiến thức về chuyên môn, thầy Hà Quang Thuy còn là người quan tâm, động viên và tận tình giúp đỡ trong những vấn đề cá nhân của các thành viên trong nhóm.

Sau hơn một năm tham gia nghiên cứu khoa học tại phòng thí nghiệm DS&KTLab thì sự đam mê và tình yêu trong nghiên cứu khoa học của Quang cũng như các thành viên trong nhóm nghiên cứu ngày càng cụ thể, rõ ràng hơn. Trước đó, khái niệm nghiên cứu khoa học với cả nhóm vẫn rất mơ hồ vì nó mang tính hàn lâm và khô khan. Nhưng chính những công trình nghiên cứu, sản phẩm khoa học công nghệ của phòng thí nghiệm và Trường Đại học Công nghệ có tính ứng dụng, thực tiễn cao đã truyền ngọn lửa khoa học cho các thành viên trong nhóm để tiếp tục đi theo con đường này.