ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐÌNH TƯỜNG

NGHIÊN CỨU MÔ HÌNH PHÂN LỚP CÂU HỎI VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐÌNH TƯỜNG

NGHIÊN CỨU MÔ HÌNH PHÂN LỚP CÂU HỎI VÀ ỨNG DỤNG

Ngành: Công Nghệ Thông Tin

Chuyên ngành: Hệ thống Thông Tin

Mã số chuyên ngành: 60480104

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Nguyễn Trí Thành

Hà Nội – 2016

LỜI CAM ĐOAN

Tác giả xin cam đoan kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân Tác giả và được sự hướng dẫn khoa học của PGS. TS. Nguyễn Trí Thành, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều trình bày là của cá nhân hoặc được tổng hợp của nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tác giả xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày tháng năm 2016

HỌC VIÊN

Nguyễn Đình Tường

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành và sâu sắc nhất tới PGS.TS Nguyễn Trí Thành, người thầy đã trực tiếp hướng dẫn tận tình và đóng góp những ý kiến quý báu cho em trong suốt quá trình thực hiện luận văn tốt nghiệp này.

Em xin gửi lời cảm ơn đến các thầy cô trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, đã tận tâm truyền đạt những kiến thức quý báu làm nền tảng cho em trong công việc và cuộc sống. Qua đây, em cũng xin gửi lời cảm ơn đến Trung tâm đào tạo E-Learning, Viện Đại học Mở Hà Nội đã tạo điều kiện giúp đỡ em trong quá trình làm thực nghiệm.

Cuối cùng, em xin được cảm ơn cha mẹ, người thân, bạn bè và đồng nghiệp của em tại bộ phận Công Nghệ trong Trung tâm đào tạo E-Learning, những người đã luôn bên em, khuyến khích và động viên em trong cuộc sống và học tập.

HỌC VIÊN

Nguyễn Đình Tường

MỤC LỤC

DANH MỤC KÝ HIỆU VÀ TỪ VIẾT TẮT

Chữ viết tắt	**Ý nghĩa**
1. Tiếng Việt	
2. Tiếng Anh	
TREC	Text Retrieval Conference
SVM	Support Vector Machine
MEM	Maximum Entropy Model
SNoW	Spare Network of Winnows

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Sự phát triển cơ sở hạ tầng công nghệ đặc biệt là công nghệ mạng đã thúc đẩy nhu cầu trao đổi, chia sẻ dữ liệu của mọi người, đưa Internet trở thành một kho dữ liệu khổng lồ. Những tri thức trong kho dữ liệu này lại cung cấp cơ sở để giải đáp các vấn đề, thắc mắc hàng ngày của con người. Với mục đích phục vụ nhiều hơn nhu cầu của con người, các hệ thống hỏi đáp tự động đã ra đời. Kiến trúc bên trong của một hệ thống hỏi đáp rất phức tạp. Những câu hỏi của người dùng sẽ được hệ thống phân tích, xử lý. Dựa vào thông tin đã được phân tích, hệ thống hỏi đáp tìm kiếm những câu trả lời tiềm năng. Cuối cùng, hệ thống trả về cho người dùng một kết quả ngắn gọn, súc tích và chính xác nhất. Để có thể tìm kiếm những câu trả lời tiềm năng thì ở giai đoạn xử lý câu hỏi, hệ thống phải phân lớp chính xác được câu hỏi. Khi đó, không gian tìm kiếm câu trả lời được giới hạn và rõ ràng hơn. Vì vậy, trong giai đoạn xử lý câu hỏi của hệ thống hỏi đáp, phân lớp câu hỏi là một nhiệm vụ rất quan trọng.

Tuy nhiên, việc nghiên cứu các giải pháp cho nhiệm vụ phân lớp gặp không ít khó khăn. Các mô hình phân lớp, giải thuật phân lớp đang áp dụng đều có những ưu điểm và nhược điểm nhất định. Bên cạnh đó, một vấn đề khác là việc xử lý ngôn ngữ tự nhiên. Trên thế giới, một số ngôn ngữ có cấu tạo từ loại, ngữ nghĩa rất đa dạng và phức tạp. Hệ thống phải xác định được các đặc trưng đó và đưa ra chiến lược xử lý phù hợp. Do đó, vấn đề này cần nhận được quan tâm và nghiên cứu nhiều hơn.

Trong luận văn nghiên cứu này, tác giả trình bày trong 4 chương với nội dung được tóm tắt như sau:

Chương 1. **Giới thiệu phân lớp câu hỏi** trình bày định nghĩa, mục tiêu trong nhiệm vụ phân lớp câu hỏi và đôi nét về khái niệm taxonomy câu hỏi. Mục tiêu của luận văn cũng được nêu trong phần cuối của chương.

Chương 2. **Các phương pháp tiếp cận bài toán phân lớp câu hỏi** nghiên cứu về các mô hình phân lớp câu hỏi đã và đang được sử dụng phổ biến như mô hình phân lớp phẳng, mô hình phân lớp cấp bậc. Chương này cũng trình bày một số giải thuật phân lớp trong học máy có giám sát và bán giám sát.

Chương 3. **Đề xuất cải tiến mô hình phân lớp** nêu lên các vấn đề trong thực tế ảnh hưởng đến kết quả phân lớp. Dựa vào một số nghiên cứu, tác giả đề xuất mô hình phân lớp cải tiến và trình bày các bước xử lý dữ liệu trong mô hình.

Chương 4. **Ứng dụng vào hệ thống hỏi đáp thắc mắc tại Trung tâm đào tạo E-Learning** giới thiệu về Trung tâm đào tạo E-Learning và thực trạng của hệ thống hỏi đáp thắc mắc. Sau đó, trình bày các kết quả thực nghiệm khi áp dụng mô hình đề xuất

với dữ liệu câu hỏi tại trung tâm. Cuối cùng là những nhận xét, đánh giá về mô hình đề xuất.

Phần **Kết luận và hướng phát triển tương lai** trình bày những kết quả đã đạt được và những hạn chế trong luận văn. Các vấn đề còn hạn chế sẽ được giải quyết trong hướng phát triển tương lai của luận văn.

Chương 1: Giới thiệu phân lớp câu hỏi

1.1 Giới thiệu

Phân lớp câu hỏi [6, tr.9-10] là nhiệm vụ gán 1 giá trị đúng hoặc sai tới mỗi cặp $(q_j, c_i) \in Q \times C$, trong đó Q là miền các câu hỏi và $C = \{C_1, C_2, \dots, C_{|C|}\}$ là tập các lớp đã được định nghĩa trước. Các lớp này mang nghĩa khác nhau ràng buộc với câu trả lời tiềm năng. Phân lớp câu hỏi là một bước xử lý quan trọng trong các hệ thống hỏi đáp. Mục đích của hệ thống hỏi đáp là đưa ra được một câu trả lời ngắn gọn, súc tích hơn là những tài liệu liên quan chứa câu trả lời. Trong ngôn ngữ tự nhiên, một câu hỏi có thể liên quan và ảnh hưởng bởi nhiều lĩnh vực khác nhau nên lượng câu trả lời liên quan cũng rất lớn. Việc phân lớp câu hỏi sẽ cung cấp các thông tin ràng buộc về loại câu trả lời. Nhờ đó, hệ thống đưa ra một hoặc nhiều chiến lược làm giảm không gian tìm kiếm các câu trả lời tiềm năng trong kho ngữ liệu khổng lồ.

Ví dụ, với câu hỏi "Điều kiện công nhận tốt nghiệp cho sinh viên Elearning?", những từ khóa như "**điều kiện**", "**tốt nghiệp**", "**sinh viên Elearning**" là cơ sở để tìm ra được câu trả lời phù hợp. Mục đích của câu hỏi này là tìm ra được điều kiện để có thể công nhận tốt nghiệp cho sinh viên chứ không phải là điều kiện vượt qua môn học hay điều kiện để đạt học bổng học tập dành cho sinh viên của Elearning. Một câu hỏi khác như "Xem thông tin tuyển sinh đào tạo Elearning ở đâu?" thì câu trả lời tiềm năng phải là một nơi công bố chính thức thông tin tuyển sinh. Địa điểm đó có thể là vị trí của cơ quan, tổ chức làm việc hoặc qua một vị trí ảo trên mạng Internet - một website điện tử. Câu trả lời cho câu hỏi này có thể là địa chỉ Tầng 3 Nhà A - B101 - Đường Nguyễn Hiền - Quận Hai Bà Trưng - Hà Nội hoặc địa chỉ website http://tuyensinh.ehou.edu.vn.

Trong phân lớp câu hỏi, việc xác định ngữ nghĩa rõ ràng của câu hỏi mang lại một lợi ích to lớn. Tuy nhiên, các câu hỏi không phải lúc nào cũng đơn giản mà thường chúng rất phức tạp và có nhiều ngữ nghĩa mập mờ, không xác định. Nếu như biết được ngữ nghĩa thì câu hỏi có thể được phân loại vào các lớp chi tiết hơn. Nhưng công việc xác định chính xác ngữ nghĩa cho câu hỏi là một thách thức không hề nhỏ.

Chẳng hạn đối với câu hỏi "Chương trình đào tạo Elearning hoạt động theo quy chế nào?", một câu trả lời liên quan tới lớp "**quy chế đào tạo**" sẽ có ích hơn là chỉ biết câu hỏi thuộc lớp quy chế. Quy chế bao gồm rất nhiều loại như quy chế đào tạo, quy chế tuyển sinh, quy chế học tập và xử lý vi phạm, quy chế thi và những quy chế khác. Nếu hệ thống đã xác định được mục đích câu hỏi là về quy chế đào tạo thì không gian tìm kiếm câu trả lời sẽ giảm đi rất nhiều. Trong quá trình phân lớp câu hỏi, một hạn chế làm ảnh hưởng tới hiệu suất phân lớp là hệ thống phải dự đoán quá nhiều lớp cho một câu hỏi trong một thời điểm. Nếu hệ thống hạn chế được số lượng lớp phải dự đoán thì hiệu suất phân lớp sẽ tăng lên. Vì vậy, ngoài việc xác định ngữ nghĩa câu hỏi, thì việc lựa

chọn sử dụng mô hình phân lớp nào cũng rất quan trọng. Đó là bước đầu thực sự cần trong nhiệm vụ phân lớp.

1.2 Tìm hiểu các loại câu hỏi

Xác định loại câu hỏi mang một ý nghĩa to lớn trong việc phân tích các câu hỏi bởi mỗi loại câu hỏi sẽ có những đặc trưng và cách tiếp cận khác nhau. Có nhiều loại câu hỏi như: câu hỏi định nghĩa, liệt kê, mô tả, sự kiện, tổng hợp, đánh giá, … Ví dụ câu hỏi định nghĩa như "**Học trực tuyến là gì**", câu hỏi liệt kê "**Những thuận lợi trong học trực tuyến là gì**". Trong một câu hỏi, người ta có thể sử dụng những từ ngữ phức tạp và đa dạng để diễn đạt cùng nội dung làm cho việc xác định lớp câu hỏi phù hợp trở lên khó khăn. Các loại câu hỏi này cần được đưa về thành hai dạng chung nhất là loại câu hỏi đơn giản và loại câu hỏi phức tạp. Mỗi loại câu hỏi cần có chiến lược phân lớp phù hợp. Hệ thống có thể không cần xử lý thêm đối với loại câu hỏi đơn giản nhưng lại yêu cầu xử lý tinh vi hơn đối với loại câu hỏi phức tạp. Việc đơn giản hóa các câu hỏi phức tạp là một chiến lược đáng được xem xét. Câu hỏi phức tạp sẽ được định nghĩa bằng tập các câu hỏi đơn giản. Câu trả lời nào trong các câu trả lời tiềm năng đáp ứng cho tập các câu hỏi đơn giản đó có thể là câu trả lời phù hợp. Vì vậy, việc xác định loại câu hỏi có ý nghĩa trong việc lựa chọn các đặc trưng phân lớp.

1.3 Taxonomy câu hỏi

Trong nhiều lĩnh vực khoa học và thực tiễn, khái niệm taxonomy mang nhiều ý nghĩa khác nhau. Nhưng về bản chất, taxonomy là sự định danh, phân loại một số tính chất, đặc điểm. Taxonomy được xem là một hình thức của hệ phân cấp và có chứa các phần tử gọi là các nút. Mối quan hệ chủ yếu giữa các nút là quan hệ cha – con và giữa các nút con đồng cấp với nhau thì không tồn tại mối quan hệ nào.

Taxonomy là một cây phân cấp các khái niệm, trong đó các nút (trừ nút gốc) biểu diễn một khái niệm và mỗi nút con có quan hệ is-a-kind-of (là một kiểu/loại của nút cha) với nút cha. Ví dụ, nút khái niệm "*number*" có các nút con chứa các khái niệm "*code*", "*count*", "*date*", "*distance*", "*money*", "*order*".

Một taxonomy được mô tả theo cấu trúc hình cây, trên đỉnh của cấu trúc là nút gốc và dưới nó là các nút con, tập nút con của các nút cha không giao nhau. Khi duyệt cây từ nút cha đến các nút con, thông tin tại các nút con chi tiết và rõ ràng hơn nút cha. Khi xác định được nút cha, các nút con cũng sẽ được xác định. Điều này mang lại hiệu quả trong tìm kiếm, truy vấn dữ liệu vì dựa vào nút cha, việc xác định miền thông tin cần tìm rõ ràng hơn và được giới hạn.

Cùng nghiên cứu về vấn đề này, hai nhà nghiên cứu Li và Roth [14, tr.5] đã đưa ra một taxonomy câu hỏi phân cấp trong hội nghị TREC. Taxonomy câu hỏi cùng bộ dữ

liệu đã được gán nhãn gồm hơn 5000 câu hỏi có độ phủ lớn các vấn đề trong thực tế. Chi tiết được thể hiện ở Bảng 1.1

Bảng 1.1. Taxonomy câu hỏi của Li và Roth

Nhãn lớp	**Ý Nghĩa**	**Số lượng câu hỏi**	**Số lượng kiểm tra**
ABBREV	**Sự tóm tắt**		
abbreviation	Tóm tắt	16	1
expansion	ý nghĩa viết tắt	70	8
ENTITY	**Thực thể**		
animal	Động vật	112	16
body	Cơ thể	16	2
color	Màu sắc	40	10
creative	Sự sáng tạo	207	0
currency	Tiền tệ	4	6
disease/ medical	Bệnh tật và y học	103	2
event	Sự kiện	56	2
food	Thực phẩm	103	4
instrument	Dụng cụ chơi nhạc	10	1
language	Ngôn ngữ	16	2
letter	Ký tự	9	0
other	Thực thể khác	217	12
plant	Thực vật	13	5
product	Sản phẩm	42	4
religion	Tín ngưỡng	4	0
sport	Thể thao	62	1
substance	Nguyên tố	41	15
symbol	Ký hiệu	11	0
technique	Kỹ thuật	38	1
term	Thuật ngữ	93	7

vehicle	Phương tiện giao thông	27	4
word	Từ ngữ	26	0
DESCRIPTION	**Mô tả**		
definition	Định nghĩa	421	123
description	Mô tả	274	7
manner	Bộ dạng cử chỉ	276	2
reason	Lý do	191	6
HUMAN	**Con người**		
group	Nhóm	189	6
individual	Cá nhân, cá thể	962	55
title	Danh nghĩa	25	1
description	Mô tả	47	3
LOCATION	**Địa điểm**		
city	Thành phố	129	18
country	Đất nước	155	3
mountain	Ngọn núi	21	3
other	Địa điểm khác	464	50
state	Bang, tỉnh thành	66	7
NUMERIC	**Số học**		
code	Mã	9	0
count	Số lượng	363	9
date	Ngày tháng	218	47
distance	Khoảng cách	34	16
money	Giá cả	71	3
order	Thứ hạng	6	0
other	Khác	52	12
period	Giai đoạn	75	8
percent	Phần trăm	27	3

speed	Tốc độ	9	6
temperature	Nhiệt độ	8	5
size	Kích thước	13	0
weight	Cân nặng	11	4

1.4 Mục tiêu của luận văn

Trong nhiều nghiên cứu về phân lớp câu hỏi, các nhà khoa học đã đề xuất áp dụng nhiều mô hình phân lớp trên các miền dữ liệu khác nhau. Ban đầu, phân lớp câu hỏi chỉ tập trung vào phân lớp phẳng nhưng sau đó có nhiều vấn đề nảy sinh cần phải được giải quyết nên phân lớp phẳng trở lên không còn phù hợp mà thay vào đó là các mô hình phân lớp khác như phân lớp cục bộ (Local Classifier), phân lớp toàn cục (Global Classifier hay Big-Bang), phân lớp phân cấp (Hierarchical Classifier), …

Mỗi hướng tiếp cận đều có những tính chất và đặc điểm khác nhau. Hướng tiếp cận phân lớp phẳng đơn giản và dễ hiểu nhưng có một nhược điểm lớn là việc phải xây dựng một bộ phân lớp để phân biệt với số lượng lớp là rất lớn. Phân lớp phẳng sử dụng toàn bộ các lớp trong một lần dự đoán nhãn lớp câu hỏi. Hướng tiếp cận phân lớp cấp bậc là một phương pháp tốt giúp tổ chức lượng lớn thông tin. Trong những năm gần đây, hướng tiếp cận này đã nhận được ngày càng nhiều sự quan tâm đặc biệt của các nhà nghiên cứu trên thế giới.

Trong hướng tiếp cận phân lớp cục bộ, thông tin được coi là trung tâm và được sử dụng để xây dựng các bộ phân lớp cục bộ theo những phương cách khác nhau. Để phân loại các phân lớp cục bộ, các nhà nghiên cứu dựa vào các cách sử dụng thông tin cục bộ là phân lớp cục bộ cho từng nút, phân lớp cục bộ cho từng nút cha, phân lớp cục bộ cho từng cấp. Phân lớp cục bộ sẽ sử dụng vòng lặp đệ quy để dự đoán các cấp độ tiếp theo. Việc phân lớp chỉ dừng lại cho đến khi một dự đoán đặc biệt được thực hiện. Nhưng vấn đề là tại các nút lá, việc dự đoán có thể không xảy ra. Điều này gây ra lỗi và nó sẽ ảnh hưởng xuống cấp nhỏ hơn tiếp theo nếu không xử lý ngăn chặn điều này. Với hướng tiếp cận phân lớp toàn cục, có hai vấn đề là cần xem toàn bộ các lớp phân cấp cùng một thời điểm và thiếu các module huấn luyện cục bộ.

Sau một số tìm hiểu, nghiên cứu về các miền câu hỏi cụ thể, tác giả thấy rằng kết quả phân lớp của một số lớp đạt tỉ lệ chính xác cao trong khi một số khác lại thấp hơn. Giả sử rằng, nếu ta tính toán, dự đoán được các lớp có độ chính xác cao và loại bỏ dữ liệu đã gán nhãn đó, ta chỉ tiến hành phân lớp với các lớp có độ chính xác thấp hơn. Phân lớp sử dụng bộ phân lớp chỉ chứa các lớp có độ chính xác kém làm tăng độ chính xác chung trong nhiệm vụ phân lớp.

Trong luận văn "**Nghiên cứu mô hình phân lớp câu hỏi và ứng dụng**" tác giả sẽ trình bày một số nghiên cứu về mô hình phân lớp câu hỏi, đề xuất mô hình phân lớp cải tiến và ứng dụng với dữ liệu cụ thể trong thực nghiệm để kiểm chứng mô hình.

Chương 2: Các phương pháp tiếp cận bài toán phân lớp câu hỏi

2.1 Mô hình phân lớp câu hỏi

2.1.1 Mô hình phân lớp phẳng

Mô hình phân lớp phẳng được biết đến như một hướng tiếp cận đơn giản trong các mô hình phân lớp. Với việc sử dụng bộ phân lớp phẳng, các mối quan hệ bên trong của nhãn lớp bị bỏ qua và khi phân lớp, toàn bộ nhãn lớp được sử dụng trong một thời điểm để dự đoán cho một dữ liệu câu hỏi.

Ví dụ, Dragomir Radev [5, tr.575] đã đưa ra một bộ phân lớp phẳng gồm 17 lớp để áp dụng trong quá trình phân lớp. Với mỗi câu hỏi, Dragomir Radev phải sử dụng tới 17 lớp trong các lần dự đoán nhãn lớp cho câu hỏi. Bộ phân lớp được trình bày trong Bảng 2.1

Bảng 2.1. Bộ phân lớp câu hỏi của Radev

STT	**Nhãn lớp**	**Ý nghĩa**
1	PERSON	Người
2	PLACE	Địa điểm
3	DATE	Ngày tháng
4	NUMBER	Số học
5	DEFINITION	Định nghĩa
6	ORGANIZATION	Tổ chức
7	DESCRIPTION	Mô tả
8	ABBREVIATION	Viết tắt
9	KNOWNFOR	Nổi tiếng
10	RATE	Tỉ lệ
11	LENGTH	Chiều dài
12	MONEY	Tiền tệ
13	REASON	Lý do
14	PURPOSE	Mục đích
15	DURATION	Thời gian
16	NOMINAL	Quy định

17	OTHER	Khác

Việc sử dụng mô hình phân lớp phẳng gây ra khó khăn trong quá trình lựa chọn đặc trưng cho nhãn lớp. Tập thuộc tính mới được lựa chọn phải chứa đặc điểm xác định, phân biệt nhãn lớp đó với các nhãn lớp đã xây dựng. Đôi khi, một số trường hợp đặc biệt có thể xảy ra và ảnh hưởng nhiều đến quá trình phân lớp. Đó là khi các thuộc tính phân biệt rất tốt nhưng lại không có ích, không mang lại hiệu quả trong việc phân lớp.

2.1.2 Mô hình phân lớp phân cấp

Mô hình phân lớp phân cấp có nhiều ưu điểm về độ chính xác, cách tổ chức thông tin và mô hình này được xem là sự bổ sung và cải tiến của một số mô hình phân lớp khác. Ý tưởng cơ bản của mô hình là giảm số lượng các nhãn lớp phải dự đoán cho mỗi câu hỏi trong từng cấp. Đầu ra của phân lớp cấp một là một tập con nhãn lớp được sử dụng làm bộ phân lớp tiếp theo. Phân lớp cấp một dự đoán câu hỏi vào các nhãn lớp tổng thể, các nhãn lớp này có độ bao quát lớn nên sẽ gặp ít rủi ro hơn nếu có nhiều đặc trưng mới.

Li và Roth[14, tr.8] đã đề xuất áp dụng mô hình phân lớp hai cấp cùng với bộ phân lớp. Mô hình kết hợp hai phân lớp liên tiếp, theo thứ tự. Đầu tiên, câu hỏi sẽ được đưa vào phân lớp cấp một - phân lớp thô, phân lớp này gồm 6 lớp là **ABBREVIATION**, **ENTITY**, **DESCRIPTION**, **HUMAN**, **LOCATION**, **NUMBERIC VALUE**. Mỗi lớp thô này lại chứa một số lớp con chi tiết hơn bên trong. Các lớp mịn này chỉ thuộc vào duy nhất một lớp thô. Tổng số lượng lớp mịn trong mô hình phân lớp phân cấp là 50 lớp đã được trình bày trong Bảng 1.1. Phân lớp cấp hai - phân lớp mịn phải dựa vào phân lớp cấp một để khi câu hỏi đi qua phân lớp cấp hai, vấn đề không rõ ràng trong câu hỏi sẽ được giải quyết.

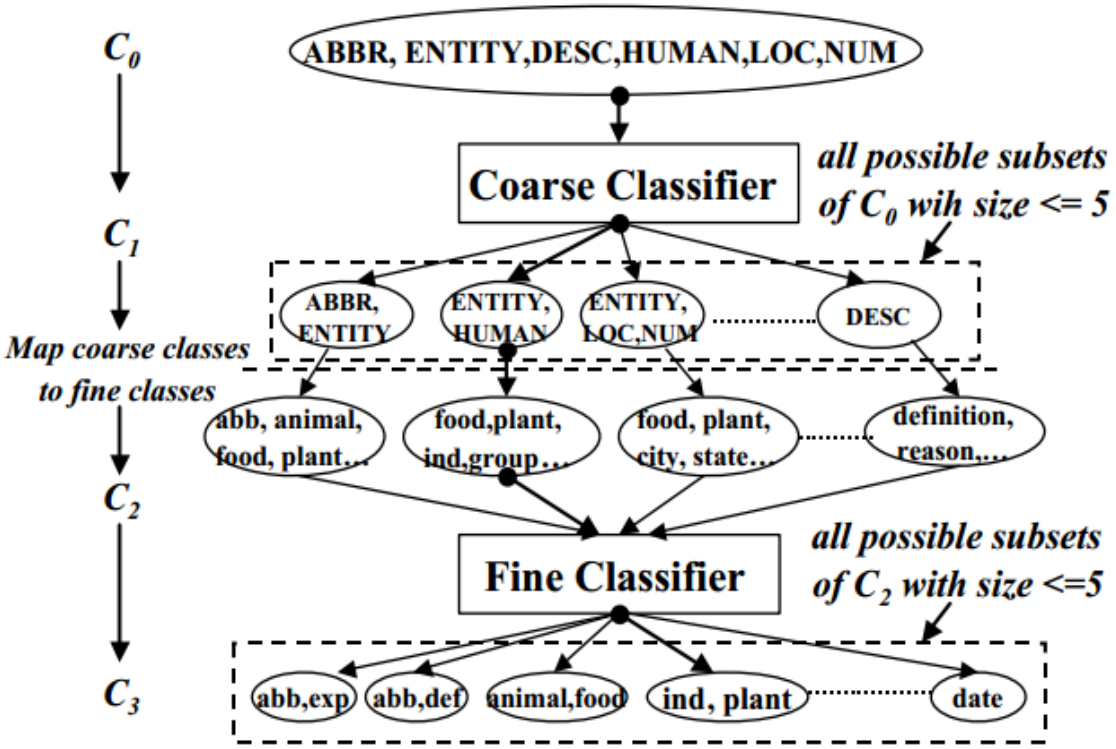
Chi tiết mô hình xử lý phân lớp được trình bày như sau: Ban đầu, định nghĩa một tập của tất cả các lớp thô $C_0 = \{c_1, c_2, c_3, \dots, c_n\}$ cho câu hỏi q bất kỳ. Phân lớp thô xác định một tập nhãn con C_1.

$$C_1 = Coarse\_Classifier(C_0, q), C_1 \subset C_0 \text{ và } |C_1| \leq 5$$

Sau đó, mỗi lớp thô trong C_1 sẽ mở rộng thành tập hợp các lớp mịn đã định nghĩa trước trong phân lớp cấp bậc. Giả sử, lớp thô c_i được ánh xạ thành tập $c'_i = \{fi_1, fi_2, fi_3, \dots, fi_m\}$ lớp mịn, tập C_2 là tập hợp của tất cả các lớp mịn được xác định bởi lớp thô trong C_1. Phân lớp mịn xác định một tập hợp nhãn con C_3 dựa vào C_2.

$$C_3 = Fine\_Classifier(C_2, q), C_3 \subset C_2 \text{ và } |C_3| \leq 5$$

Kết thúc quá trình, ta có tập C_1 và C_3 là tập nhãn lớp được dự đoán qua hai lần phân lớp. Các nhãn lớp này được sử dụng để đánh giá hiệu quả.



Hình 2.1. Mô hình phân lớp của Li và Roth

2.2 Giải thuật phân lớp câu hỏi

Về cơ bản, phân lớp câu hỏi thường sử dụng hai hướng tiếp cận chính là hướng tiếp cận dựa trên luật và hướng tiếp cận dựa trên học máy. Bên cạnh đó, sự kết hợp của hướng tiếp cận dựa trên luật và học máy cũng đưa đến những hướng tiếp cận mới. Trong hướng tiếp cận dựa trên luật, các luật được xây dựng một cách thủ công và cần nhiều thời gian để tạo ra các luật áp dụng cho dữ liệu thực tế. Việc phân lớp chỉ đạt hiệu quả tốt trên một số miền đặc biệt với bộ dữ liệu chuẩn, còn đối với các bộ dữ liệu mới thì kết quả đạt được lại không mấy khả quan. Vì vậy, mô hình hóa luật là việc làm cần thiết khi áp dụng hướng tiếp cận dựa trên luật nhưng sẽ vô cùng khó khăn.

Hướng tiếp cận dựa trên học máy thì gần như đối lập với hướng tiếp cận dựa trên luật trong việc lựa chọn các đặc trưng để xây dựng bộ phân lớp. Các đặc trưng trong câu hỏi được trích xuất để huấn luyện bộ phân lớp. Công việc không phải thực hiện một cách thủ công hoàn toàn nên sẽ không mất nhiều thời gian. Ví dụ, cùng một nội dung câu hỏi nhưng được biểu diễn thành các câu hỏi có cú pháp khác nhau. Các câu hỏi được gán cùng một nhãn lớp nhưng các đặc trưng trích xuất từ câu hỏi thì có một số khác nhau. Do vậy, điều này sẽ gây khó khăn đối với hướng tiếp cận dựa trên luật nhưng lại

rất đơn giản đối với hướng tiếp cận học máy. Các thực nghiệm áp dụng hướng tiếp cận dựa trên học máy thành công đã được công bố trong nhiều bài báo. Trong luận văn này, tác giả cũng xin trình bày chủ yếu về hướng tiếp cận dựa trên học máy.

2.2.1 Giải thuật học máy có giám sát

Trong học máy có giám sát, chương trình học sẽ được cung cấp hai tập dữ liệu, một tập dữ liệu huấn luyện và một tập dữ liệu kiểm tra. Ý tưởng của phương pháp này là chương trình học sẽ "học" từ những dữ liệu đã được gán nhãn lớp trong tập dữ liệu huấn luyện để phát hiện ra một quy luật, một chương trình hay một thủ tục phân loại những dữ liệu mới.

Ban đầu, tập dữ liệu huấn luyện sẽ chứa n cặp $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$, trong đó mỗi x_i là một tập các đặc trưng cho từng dữ liệu huấn luyện riêng, y_i là nhãn lớp đã gắn cho dữ liệu x_i. Ví dụ, $x_i$có một số đặc trưng của sinh viên đang tham gia quá trình học tập như điểm tổng kết, tín chỉ tích lũy, bị kỷ luật, được khen thường, hạnh kiểm. y_i là nhãn lớp tương ứng cho sinh viên được tốt nghiệp hay không được tốt nghiệp. Tập dữ liệu kiểm tra là một tập khác với m đặc trưng mà chưa được gán nhãn. Mục đích là gán nhãn cho tập dữ liệu kiểm tra như được tốt nghiệp hoặc không được tốt nghiệp bằng cách suy diễn dựa vào những đặc trưng đã được học trong tập dữ liệu huấn luyện.

Để làm được như vậy, ta cần sử dụng các giải thuật kết hợp với mô hình phân lớp cụ thể. Mỗi bài toán có thể lựa chọn áp dụng các giải thuật khác nhau. Hiện nay, một số giải thuật phân lớp phổ biến được sử dụng trong hướng tiếp cận học máy có giám sát có thể kể tới như Support Vector Machines (SVM), Maximum Entropy Model (MEM), Spare Network of Winnows (SNoW), ... Trong chương này, tác giả xin trình bày về giải thuật SVM sẽ sử dụng trong thực nghiệm.

Giải thuật SVM

Support Vector Machines (SVM) được Cortes và Vapnik giới thiệu lần đầu vào năm 1995, là giải thuật mạnh mẽ trong việc giải quyết các bài toán có số chiều lớn. SVM là phương pháp phân lớp nhanh, có hiệu suất cao và giải quyết vấn đề overfitting rất tốt. Ý tưởng của SVM là tìm ra một siêu phẳng với biên độ cao nhất, tức là siêu phẳng này phải tối ưu, có khoảng cách tới điểm gần nhất là lớn nhất.

Cho tập dữ liệu D=$\{(x_i, y_i), i = 1, \dots, n\}$ với $x_i = (x_{i1}, \dots, x_{ih})$ là mẫu có h chiều và $y_i \in \{-1, +1\}$ là nhãn tương ứng xác định cho x_i, nó nhận giá trị là dữ liệu dương hay âm. Bộ phân lớp tuyến tính được xác định bằng siêu phẳng:

$$\{x: f(x) = w^T x + w_0\}$$

Một dữ liệu thuộc lớp dương $y_i = +1$ nếu $f(x) \geq +1$ và nếu $f(x) \leq -1$ thì dữ liệu thuộc lớp âm $y_i = -1$. Đây là những điều kiện cần và đủ để tìm ra siêu phẳng tối ưu. Để xác định w^T và w_0 ta dùng bất đẳng thức sau đây:

$$y_i(w^T x_i + w_0) \geq 1, \forall i = 1, 2, \dots, n$$

Hàm tuyến tính tối ưu đạt được bằng cách tối thiểu hóa phương trình sau:

$$\min \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + w_0) - 1) \; với \; w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Và $\alpha_i \geq 0 \; với \; i = 1, 2, \dots, n$ là hệ số Lagrange. Đây là chiến lược để tìm ra giá trị cực đại và cực tiểu cục bộ theo một số ràng buộc. Trong việc phân lớp, các không gian đặc trưng nên ánh xạ sang không gian có số chiều lớn vì điều này sẽ làm SVM luôn đạt được hiệu suất cao.

Việc ánh xạ không gian đặc trưng còn có tên gọi khác là hàm nhân. Các hàm nhân được sử dụng trong nghiên cứu phân lớp câu hỏi rất đa dạng. Mỗi hàm nhân có hiệu suất phân lớp khác nhau tùy thuộc vào nghiên cứu. Ba hàm nhân phổ biến là nhân tuyến tính, nhân đa thức và nhân xích ma. Nhân tuyến tính thường được sử dụng trong trường hợp có các vector dữ liệu lớn nhưng thưa thớt như phân loại văn bản. Nhân đa thức thì phổ biến trong phân tích xử lý ảnh và nhân xích ma thì chủ yếu sử dụng cho mạng nơ-ron. Nhân tuyến tính là hàm đơn giản trong các loại hàm nhân nên trong nhiều thực nghiệm, nó là lựa chọn được sử dụng nhiều nhất.

2.2.2 Giải thuật học máy bán giám sát

Trong học máy bán giám sát, ý tưởng đầu tiên về việc tận dụng các đặc trưng có trong dữ liệu chưa gán nhãn chính là việc tự học hay là tự huấn luyện. Chương trình học sẽ sử dụng một ít dữ liệu đã được gán nhãn có độ tin cậy và chính xác cao. Sau đó chương trình sẽ tự động thu nhận các dữ liệu khác trong quá trình kiểm tra dữ liệu chưa gán nhãn. Việc sử dụng dữ liệu chưa gán nhãn kết hợp với dữ liệu đã gán nhãn trong học máy bán giám sát giúp hạn chế những lỗi chủ quan của con người khi xây dựng tập dữ liệu huấn luyện ban đầu.

Giả sử ta có được các tập dữ liệu sau $X_1 = (x_1, x_2, \dots, x_l)$ tương ứng với nó là các nhãn $Y_1 = (y_1, y_2, \dots, y_l)$ và tập $X_u = (x_{l+1}, x_{l+2}, \dots, x_{l+u})$ chưa được gán nhãn. Bước đầu tiên, ta sẽ tiến hành huấn luyện với dữ liệu đã gán nhãn và có được bộ phân lớp như sau $(x'_1, y_1), (x'_2, y_2), \dots, (x'_n, y_n)$ trong đó x'_i là tập các đặc trưng dùng để xác định nhãn lớp y_i. Ta dùng bộ phân lớp này để tiến hành gán nhãn cho từng phần tử x'_u trong tập dữ liệu chưa được gán nhãn X_u. Với mỗi dữ liệu vừa mới gán nhãn, ta lại sử dụng

học máy có giám sát để huấn luyện lại bộ phân lớp và tiếp tục dùng bộ phân lớp mới dự đoán nhãn cho phần tử tiếp theo trong tập dữ liệu X_u. Quá trình này sẽ lặp đi lặp lại đến khi duyệt hết dữ liệu trong tập dữ liệu chưa gán nhãn X_u.

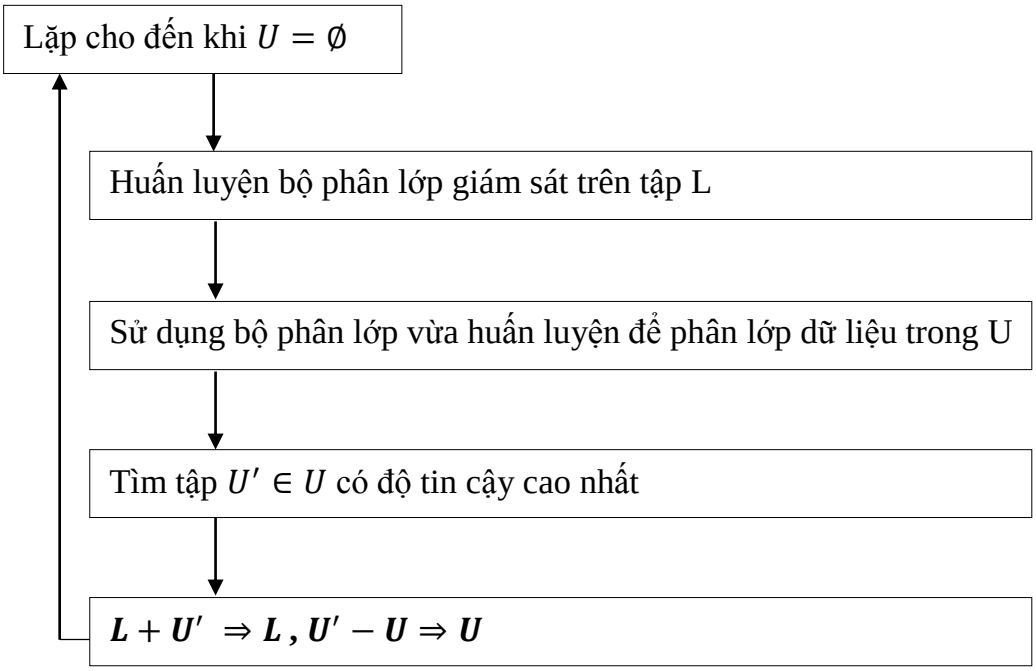
Các giải thuật điển hình trong hướng tiếp cận học máy bán giám sát được kể đến như Self-training, Co-training, Tri-training, …

Giải thuật Self-training

Self-training là giải thuật cơ bản trong học máy bán giám sát và được sử dụng phổ biến. Nó bắt đầu bằng cách huấn luyện chỉ một ít dữ liệu đã được gán nhãn. Trong mỗi bước tiếp theo, nếu một dữ liệu chưa gán nhãn được dự đoán, gán nhãn chính xác và có độ tin cậy cao theo bộ huấn luyện hiện tại thì dữ liệu này được thêm vào tập dữ liệu đã được gán nhãn để xây dựng bộ phân lớp. Sau đó phương pháp học có giám sát sẽ huấn luyện lại bộ phân lớp với dữ liệu đã gán nhãn ban đầu và dữ liệu được bổ sung thêm. Giải thuật được mô tả trong Hình 2.2.

Đầu vào: L là tập dữ liệu đã được gán nhãn, U là tập dữ liệu chưa được gán nhãn.

Đầu ra: Bộ phân lớp đã được huấn luyện.



Hình 2.2. Sơ đồ thực hiện giải thuật Self-training

Giải thuật Co-traning

Co-training là giải thuật được áp dụng rộng rãi trong lĩnh vực xử lý ngôn ngữ tự nhiên. Ý tưởng giải thuật là sử dụng hai tập chứa đặc trưng riêng biệt, không trùng lặp với mỗi dữ liệu đã được gán nhãn là tập ${x^1}_i$ và ${x^2}_i$. Hai tập đặc trưng này được sử dụng để tạo ra hai bộ phân lớp khác nhau. Các bộ phân lớp này sẽ dự đoán dữ liệu chưa gán

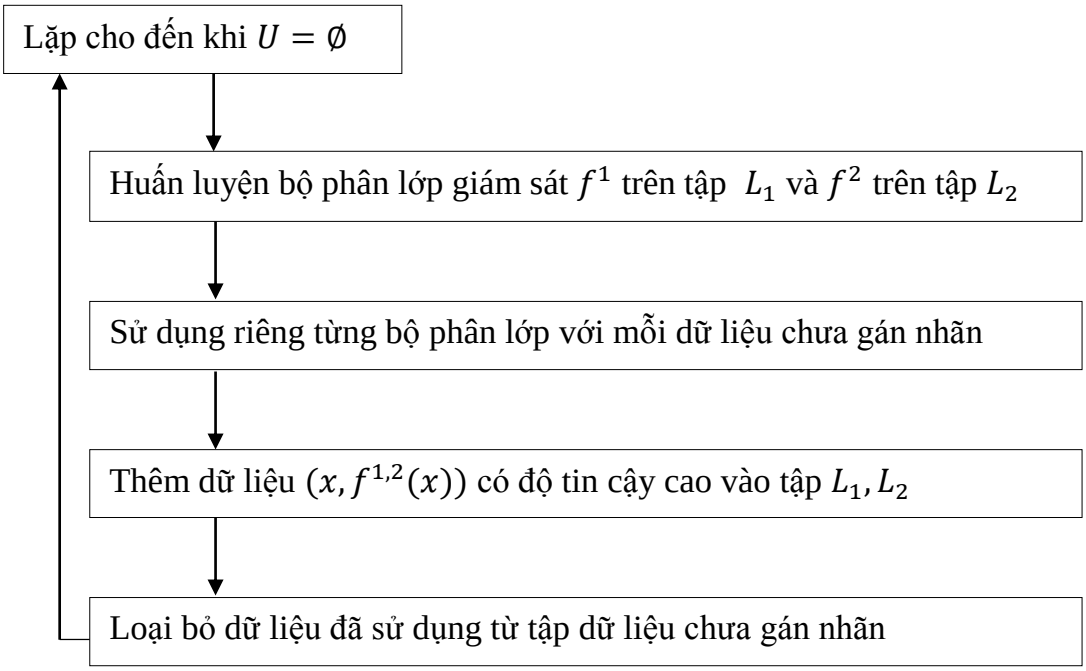
nhãn và tìm ra dữ liệu vừa gán nhãn tin cậy để thêm vào tập dữ liệu đã được gán nhãn. Để nâng cao độ chính xác, việc thêm dữ liệu vào tập dữ liệu đã gán nhãn chỉ được tiến hành nếu như với một dữ liệu chưa gán nhãn mà hai bộ phân lớp cho cùng một kết quả. Bước tiếp theo, các bộ phân lớp sẽ được huấn luyện lại với dữ liệu vừa được thêm vào tập dữ liệu đã gán nhãn. Quá trình lặp này sẽ diễn ra đến khi dữ liệu trong tập chưa gán nhãn đã được sử dụng hết. Giải thuật được như trong Hình 2.3.

Đầu vào: Tập dữ liệu đã gán nhãn $L = \{(x_i, y_i)$ với $i = 1, 2, \dots, n\}$

Tập dữ liệu chưa gán nhãn là $U = \{(x_j)$ với $j = 1, 2, \dots, n\}$.

$$L_1 = L_2 = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$$

Đầu ra: 2 bộ phân lớp đã được huấn luyện



Hình 2.3. Sơ đồ giải thuật Co-training

Vấn đề của giải thuật Co-training là khi sử dụng hai tập đặc trưng riêng biệt sẽ gây ra hạn chế là các đặc trưng của một tập bất kỳ không thể biểu diễn đầy đủ thông tin dữ liệu. Do vậy, để hai bộ phân lớp có thể dự đoán được cùng một kết quả là điều khó khăn.

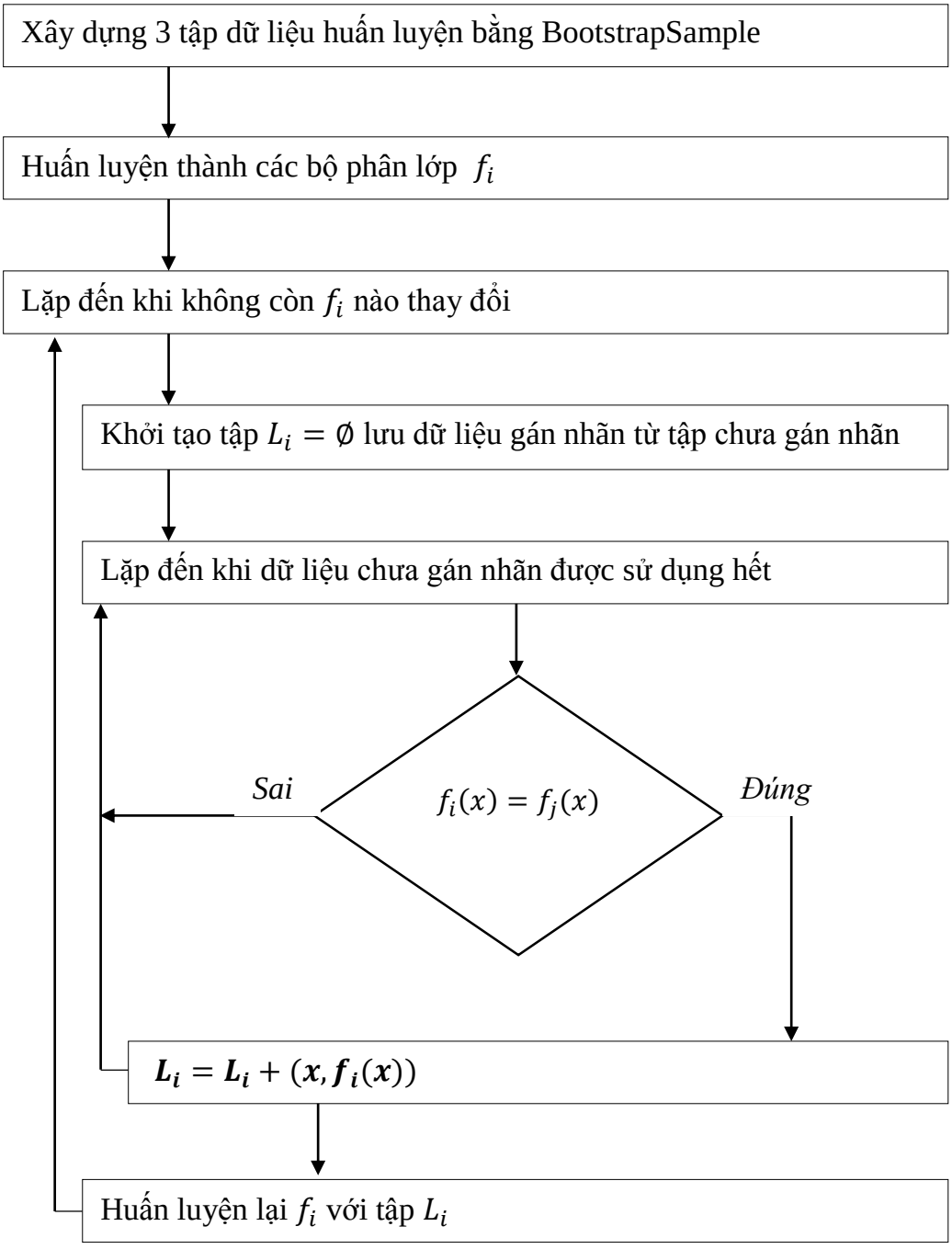
Giải thuật Tri-training

Được coi là một phiên bản của giải thuật Co-training, thay vì chỉ sử dụng hai tập đặc trưng thì Tri-training sử dụng ba tập đặc trưng. Giải thuật này được đề xuất và đặt tên bởi Zhi-Hua Zhou và các đồng nghiệp. Khởi tạo ban đầu, ba tập dữ liệu được tạo ra bằng cách sử dụng BootstrapSample - phương pháp lấy mẫu từ tập dữ liệu đã được gán nhãn và được huấn luyện thành ba bộ phân lớp. Việc lựa chọn sử dụng ba tập để áp dụng

dựa theo nguyên tắc đa số thắng thiểu số, nếu hai trong ba bộ phân lớp bất kỳ độc lập nhau có chung dự đoán với cùng một dữ liệu chưa gán nhãn thì bộ phân lớp còn lại trong ba bộ phân lớp sẽ bị áp đặt việc huấn luyện theo dữ liệu đó. Giải thuật được mô tả trong Hình 2.4.

Đầu vào: L là tập dữ liệu đã được gán nhãn, U là tập dữ liệu chưa được gán nhãn.

Đầu ra: Các bộ phân lớp đã được huấn luyện



Hình 2.4. Sơ đồ giải thuật Tri-training

Giải thuật Tri-training đã thể hiện được tính hiệu quả so với các giải thuật khác. Tuy nhiên, một vấn đề cố hữu vẫn có thể xảy ra là việc không sử dụng hết toàn bộ thông

tin dữ liệu đã gán nhãn trong quá trình tạo thành các tập đặc trưng. Việc sử dụng phương pháp lấy mẫu BootstrapSample có thể bỏ qua các đặc trưng của một số lớp dẫn đến việc làm cho bộ phân lớp có kết quả kém. Vấn đề này có thể gặp trong một số miền dữ liệu đặc biệt.

Để giải quyết vấn đề này, PGS.TS. Nguyễn Trí Thành và đồng nghiệp[11] đã đề xuất cải tiến giải thuật này bằng cách xây dựng ba tập huấn luyện lấy toàn bộ dữ liệu từ tập dữ liệu đã gán nhãn nhưng khác nhau về mô hình trích xuất đặc trưng. Tập đặc trưng thứ nhất được xây dựng dựa theo mô hình bag-of-word. Tập thứ hai xây dựng theo mô hình bag-of-POS và tập đặc trưng thứ ba là kết hợp cả hai mô hình bag-of-word và bag-of-POS. Để huấn luyện, tập đặc trưng thứ nhất và thứ hai sử dụng SVM. Tập đặc trưng thứ ba sử dụng MEM. Kết quả thực nghiệm của các nhà nghiên cứu đã cho thấy rằng độ chính xác của phân lớp tăng khi tận dụng tốt dữ liệu chưa gán nhãn.

Chương 3: Đề xuất cải tiến mô hình phân lớp

3.1 Thực trạng

Trên thực tế, phân lớp câu hỏi muốn đạt kết quả tốt cần phải dựa vào nhiều yếu tố khác nhau. Một trong những yếu tố quan trọng đó là chất lượng của mỗi nhãn lớp. Dù các nhãn lớp tốt hay kém, chúng đều có ảnh hưởng đến chất lượng bộ phân lớp. Chất lượng của nhãn lớp lại liên quan tới đặc trưng trích xuất từ câu hỏi. Những đặc trưng này có thể đóng vai trò tích cực hoặc tiêu cực khi dự đoán nhãn lớp. Vì vậy, để nâng cao chất lượng của từng nhãn lớp, người ta thường xử lý lựa chọn và làm giàu đặc trưng để đưa ra được tập tối ưu các đặc trưng liên quan và có giá trị nhất cho việc dự đoán nhãn lớp đó.

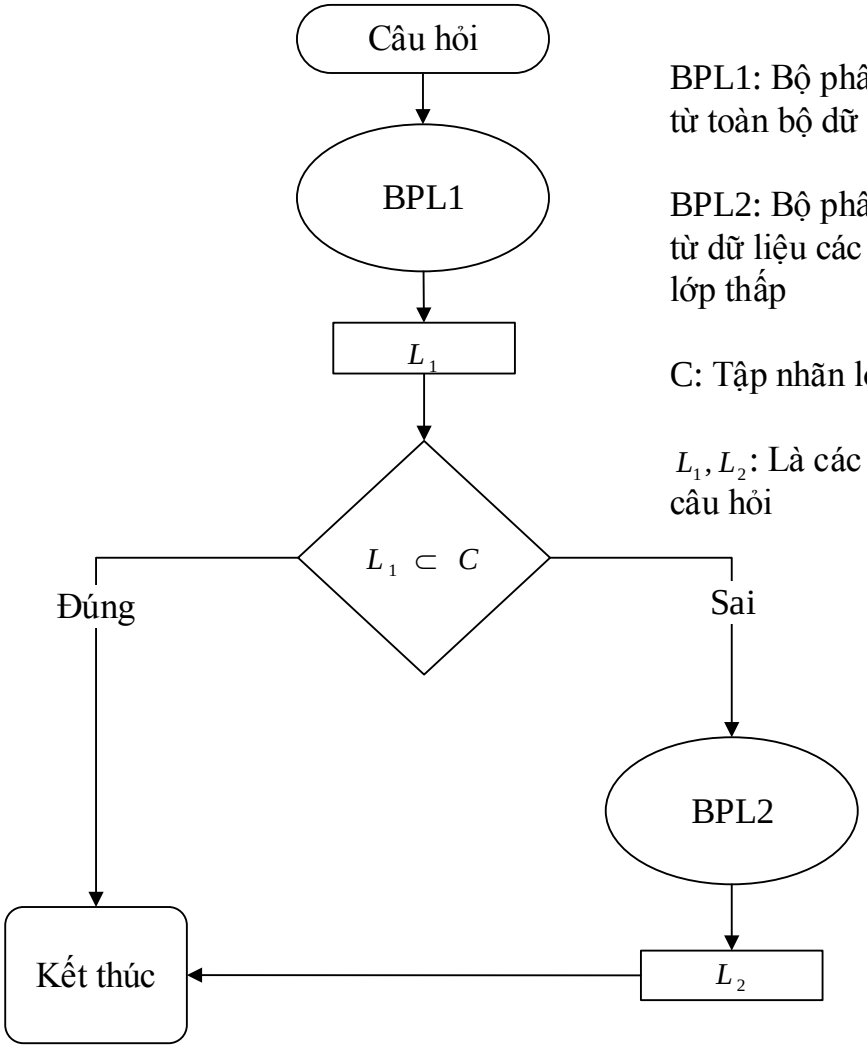
Bên cạnh chất lượng thì số lượng nhãn lớp trong bộ phân lớp cũng cần được quan tâm. Trong một số hệ thống, việc phân lớp chỉ đạt được kết quả tốt khi tiến hành phân lớp với số lượng nhãn lớp nhất định. Việc sử dụng nhiều nhãn lớp có thể gây ảnh hưởng đến hiệu suất làm việc. Vì vậy, giảm số lượng các nhãn lớp phải dự đoán cho một câu hỏi giúp không gian nhãn lớp tiềm năng với câu hỏi đó giảm xuống và tăng hiệu suất làm việc của hệ thống.

Ngày nay, nghiên cứu về phân lớp câu hỏi không chỉ tập trung vào các nhãn lớp và đặc trưng của câu hỏi mà còn xuất hiện nhiều hơn các nghiên cứu về cách tổ chức các mô hình phân lớp, đặc điểm của các miền câu hỏi hay việc sử dụng các giải thuật phân lớp. Nổi bật nhất có lẽ là việc áp dụng mô hình phân lớp theo thứ bậc. Mô hình phân lớp này có nhiều ưu điểm so với mô hình phân lớp phẳng. Có nhiều nghiên cứu mới xuất hiện xung quanh việc áp dụng mô hình này như phát triển gán đa nhãn lớp, cải tiến độ chính xác của mô hình. Ngoài ra còn xuất hiện thêm mô hình được gọi là phân lớp cấu trúc (Structured Classification). Phân lớp cấp bậc được xem như một trường hợp đặc biệt của phân lớp cấu trúc.

3.2 Mô hình đề xuất

Dựa trên việc phân tích đặc điểm của các bộ phân lớp cũng như trên các miền câu hỏi khác nhau, kết quả của quá trình phân lớp xuất hiện các nhãn lớp có độ chính xác khác nhau. Cụ thể, trong một số miền câu hỏi, một số nhãn lớp có độ chính xác cao hơn các nhãn lớp còn lại. Dựa vào những tính toán và đặc trưng của miền, có thể dự đoán được một số nhãn lớp có độ phân lớp chính xác cao. Trong mô hình đề xuất này, tác giả xây dựng hai bộ phân lớp khác nhau. Bộ phân lớp cấp một được huấn luyện từ toàn bộ dữ liệu câu hỏi huấn luyện. Bộ phân lớp cấp hai được xây dựng thông qua việc áp dụng giải thuật tham lam để tìm ra các nhãn lớp có độ phân lớp chính xác cao. Từ đó, tác giả tiến hành lọc bỏ dữ liệu nhãn lớp vừa tìm được ra khỏi tập dữ liệu huấn luyện. Tập dữ liệu huấn luyện còn lại được huấn luyện để tạo thành bộ phân lớp cấp hai cho

mô hình. Câu hỏi đi qua bộ phân lớp cấp một và được gán nhãn lớp. Sau đó, kiểm tra nhãn lớp vừa được gán có phải là nhãn lớp có độ phân lớp chính xác cao hay không. Nếu nhãn lớp câu hỏi không thuộc vào những nhãn lớp đó thì tiến hành phân lớp câu hỏi bằng bộ phân lớp cấp hai.



Hình 3.1. Mô hình phân lớp đề xuất

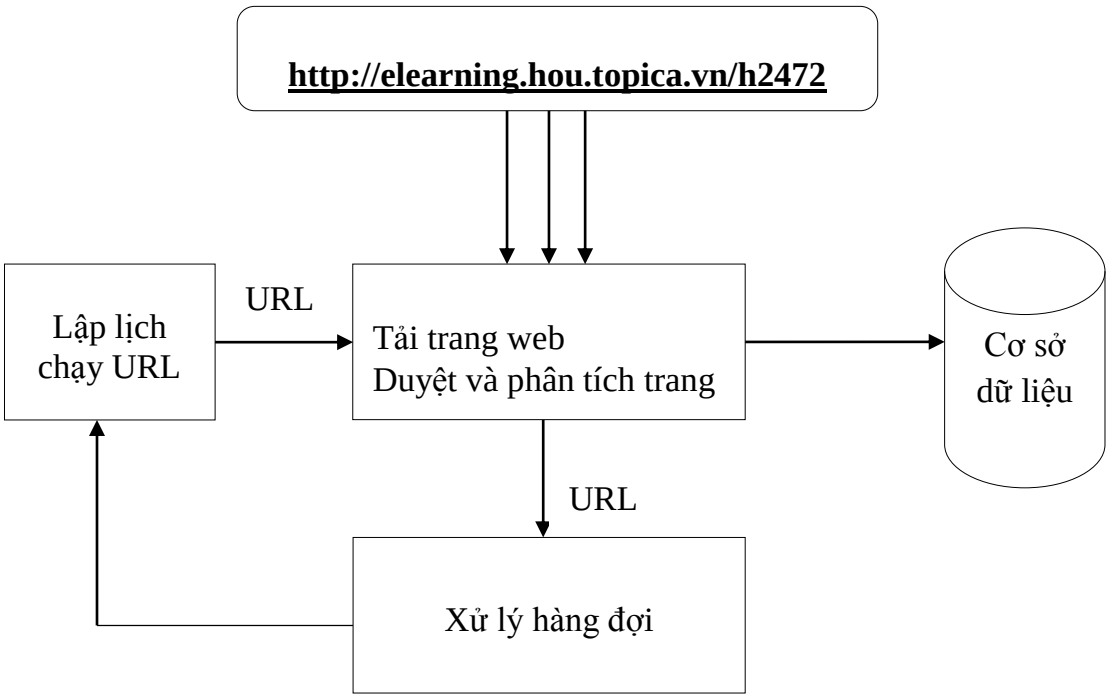
Trong thực nghiệm với mô hình đề xuất, tác giả cũng sử dụng dữ liệu thực nghiệm của Li và Roth để đánh giá mô hình được khách quan hơn.

3.3 Mô hình xử lý dữ liệu

3.3.1 Thu thập dữ liệu

Dữ liệu huấn luyện có thể được lấy từ một hoặc nhiều nguồn khác nhau phụ thuộc vào nhu cầu cũng như mục đích sử dụng của hệ thống. Có một nguồn tài nguyên mà nhiều lĩnh vực thường tìm kiếm đó là Internet. Kho dữ liệu này rất đa dạng và vô cùng quý giá. Đầu tiên, ta cần xác định rõ tài nguyên cần lấy là gì và nơi cung cấp tài nguyên đó, sau đó, sử dụng chương trình để tải về tự động và lưu lại dữ liệu. Chương trình Crawler sẽ đảm nhận công việc này. Crawler là một trong những thành phần quan trọng

trong các bộ máy tìm kiếm như Google, Bing. Chương trình Crawler có thể được viết bằng nhiều ngôn ngữ khác nhau như java, php, nodejs, … Khởi đầu từ một trang nào đó và sử dụng các siêu liên kết trong trang để lấy nội dung câu hỏi và đi tới các trang tiếp theo. Quá trình này sẽ tiếp tục với các trang mới và các trang này lại cung cấp những siêu liên kết mới, nội dung mới. Các siêu liên kết bị trùng sẽ không được lưu vào trong hàng đợi xử lý. Công việc Crawler kết thúc khi đạt được số lượng dữ liệu nhất định hoặc không còn siêu liên kết mới nào trong hàng đợi chưa được duyệt.



Hình 3.2. Mô hình xử lý của chương trình Crawler

Trên thực tế, công việc Crawler còn nhiều vấn đề phức tạp hơn như việc phải xác thực tài khoản nếu nội dung dữ liệu muốn lấy về có liên quan đến các vấn đề bảo mật, các chuẩn của siêu liên kết, hiểu cấu trúc của trang web, lưu câu hỏi vào trong cơ sở dữ liệu… Việc đầu tiên là tìm một tài khoản có quyền truy cập vào nguồn dữ liệu xác định. Sau khi đăng nhập để xác thực với nguồn cung cấp dữ liệu xong, ta nhận được thông tin cookie, session của nguồn cung cấp dữ liệu được gửi về qua trình duyệt. Để trình Crawler chạy, ta cần gắn cookie, session đã đăng nhập. Sau đó, việc tiếp theo là tìm hiểu cấu trúc HTML để lấy được nội dung dữ liệu có ích chứa trong các thẻ và các siêu liên kết đến trang tiếp theo.

Sau khi thu thập dữ liệu xong, cần tổ chức lưu trữ nơi phù hợp để tiện cho việc lấy ra xử lý.

3.3.2 Xử lý dữ liệu

3.3.2.1 Xử lý thô

Là bước tiền xử lý sau khi thu thập dữ liệu từ nguồn cung cấp. Dữ liệu lấy về có số lượng rất lớn nhưng sau bước xử lý này có thể giảm đi vì ta không thể biết trước được có bao nhiêu dữ liệu bị trùng lặp, vô nghĩa. Để xây dựng tập dữ liệu huấn luyện tốt và chất lượng, việc tồn tại của các dữ liệu câu hỏi gây nhiễu cần được loại bỏ.

3.3.2.2 Xây dựng tập nhãn lớp và gán nhãn lớp câu hỏi

Đây là bước quan trọng để sử dụng tập dữ liệu đã qua xử lý thô làm dữ liệu huấn luyện bộ phân lớp. Các công việc được thực hiện tuần tự để đưa ra hai tập dữ liệu bao gồm tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Công việc đầu tiên là xây dựng tập nhãn lớp phù hợp. Mỗi nhãn lớp có những tính chất riêng cần xác định qua việc duyệt tập dữ liệu. Việc duyệt này cần thực hiện một cách thận trọng và có thể phải làm đi làm lại nhiều lần mới đưa ra được tập nhãn lớp chính xác. Công việc tiếp theo là gán nhãn lớp mỗi câu hỏi trong tập dữ liệu theo tập nhãn lớp đã xây dựng. Công việc này có thể thực hiện cùng lúc với việc tạo ra tập nhãn lớp nhưng ta nên làm riêng từng công việc một để tránh nhầm lẫn vì cả hai công việc này đều phải thực hiện bằng phương pháp thủ công.

Cuối cùng, dựa trên tập dữ liệu câu hỏi đã được gán nhãn lớp, ta chia ra hai tập là tập dữ liệu huấn luyện và tập dữ liệu kiểm tra theo tỉ lệ nhất định. Nếu số lượng câu hỏi trong tập dữ liệu thu thập được lớn, ta có thể chia theo tỉ lệ 70% câu hỏi huấn luyện và 30% câu hỏi kiểm tra. Bên cạnh đó cũng cần đảm bảo độ phủ của các nhãn lớp trong cả hai tập dữ liệu. Ta không nên để toàn bộ câu hỏi của một nhãn lớp chỉ tập trung ở một tập dữ liệu dù việc xây dựng hai tập dữ liệu sử dụng phương pháp ngẫu nhiên.

3.3.2.3 Gán nhãn từ loại tiếng Việt

Trong ngôn ngữ tự nhiên, đặc biệt là trong Tiếng Việt và một số ngôn ngữ khác, hệ thống từ loại rất đa dạng và phức tạp chẳng hạn như danh từ, động từ, tính từ, số từ, lượng từ, phó từ, … Một số từ loại lại được chia thành các nhóm từ loại nhỏ liên quan. Ví dụ, Danh từ được chia thành danh từ chỉ sự vật, danh từ chung, danh từ riêng, … Chúng ta cũng cần chú ý về cấu trúc của từ như từ đơn, từ ghép, từ láy, từ tượng thanh, tượng hình, ... Các từ đứng cạnh nhau có thể có nghĩa riêng và khi ghép vào thì lại mang một nghĩa khác. Vì vậy, gán nhãn từ loại tiếng Việt để xác định từ loại và tạo ra dấu hiệu chung phân biệt, giúp ánh xạ sang ma trận vector được chính xác hơn.

Ví dụ, đối với câu hỏi: **Quy chế hoạt động của diễn đàn,** hệ thống sẽ xử lý và câu hỏi được chuyển đổi thành **Quy_chế/**Danh_tu **hoạt_động/**Dong_tu **của/**Khong **diễn_đàn/**Danh_tu. Nếu để dữ liệu câu hỏi dạng đơn thuần thì từ "**Quy**" và "**chế**" có những ý nghĩa riêng hoặc vô nghĩa. Máy tính sẽ hiểu rằng đây là hai từ tách biệt nhau

và khi chuyển đổi sang ma trận vector, nó không có ý nghĩa trong việc tạo ra các đặc trưng trong phân lớp câu hỏi. Ngược lại, khi ghép hai từ "Quy" và "chế" lại với nhau, ta được một từ ghép có ý nghĩa và nó đóng vai trò là một danh từ..

3.3.2.4 Định dạng dữ liệu theo chuẩn SVM

Bước tiếp theo, toàn bộ câu hỏi sẽ được ánh xạ sang ma trận vector. Mỗi câu hỏi được xem như một tập hợp vector các giá trị đặc trưng. Khi áp dụng các giải thuật khác nhau, dữ liệu đầu vào phải được định dạng đúng.

Thuật toán SVM được sử dụng trong thực nghiệm yêu cầu dữ liệu đầu vào như sau:

[**nhãn lớp**] [$\boldsymbol{index_1}$]:[$\boldsymbol{value_1}$] [$\boldsymbol{index_2}$]:[$\boldsymbol{value_2}$] … **[**$\boldsymbol{index_n}$]: [$\boldsymbol{value_n}$]

[**nhãn lớp**]: là giá trị lớp ứng với nhãn lớp câu hỏi và là mục đích cần hướng tới của tập huấn luyện.

[$\boldsymbol{index_i}$]: là các đặc trưng trong câu hỏi nhằm xác định cho nhãn lớp, là giá trị đại diện cho từ loại đó.

[$\boldsymbol{value_i}$]: thể hiện tần số xuất hiện của [$\boldsymbol{index_i}$] trong câu.

Các nhãn lớp, từ loại trong câu sẽ được ánh xạ vào các tập hợp để lưu trữ trong quá trình chuyển đổi. Đại diện cho mỗi nhãn lớp, từ loại là một giá trị số tương ứng trong tập hợp. Khi xuất dữ liệu ra theo định dạng của SVM, các nhãn lớp và đặc trưng được ánh xạ ngược thành giá trị đại diện theo quy tắc sắp xếp giá trị vector từ nhỏ đến lớn, không có hai vector đặc trưng trùng nhau trong một câu. Vấn đề trùng nhau được xử lý bằng cách đánh trọng số cho vector. Do giá trị vector của các nhãn lớp không ảnh hưởng đến giá trị vector từ loại trong câu hỏi nên vector các nhãn lớp sẽ được lưu riêng và các vector từ đặc trưng sẽ được lưu riêng nhằm làm giảm không gian lưu trữ.

3.3.2.5 Tìm kiếm nhãn lớp có độ phân lớp chính xác cao

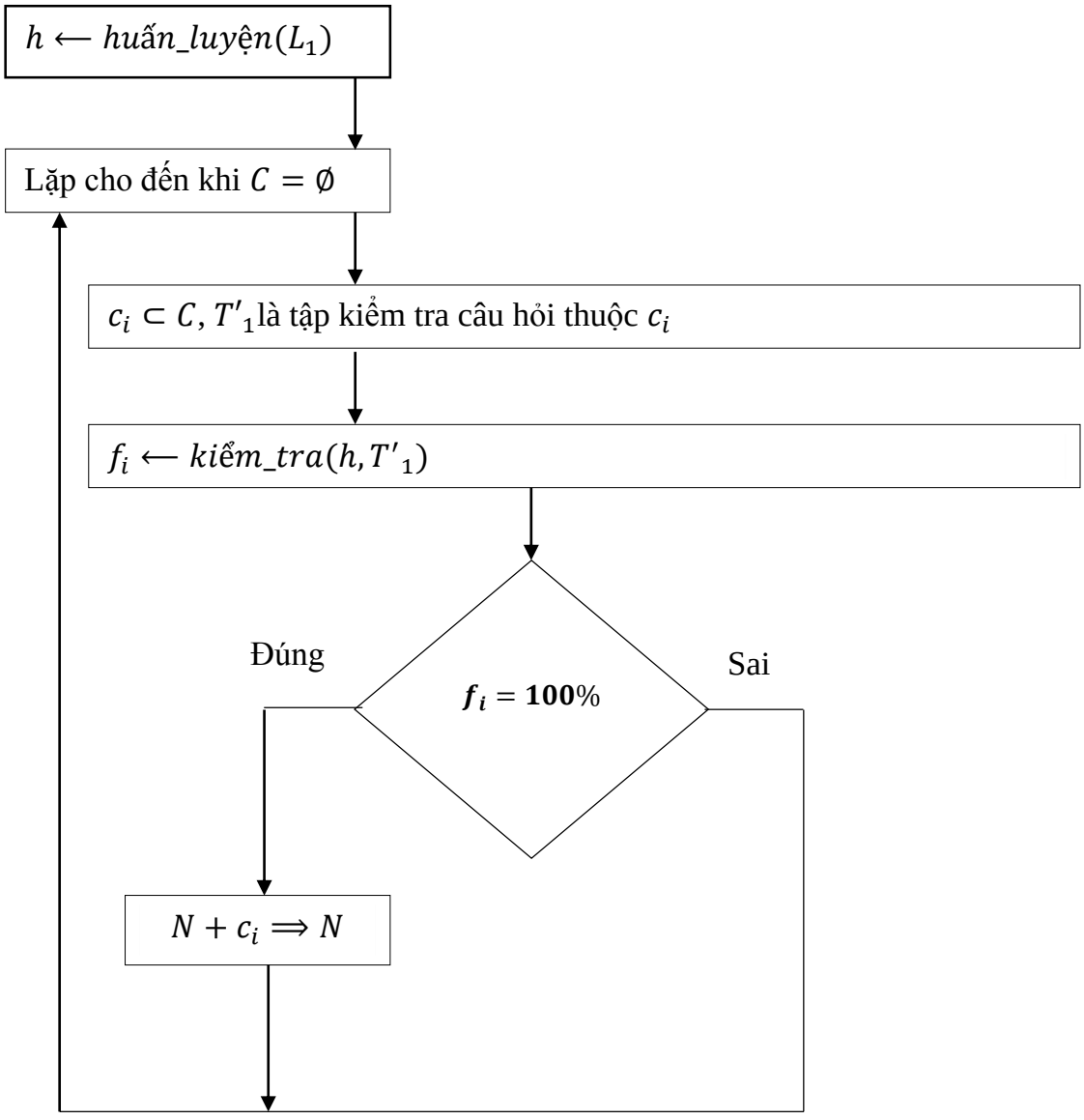
Trong bước xử lý này, tác giả áp dụng giải thuật tham lam (Greedy Algorithms) trong việc tìm kiếm các nhãn lớp yêu cầu. Tư tưởng của giải thuật là ở mỗi bước, giải thuật lựa chọn giải pháp tối ưu trong thời điểm hiện tại. Việc quyết định có sử dụng giải pháp đó hay không sẽ đưa ra hướng đi tiếp theo của giải thuật. Giải thuật này có một nhược điểm là khó đưa ra được giải pháp tối ưu toàn cục vì giải thuật chỉ xem xét quyết định ở bước hiện tại mà không cân nhắc đến toàn bộ quyết định đã thực hiện trong quá trình chạy.

Giải thuật tham lam tìm kiếm các nhãn lớp với dữ liệu chỉ trong tập dữ liệu huấn luyện. Tập dữ liệu huấn luyện ban đầu được tách thành tập dữ liệu huấn luyện mới và tập dữ liệu kiểm tra mới theo tỉ lệ phù hợp. Ở mỗi bước, giải thuật chọn ra một nhãn lớp

và loại bỏ toàn bộ dữ liệu câu hỏi được gán nhãn lớp đó trong tập dữ liệu huấn luyện mới và tập dữ liệu kiểm tra mới. Nếu kết quả phân lớp cho độ chính xác tăng lên so với kết quả phân lớp trước đó thì ta sẽ giữ nguyên hai tập dữ liệu. Ngược lại, nếu phân lớp cho kết quả độ chính xác giảm xuống so với kết quả ở nhãn lớp trước đó thì ta sẽ khôi phục lại dữ liệu câu hỏi đã bị xóa từ 2 tập dữ liệu mới. Quá trình sẽ được lặp lại với nhãn lớp đến khi duyệt hết toàn bộ tập nhãn lớp. Sơ đồ giải thuật được trình bày trong Hình 3.3.

Đầu vào: Tập dữ liệu huấn luyện ban đầu được tách thành tập dữ liệu huấn luyện L_1, tập dữ liệu kiểm tra mới T_1, tập nhãn lớp C, tập $N = \emptyset$

Đầu ra: Tập nhãn lớp N có độ phân lớp chính xác cao.



Hình 3.3. Sơ đồ giải thuật tham lam loại bỏ nhãn lớp có độ phân lớp chính xác cao

Tập dữ liệu huấn luyện ban đầu được sử dụng trong giải thuật tìm kiếm nhãn lớp có độ phân lớp chính xác cao còn được gọi là tập dữ liệu phát triển. Khi chia tách tập dữ liệu phát triển cần đảm bảo độ phủ của nhãn lớp trong mỗi tập dữ liệu. Kết thúc giải thuật, tập nhãn lớp có độ phân lớp chính xác cao được tìm thấy và sử dụng khi xây dựng bộ phân lớp cấp hai.

Chương 4: Ứng dụng vào hệ thống hỏi đáp thắc mắc tại Trung tâm đào tạo E-Learning

4.1 Giới thiệu trung tâm E-Learning

Trung tâm đào tạo E-Learning được ra đời năm 2009, nhằm thực hiện nhiệm vụ đào tạo từ xa theo phương thức E-Learning của Viện Đại học Mở Hà Nội. Qua một thời gian triển khai và tổ chức đào tạo, trung tâm cũng có một số thành tựu nhất định đóng góp vào sự phát triển chung của Viện Đại học Mở Hà Nội. Hiện nay, trung tâm đào tạo 6 ngành học: Quản trị kinh doanh, Kế toán, Công nghệ Thông tin, Tài chính Ngân hàng, Luật kinh tế và Ngôn ngữ Anh với số lượng lớn học viên đăng ký theo học tại nhiều đơn vị liên kết trên cả nước. Với phương pháp học Elearning, sinh viên cần chuẩn bị các thiết bị học tập: Máy tính hoặc các phương tiện di động có kết nối mạng. Tuy nhiên nếu có điều kiện đầu tư vào phương tiện hiện đại thì các môn học sẽ sinh động, dễ tiếp thu hơn. Sinh viên có thể học ở bất kỳ đâu khi có máy tính kết nối mạng. Sinh viên có thể trao đổi trực tiếp với giảng viên bằng hình ảnh qua webcam, bằng tiếng nói hoặc bằng gửi tin nhắn trong buổi học trên lớp Vclass theo lịch. Nếu chưa tới buổi học, sinh viên có thể đặt câu hỏi trên diễn đàn. Trong quá trình học, học viên có nhu cầu học tập trung có thể đăng ký với nhà trường để bố trí lớp. Nhà trường cũng thực hiện đánh giá thường xuyên qua các bài kiểm tra trắc nghiệm khi kết thúc mỗi bài học và bài thi kết thúc môn học.

Theo đặc thù của các ngành đào tạo cũng như chủ trương của nhà trường, trung tâm hết sức tạo điều kiện cho các học viên đã có bằng Trung cấp, Cao đẳng hoặc bằng Đại học khác theo học để nâng cao trình độ của mình thông qua việc miễn giảm các môn học, giảm thời gian theo học của học viên. Để làm được điều đó, trung tâm dưới sự tư vấn của Hội Đồng Đào Tạo Trực Tuyến Viện Đại học Mở Hà Nội đã đưa ra được chương trình thống nhất cho toàn bộ khóa học, kết hợp với việc phân loại học viên theo các trình độ khác nhau.

- Tốt nghiệp Phổ thông trung học
- Tốt nghiệp Trung cấp khác ngành học
- Tốt nghiệp Trung cấp cùng ngành học
- Tốt nghiệp Cao đẳng khác ngành học
- Tốt nghiệp Cao đẳng cùng ngành học
- Đã có bằng đại học khác
- Học song bằng

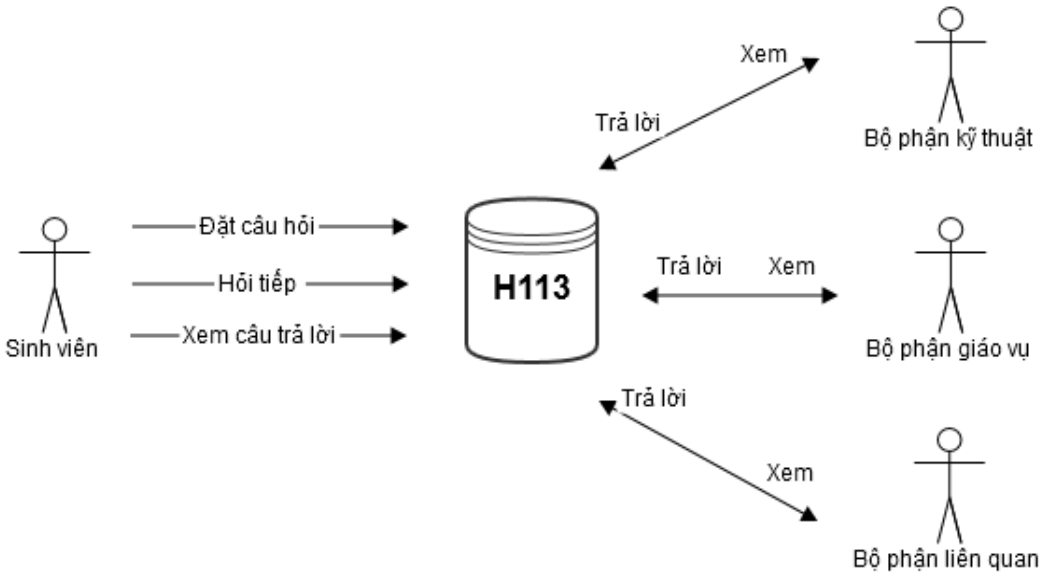
Dựa trên từng thành phần học viên, trung tâm sẽ cho phép học viên được miễn giảm một số môn học theo quy định, từ đó tiến tới giảm bớt thời lượng học tập cho từng loại học viên khác nhau. Việc xếp lịch học của từng học viên tại các học kỳ cũng được

tiến hành theo đúng những môn mà học viên phải học, và học phí của học viên tại từng kỳ học cũng được tính theo số môn học.

4.2 Tình trạng hệ thống hỏi đáp

Hệ thống hỏi đáp là một phương thức hỗ trợ cho sinh viên khi tham gia học tập trong môi trường học tập trực tuyến. Nó là một thành phần trong hệ thống EHOU của Trung tâm Đào tạo trực tuyến E-Learning, Viện Đại học Mở Hà Nội. Hệ thống còn có tên gọi khác là H113 – một điểm kết nối giữa sinh viên và nhà trường nhằm mục đích hỗ trợ trong quá trình học tập. Tên gọi H113 có ý nghĩa trong việc xử lý câu hỏi của sinh viên. Hệ thống cần một ngày để xử lý và giải đáp cho sinh viên đối với những câu hỏi về vấn đề kỹ thuật và đào tạo, và ba ngày để xử lý đối với những câu hỏi liên quan đến chuyên môn.

Chức năng chính của H113 là hỗ trợ học tập cho sinh viên bất cứ khi nào có vấn đề trong quá trình học tập, sinh viên có thể đặt câu hỏi cho bộ phận quản lý. Mỗi câu hỏi, thắc mắc của sinh viên được tiếp nhận và trả lời bởi một hoặc nhiều bộ phận liên quan. Việc giải quyết tốt các vấn đề nảy sinh trong quá trình học sẽ giúp sinh viên có được sự thoải mái nhất để tham gia học tập.



Hình 4.1. Mô hình xử lý của hệ thống H113 hiện tại

Khi có một câu hỏi, bộ phận quản lý chính sẽ tiếp nhận câu hỏi và đồng thời chuyển những thông tin, nội dung câu hỏi đến bộ phận liên quan để xử lý. Bộ phận xử lý sau khi đã nhận được câu hỏi thì phải tìm hiểu, xác định rõ vấn đề mà sinh viên đang gặp phải, đưa ra phương hướng giải quyết vấn đề và trả lời cho sinh viên ngay trên hệ thống.

Trong một vài trường hợp, bộ phận xử lý có thể trực tiếp liên hệ lại với sinh viên để giải thích cũng như tư vấn, đưa ra các biện pháp giải quyết chính xác nhất.

Ngoài vai trò là nơi tiếp nhận thông tin, xử lý thông tin và chỉ kết thúc chu trình khi xác định sinh viên đã hoàn toàn hài lòng, hệ thống H113 còn lưu trữ lại tất cả thông tin liên quan làm cơ sở dữ liệu cho sinh viên khác tham khảo.

Để xử lý tốt các câu hỏi, hệ thống cần sự phối hợp của nhiều bộ phận với nhau và áp dụng các quy trình đã đề ra. Sinh viên và người quản lý câu hỏi cũng cần nắm rõ một vài các yêu cầu như sau:

- Về phía sinh viên, khi gặp phải một vấn đề nào đó cần phải có kỹ năng mô tả chi tiết, có thể kèm theo hình ảnh là tốt nhất.
- Về phía người quản lý, cần phát triển cho mình kỹ năng xử lý những tình huống thường gặp, hiểu rõ quy trình, biết cách sử dụng các công cụ trợ giúp. Có kỹ năng chuyên nghiệp trong việc giao tiếp, biên soạn các tài liệu trả lời, hỗ trợ cho sinh viên. Khả năng nắm bắt thông tin từ câu hỏi để chuyển giao và phân phối câu hỏi cho các thành viên trong nhóm, tổ tư vấn để trả lời. Cập nhật thông tin, kiểm soát nhắc nhở các thành viên trả lời khi sinh viên đặt câu hỏi. Tìm giải pháp cải tiến phương pháp hỗ trợ.

Hiện tại, bộ phận trực hệ thống để tiếp nhận câu hỏi cũng chưa thể xử lý câu hỏi một cách nhanh chóng nhất. Có những trường hợp, cán bộ trực hệ thống không tiếp nhận kịp thời câu hỏi của sinh viên làm cho câu hỏi bị ngưng trệ trong quá trình xử lý. Mặc dù công việc của cán bộ trực hệ thống khá đơn giản, chỉ tiếp nhận câu hỏi và chuyển tới bộ phận liên quan. Nhưng nó lại là khâu đầu vào của quá trình hoạt động. Việc áp dụng mô hình phân lớp câu hỏi cho hệ thống này giúp cho khâu tiếp nhận câu hỏi không còn phụ thuộc quá nhiều vào cán bộ trực. Câu hỏi sẽ được tự động phân lớp và chuyển tới bộ phận liên quan mà không cần qua bước trung gian.

Trường hợp khi sinh viên có vấn đề thắc mắc trong quá trình học tập và muốn đặt câu hỏi lên hệ thống. Hệ thống sẽ thực hiện phân lớp tiêu đề câu hỏi trước để đưa ra một vài gợi ý về trường hợp mà sinh viên gặp phải. Điều này giúp sinh viên không cần đặt câu hỏi lên hệ thống mà vẫn lấy được những lời giải đáp thắc mắc từ câu hỏi đã được sinh viên khác thắc mắc. Hơn nữa, hệ thống sẽ không phải lưu thêm câu hỏi trùng lặp mà sinh viên đưa lên.

4.3 Chuẩn bị dữ liệu thực nghiệm

4.3.1 Thu thập dữ liệu

Do hệ thống hỏi đáp H113 tại Trung tâm E-Learning mới đi vào hoạt động phục vụ cho chương trình đào tạo mới của Viện Đại học Mở Hà Nội nên số lượng câu hỏi

được lưu trữ trong cơ sở dữ liệu chưa đủ để đáp ứng việc xây dựng bộ phân lớp. Vì vậy, để đảm bảo số lượng câu hỏi ở một mức phù hợp thì cần thu thập dữ liệu từ nguồn các đơn vị liên kết với trung tâm.

HOU-Topica là một đơn vị liên kết của Trung tâm E-Learning cùng chung mục tiêu đào tạo sinh viên trực tuyến. Đơn vị đã hoạt động được một thời gian trước khi chương trình đào tạo trực tuyến mới được xây dựng. Hiện nay Trung tâm E-Learning đã dừng tuyển sinh đối với chương trình liên kết với HOU-Topica. Trong thời kỳ liên kết, hai bên cũng đưa ra một số hệ thống hỏi đáp đóng vai trò là phương tiện trao đổi giữa sinh viên và nhà trường có tên gọi H2472. Mô hình hoạt động của hệ thống này chỉ dừng lại ở mức hỏi đáp thông thường mà hầu hết các hệ thống hỏi đáp thắc mắc đều áp dụng đó là: Người dùng đưa câu hỏi lên hệ thống, các chuyên gia sẽ đưa ra câu trả lời hoặc góp ý. Sau đó, hệ thống báo lại cho người đặt câu hỏi biết là câu hỏi đã được trả lời. Nếu không còn vấn đề gì cần hỏi lại từ phía người đặt câu hỏi thì câu hỏi sẽ đóng lại. Trong hệ thống H113 và H2472, người dùng đưa câu hỏi lên hệ thống chính là sinh viên và các chuyên gia trả lời là các cán bộ kỹ thuật, giảng viên, cố vấn học tập, … Dựa trên cơ sở đó, dữ liệu câu hỏi của hai hệ thống này về cơ bản có thể tương đồng và phù hợp để đưa vào làm dữ liệu phân lớp trong thực nghiệm.

Do lượng câu hỏi trong hệ thống H2472 tại HOU –Topica khá lớn và đều được lưu trữ trên website **http://elearning.hou.topica.vn/h2472** nên việc lấy dữ liệu thủ công là phương án không khả thi. Sử dụng chương trình Crawler là phương án khả thi hơn để đạt được kết quả mong muốn. Các vấn đề cần xử lý là tìm hiểu cấu trúc của trang, bóc tách từng thẻ trong cấu trúc để tìm thẻ lưu trữ nội dung câu hỏi, thẻ sử dụng để sang trang tiếp theo, xử lý xác thực, …

```
<tr>
    <td>55010</td>
    <td><a href="./?act=answers&do=detail&id=55010" title="">ACC505: Không vào được lớp học lại</a></td>
    <td>H2472 - Giải đáp chung</td>
  <td nowrap="nowrap">thomnt4375</td>
    <td nowrap="nowrap">09/04/16 16:09</td>
    <td style="white-space:nowrap;">
  40 giờ 31 phút                                    </td>
    <td nowrap="nowrap">Bùi Quỳnh Anh<br>qlht</td>

    <td nowrap="nowrap"><strong><img src='1luc.png' height='18px' align='absmiddle' />  Đã trả lời</strong></td>
</tr>

<tr>
```

Hình 4.2. Cấu trúc HTML lưu trữ nội dung câu hỏi

```
<div class="pagination">
    <b>
        <font size="1" face="Tahoma" color="#FF0000">1</font>
    </b>
    ,
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=20">
        <font size="1" face="Tahoma" color="#FF0000">2</font>
    </a>
    ,
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=40">
        <font size="1" face="Tahoma" color="#FF0000">3</font>
    </a>
    ...
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=8220">
        <span style="text-decoration: none">
            <span>412</span>
        </span>
    </a>
    ,
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=8240">
        <span style="text-decoration: none">
    </a>
    ,
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=8260">
        <span style="text-decoration: none">
            <span>414</span>
        </span>
    </a>
    <a href="?subject=0&gv=0&topic=&delay=&status=&searchtext=&attach=&do=&p=20">
</div>
```

Hình 4.3. Cấu trúc HTML phân trang dữ liệu

```
public Config() {
    this.name = "null";
    this.defaultCharset = "UTF-8";
    this.url = "";
    this.delay = 1000L;
    // this.nextPageRegex = "";
    this.elementSelector = "body";
    // this.maxPageNo = 1;
    Map<String, String> c=new HashMap<String, String>();
    c.put("URLCookie", "http%3A%2F%2Felearning.hou.topica.vn%2Fh2472%2F%3Fsubject%3D0
    c.put("_pk_id.2.7a97", "ea24380a2f47ddfd.1445912340.15.1451266843.1451266836.");
    c.put("_pk_ref.2.7a97", "%5B%22%22%2C%22%22%2C1451266836%2C%22http%3A%2F%2Felc.eh
    c.put("MoodleSession", "j2cg9drgfrnf0vk5jgatg0iqi7");
    c.put("MoodleSessionTest", "jqfBL10i8K");
    c.put("SID", "AcAAP%2F%2FSnJqBvbeNi6Och6aho7Kev6W3oJKYnKi%2FsqaalZuiq7WEoLp8d4WTk
    c.put("Iq6cJ0116W", "HY7dSoNgAEAfc3uBIW0TtwtHCN7W5x%2BKI%2B3T8i%2BXinbXjBaDAqdMzR
    c.put("MOODLEID_", "%25E2%25C2%2516X%25A7%2525%25E4%255E%25B2");
    c.put("user", "http%3A%2F%2Felearning.hou.topica.vn%2F");
    this.cookies = c;
}
```

Hình 4.4. Module xử lý xác thực crawler

```
try {
    // Get question
    DBConnect db = new DBConnect();
    Elements urls = doc.select(config.getElementSelector()).select(
            "table.list a[href]");
    if (urls != null) {
        for (Element element : urls) {
            db.insert(element.text());
        }
    }

    // Get Pagination
    urls = doc.select(config.getElementSelector()).select(
            ".pagination a[href]");
    if (urls != null) {
        for (Element element : urls) {
            String url = element.attr("abs:href").trim();
            url = url.replace("../", "");
            url = url.replace("./", "");

            if (url != null && url.trim().length() > 0) {

                // if (url.matches(config.getUrlRegex()))
                urlSet.add(URL.valueOf(url));
            }
        }
    }

} catch (RuntimeException e) {
```

Hình 4.5. Module xử lý dữ liệu crawler

Sau khi áp dụng các phương pháp chạy Crawler, kết quả đưa ra được là một tập gồm hơn 4000 câu hỏi ở dạng thô chưa xử lý. Ở giai đoạn tiếp theo, tập câu hỏi sẽ được xử lý để xây dựng tập dữ liệu chuẩn dùng trong phân lớp.

4.3.2 Xử lý dữ liệu

4.3.2.1 Xử lý thô

Với hơn 4000 câu hỏi đã được lấy về từ website của đơn vị liên kết, sau khi tiến hành xử lý sàng lọc, kiểm tra và loại bỏ các câu trùng lặp, vô nghĩa, hệ thống còn lại 1509 câu hỏi. Số lượng câu hỏi như vậy chỉ ở mức trung bình để tạo ra một bộ phân lớp tốt.

4.3.2.2 Xây dựng tập nhãn lớp và gán nhãn lớp câu hỏi

Từ tập dữ liệu đã xử lý thô, tác giả xây dựng tập nhãn lớp bằng cách duyệt qua từng câu. Sau một số lần duyệt toàn bộ tập dữ liệu một cách cẩn thận, tập nhãn lớp được hình thành với 22 nhãn lớp. Công việc tiếp theo là gán nhãn lớp cho tập dữ liệu hơn 1509 câu hỏi. Chi tiết các nhãn lớp cũng như số lượng câu hỏi đã gán cho từng nhãn lớp được trình bày trong Bảng 4.1.

Bảng 4.1. Bộ phân lớp câu hỏi tại trung tâm E-Learning

STT	Nhãn lớp	Ý nghĩa	Số lượng câu
1	CVHT	Cố vấn học tập	224
2	Dang_nhap	Đăng nhập	15
3	Dao_tao	Đào tạo	83
4	Diem	Điểm	153
5	Dien_dan	Diễn đàn	61
6	H113	H113	43
7	He_thong	Hệ thống	67
8	Hoc_phi	Học phí	49
9	Ky_thuat	Kỹ thuật	52
10	Lich	Lịch	31
11	Mail	Hòm thư cá nhân	28
12	Mien_mon	Miễn môn	43
13	Mobile	Ứng dụng trên điện thoại	34
14	Nhom	Nhóm	71
15	Tai_lieu	Tài liệu	122
16	The	Thẻ	15
17	Thi	Thi	83
18	Tot_nghiep	Tốt nghiệp	71
19	Trac_nghiem	Trắc nghiệm	203
20	Tu_luan	Tự luận	29
21	Tuyen_sinh	Tuyển sinh	9
22	VClass	Lớp học Online	23

Cuối cùng, tác giả xây dựng tập dữ liệu huấn luyện và tập dữ liệu kiểm tra từ tập 1509 câu hỏi đã được gán nhãn. Tỉ lệ cụ thể được chia là 90% câu hỏi huấn luyện và 10% câu hỏi kiểm tra. Như vậy, tập dữ liệu huấn luyện có 1359 câu hỏi và tập dữ liệu kiểm tra có 150 câu hỏi.

4.3.2.3 Gán nhãn từ loại tiếng Việt

Với mỗi câu hỏi trong tập dữ liệu huấn luyện và tập dữ liệu kiểm tra, tác giả tiến hành chuẩn hóa từ loại, ngữ nghĩa tiếng Việt. Tác giả sử dụng chương trình vnTagger

của Lê Hồng Phương[10] được phát triển bằng ngôn ngữ java để xử lý ngôn ngữ tự nhiên, gán nhãn cho các từ loại tiếng Việt. Chương trình có độ chính xác khá cao khoảng 95%. Để thuận tiện cho việc quản lý và thống kê, câu hỏi được xử lý và lưu trong cơ sở dữ liệu.

```
Dien_dan Quy_chế/N hoạt_động/V của/E diễn_đàn/N
Diem Xử_lý/V việc/N chấm/V bài/N khi/N ngân_hàng/N câu_hỏi/N thay_đổi/V đáp_án/N như_thế_nào/X
Nhom Em/N không/R có/V tên/N trong/E nhóm/N
H113 Tôi/P chưa/R được/V kích_hoạt/V tài_khoản/N H113/Np
He_thong Khi/N gửi/V bài_tập/N đi/V mất/V nhiều/A thời_gian/N chờ/V hệ_thống/N
Dao_tao Khi/N nào/P được/V đăng_ký/V học/V vượt/V
Hoc_phi Đổi/V ngày/N đóng/V học_phí/N thì_có/X được/V thông_báo/V
Dao_tao Mua/V học/V liệu/V bắt_buộc/V là/V quy_định/N của/E Viện/N hay/C của/E riêng/A ELearning/Np
Hoc_phi Số/N tài_khoản/N ngân_hàng/N để/E đóng/V học_phí/N
H113 Test/V hệ_thống/N H113/Np
Diem Điểm/N sẽ/R được/V thể_hiện/V ở/E đâu/P
Tai_lieu Học/V liệu/V của/E trường/N có/V khác/A với/E tài_liệu/N trên/E web/N
CVHT Về/V mã_số/N học_viên/N khác/A với/E quy_định/N
Dao_tao Chất_lượng/N học_viên/N của/E chương/N trình/V Ehou/Np
Mail Cần/V hỗ_trợ/V reset/N password/V vào/E mail/N học_viên/N
Tuyen_sinh Tại_sao/X phải/V ký/V cam_kết/V hoàn_thiện/V hồ_sơ/N khi/N xét/V tuyển/V
CVHT Có_thể/R chuyển/V nơi/N học_tập/V
CVHT Thắc_mắc/V về/E các/L đề_tài/N bình_luận/A của/E giảng_viên/N
CVHT Trong/E hồ_sơ/N cá_nhân/N quê_quán/N không/R đúng/A
CVHT Sao/P kế_hoạch/N tuần/N 6/M lại/R bỏ/V trống/A trong/E khi/N tuần/N 7/M bị/V dồn/V
Dien_dan Thay_đổi/V hình_ảnh/N đại_diện/V trên/E diễn_đàn/N
Tot_nghiep Tải/N ảnh/N tốt_nghiệp/V ở/E đâu/P
```

Hình 4.6. Câu hỏi đã được gán nhãn từ loại

4.3.2.4 Định dạng dữ liệu theo chuẩn SVM

Công việc cuối cùng là ánh xạ đặc trưng câu hỏi đã được gán nhãn lớp, từ loại sang vector theo định dạng dữ liệu đầu vào chuẩn cho SVM. Để tạo ra được dữ liệu đầu vào này, tác giả đã xây dựng một số công cụ chuyển đổi lớp nhãn, từ loại thành các giá trị đặc trưng sử dụng ngôn ngữ java.

```
while ((strLine = br.readLine()) != null) {
    String[] strWord = strLine.toString().split(" ");
    String[] strWordNew = Arrays.copyOfRange(strWord, 1,
            strWord.length);
    for (String s : strWordNew) {
        if (!featureMap.containsKey(s)) {
            featureMap.put(s, counter);
            counter++;
        }
    }
}
```

Hình 4.7. Module xử lý đưa nhãn, từ loại vào tập hợp

```
while ((strLine = br.readLine()) != null) {
    String[] strWord = strLine.toString().split(" ");
    ArrayList<Integer> stNumber = new ArrayList<Integer>();
    firstKey = strWord[0];
    for (String s : strWord) {
        if (s == firstKey) {
            if (!storeLabel.containsKey(firstKey)) {
                storeLabel.put(firstKey, counterLabel);
                stNumber.add(counterLabel);
                counterLabel++;
            } else {
                stNumber.add(storeLabel.get(s));
            }
            continue;
        }
        stNumber.add(storeVector.get(s));
    }
    Collections.sort(stNumber.subList(1, stNumber.size()));
    storeWordList.addAll(getTermFrequency(stNumber));
}
storeWordList.remove(storeWordList.size() - 1);
```

Hình 4.8. Module xuất dữ liệu theo định dạng SVM

Kết quả của quá trình xử lý dữ liệu là 1 file đầu vào có dữ liệu như Hình 4.9.

```
30 3387:1 3720:1 10619:1 10620:1 10833:1 10834:1 12853:1 12885:1 14459:1 14460:1 17683:1 34185:1 34187:1
39 64:1 875:1 5842:1 5927:1 9340:1 9378:1 12591:1 12592:1 15449:1 20040:1 20041:1 26212:1 26213:1 29749:1
4 1079:1 1449:1 25791:1 25793:1 33562:1 34633:1 34636:1
40 1079:1 1087:1 3387:1 3612:1 9728:1 9729:1 15204:1 15222:1 33727:1 35412:1 35414:1
21 64:1 556:1 3387:1 4161:1 5842:1 6194:1 14306:1 15653:1 15655:1 16686:1 16687:1 21720:1 21721:1 34633:1
36 1079:1 1087:1 3387:2 4854:1 4963:1 6236:1 6246:1 16140:1 20738:1 20740:1 23990:1 23992:1 24719:1 24720
7 64:1 871:1 3387:2 3447:1 4243:1 5290:1 5295:1 6236:1 6246:1 7802:1 10230:1 10231:1 16557:1 16558:1 2970
5 5559:1 5587:1 5842:1 6150:1 8668:1 11366:1 11380:1 18364:1 18367:1 25533:1 25535:1 28666:1 28689:1 2882
18 64:1 128:1 3387:1 3520:1 7798:1 7799:1 8469:1 8597:1 20008:1 20009:1 20419:1 20420:1 20545:1 20546:1 2
39 3387:1 4417:1 5842:1 5850:1 9340:1 9486:1 18531:1 18559:1 20936:1 27256:1 27262:1 31295:1 31296:1 3538
39 64:1 75:1 2093:1 2316:1 3387:1 4893:1 5542:1 5545:1 7104:1 7105:1 9340:1 9457:1 22853:1 25192:1 25193:
5 1079:1 1502:1 2045:1 2078:1 3178:1 3252:1 13758:1 28068:1 28069:1 28991:1 28999:1 29953:1 29959:1 30447
31 2093:1 2199:1 3387:1 4414:1 13124:1 13127:1 14594:1 17418:1 17432:1 18474:1 18475:1 24927:1 24928:1 32
14 3387:1 3865:1 12122:1 12132:1 12749:1 12751:1 17498:1 17499:1 20408:1 20409:1 21901:1 21908:1 29953:1
30 3387:1 4494:1 5842:1 5850:1 19692:1 26904:1 26905:1 33638:1 33639:1 34633:1 35171:1
5 1079:1 1505:1 2409:2 2620:1 2735:1 6979:1 6981:1 7097:1 7100:1 10707:1 10708:1 13252:1 17534:1 17535:1
```

Hình 4.9. Nội dung file theo định dạng SVM

Sử dụng file dữ liệu định dạng đã được tạo ra để thực nghiệm với SVM. Tác giả sử dụng libSVM[3] để tạo bộ phân lớp từ dữ liệu huấn luyện. Các thông số sử dụng trong thực nghiệm để mặc định. Khi chạy dữ liệu huấn luyện, ta được kết quả là một file mô hình. File này chính là bộ phân lớp được dùng để đánh giá dữ liệu kiểm tra và cho ra độ chính xác, số lượng câu hỏi được gán đúng nhãn.

4.3.2.5 Tìm kiếm nhãn lớp có độ phân lớp chính xác cao

Áp dụng mô hình tìm kiếm nhãn lớp sử dụng giải thuật tham lam với tập dữ liệu chuẩn của Li và Roth. Tập dữ liệu huấn luyện ban đầu gồm 5452 câu hỏi được dùng làm tập dữ liệu phát triển. Tác giả chia tập dữ liệu này thành là tập dữ liệu huấn luyện mới gồm 4899 câu hỏi và tập dữ liệu kiểm tra mới gồm 553 câu hỏi. Sau khi áp dụng giải thuật, kết quả các nhãn lớp được thể hiện trong Bảng 4.2.

Bảng 4.2. Thống kê kết quả tìm kiếm nhãn lớp có độ phân lớp chính xác cao với dữ liệu chuẩn của Li và Roth

STT	Nhãn lớp	Số câu hỏi kiểm tra	Độ chính xác (%)
1	ABBR:abb	4	100
2	ABBR:exp	6	50
3	DESC:def	42	92.85
4	DESC:desc	26	50
5	DESC:manner	24	100
6	DESC:reason	22	81.81
7	ENTY:animal	6	50
8	ENTY:body	1	0
9	ENTY:color	7	85.71
10	ENTY:cremat	27	62.96
11	ENTY:currency	1	100
12	ENTY:dismed	1	100
13	ENTY:event	1	100
14	ENTY:food	9	55.55
15	ENTY:instru	1	100
16	ENTY:lang	1	100
17	ENTY:letter	1	100
18	ENTY:other	21	38.09
19	ENTY:plant	2	0
20	ENTY:product	3	33.33
21	ENTY:religion	1	0
22	ENTY:sport	6	83.33
23	ENTY:substance	3	66.66
24	ENTY:symbol	1	100
25	ENTY:techmeth	5	40
26	ENTY:termeq	11	45.45

27	ENTY:veh	4	25
28	ENTY:word	2	50
29	HUM:desc	6	100
30	HUM:gr	19	52.63
31	HUM:ind	104	89.42
32	HUM:title	4	50
33	LOC:city	9	100
34	LOC:country	21	90.47
35	LOC:mount	5	100
36	LOC:other	42	83.33
37	LOC:state	7	71.42
38	NUM:code	1	100
39	NUM:count	40	95
40	NUM:date	22	86.36
41	NUM:dist	6	50
42	NUM:money	5	80
43	NUM:ord	1	100
44	NUM:other	5	80
45	NUM:perc	4	75
46	NUM:period	9	100
47	NUM:speed	1	100
48	NUM:temp	1	0
49	NUM:volsize	1	100
50	NUM:weight	1	0

Tác giả xác định được tập nhãn lớp có độ phân lớp chính xác cao trong bộ dữ liệu của Li và Roth gồm 17 nhãn lớp có độ chính xác 100% trong Bảng 4.2.

Tương tự, tác giả áp dụng giải thuật tham lam để tìm các nhãn lớp có độ phân lớp chính xác cao trong tập dữ liệu phát triển tại trung tâm E-Learning. Tập dữ liệu huấn luyện ban đầu gồm 1359 câu hỏi được chia thành tập dữ liệu huấn luyện mới gồm 1222

câu hỏi và tập dữ liệu kiểm tra mới gồm 137 câu hỏi. Kết quả chi tiết được thể hiện trong Bảng 4.3.

Bảng 4.3. Thống kê kết quả tìm kiếm nhãn lớp có độ phân lớp chính xác cao với dữ liệu tại trung tâm E-Learning

STT	**Nhãn lớp**	**Số câu hỏi kiểm tra**	**Độ chính xác (%)**
1	Tuyen_sinh	2	100
2	The	4	100
3	Dang_nhap	1	0
4	VClass	1	100
5	Mail	3	33.33
6	Lich	2	50
7	Tu_luan	1	100
8	Mobile	3	0
9	Mien_mon	1	100
10	H113	8	87.5
11	Hoc_phi	2	100
12	Ky_thuat	9	33.33
13	Dien_dan	3	100
14	He_thong	7	71.42
15	Nhom	8	100
16	Tot_nghiep	4	100
17	Thi	7	85.7
18	Dao_tao	9	44.44
19	Tai_lieu	13	61.53
20	Diem	13	84.61
21	CVHT	16	93.75
22	Trac_nghiem	20	75

Theo thống kê trên, các nhãn lớp có độ phân lớp chính xác cao đã được xác định là "**Tuyển sinh**", "**Thẻ**", "**VClass**", "**Tự luận**", "**Miễn môn**", "**Học phí**", "**Diễn đàn**",

"**Nhóm**", và "**Tốt nghiệp**". Các lớp này sẽ bị loại ra khỏi tập dữ liệu huấn luyện ban đầu để xây dựng bộ phân lớp cấp hai.

4.3.3 Áp dụng mô hình đề xuất

Để áp dụng mô hình để xuất, tác giả đã tạo ra hai bộ phân lớp. Bộ phân lớp cấp một được tạo từ toàn bộ câu hỏi huấn luyện ban đầu. Bộ phân lớp cấp hai được tạo từ tập dữ liệu câu hỏi huấn luyện đã loại bỏ các câu hỏi được gán nhãn lớp có độ phân lớp chính xác cao. Các câu hỏi trong tập dữ liệu kiểm tra sẽ đi qua lần lượt hai bộ phân lớp. Nếu câu hỏi được gán nhãn lớp thuộc các lớp có độ phân lớp cao thì câu hỏi đó không cần phải phân lớp với bộ phân lớp cấp hai. Ngược lại, các câu hỏi kiểm tra sẽ đi tiếp qua bộ phân lớp cấp hai. Kết quả áp dụng mô hình đề xuất được trình bày trong phần tiếp theo của luận văn.

4.4 Kết quả thực nghiệm

4.4.1 Cài đặt môi trường thực nghiệm

Thực nghiệm tiến hành trên máy chủ Linux có cấu hình được trình bày trong Bảng 4.4.

Bảng 4.4. Cấu hình máy chủ trong thực nghiệm

STT	**Thông số phần cứng**	
1	CPU	Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
2	RAM	2Gb
3	HDD	15Gb
	Thông số phần mềm	
4	Hệ điều hành	CentOS 6.5
5	Thư viện	libsvm v2.9
6	Gói hỗ trợ	Các gói cần thiết trong quá trình chạy như gcc, gcc-c+, gmp, libstdc-devel, glibc-devel

4.4.2 Thực nghiệm với dữ liệu của Li và Roth

Trong bộ dữ liệu của Li và Roth, tập dữ liệu bao gồm 5952 câu hỏi. Hai bộ phân lớp được xây dựng theo như trong mô hình đề xuất. Kết quả phân lớp với mỗi bộ phân lớp được trình bày trong Bảng 4.5.

Bảng 4.5. Kết quả thực nghiệm với dữ liệu của Li và Roth

STT	Bộ phân lớp	Số lượng câu hỏi huấn luyện	Số lượng câu hỏi kiểm tra	Số lượng câu đúng	Độ chính xác(%)
1	Bộ phân lớp cấp một	5452	500	39/54	72.22
2	Bộ phân lớp cấp hai	4642	446	373/446	83.63
	Tổng			**412/500**	**82.4**

Sau khi phân lớp câu hỏi qua bộ phân lớp cấp một, tác giả nhận được 39 câu hỏi chính xác trong 54 câu hỏi kiểm tra có độ chính xác cao. Trong thực nghiệm tìm kiếm nhãn lớp có độ chính xác cao, tác giả đã tìm ra tập nhãn lớp yêu cầu với dữ liệu của Li và Roth. Bộ phân lớp cấp hai sẽ được xây dựng bằng cách loại bỏ các câu hỏi đã gán nhãn lớp này. Cụ thể, số lượng câu hỏi trong tập dữ liệu huấn luyện phải loại đi là 810 câu hỏi. Khi qua bộ phân lớp cấp hai, số lượng câu chính xác nhận được là 373 câu hỏi. Như vậy trong tổng số 500 câu hỏi kiểm tra, có 412 câu hỏi được gán chính xác nhãn lớp (gồm 39 câu hỏi kiểm tra thuộc các nhãn lớp có độ phân lớp chính xác cao và 373 câu hỏi chính xác ở phân lớp cấp hai) và cho kết quả độ chính xác tăng lên **82.4%** đối với tập dữ liệu này.

4.4.3 Thực nghiệm với dữ liệu tại Trung tâm E-Learning

4.4.3.1. Phương pháp K-Fold Cross Validation

Cross validation là phương pháp chia tập dữ liệu gốc thành tập dữ liệu huấn luyện để huấn luyện mô hình và tập dữ liệu kiểm tra để đánh giá mô hình đó. Phương pháp này thường được áp dụng khi mà tập dữ liệu gốc có kích thước vừa và nhỏ. Mục đích của phương pháp cross validation là đánh giá công bằng vai trò mỗi dữ liệu. Tập dữ liệu gốc được chia ngẫu nhiên thành k tập dữ liệu có kích thước bằng nhau. K tập dữ liệu cần đảm bảo không giao nhau. Trong k lần lặp, 1 tập dữ liệu sẽ làm nhiệm vụ kiểm tra mô hình được tạo bởi k-1 tập dữ liệu còn lại. Độ phân lớp chính xác cuối cùng là trung bình cộng của độ phân lớp chính xác tại k vòng lặp.

Tác giả chọn $k = 10$ và áp dụng với tập dữ liệu 1509 câu hỏi tại trung tâm E-Learning. Độ phân lớp chính xác trung bình nhận được là **79.32%** (1197/1509). Kết quả này được dùng để so sánh với mô hình cải tiến trong thực nghiệm tiếp theo.

4.4.3.2. Phương pháp cải tiến

Sau quá trình xử lý dữ liệu câu hỏi, các tập dữ liệu huấn luyện và dữ liệu kiểm tra được tạo ra. Với 22 nhãn lớp bao phủ 1359 câu hỏi huấn luyện và 150 câu hỏi kiểm tra. Kết quả phân lớp không tổ chức theo mô hình phân cấp đạt 80.7%. Kết quả này được dùng để so sánh với kết quả thực nghiệm sử dụng mô hình phân lớp cải tiến.

Tác giả xây dựng bộ phân lớp cấp một dựa trên 1359 câu hỏi huấn luyện. Tiếp theo, tác giả áp dụng giải thuật tham lam để tìm kiếm các nhãn lớp có độ phân lớp chính xác cao rồi loại bỏ các nhãn lớp này khỏi tập dữ liệu huấn luyện. Tổng số câu hỏi huấn luyện thuộc nhãn lớp có độ phân lớp chính xác cao là 517 câu hỏi. Để xây dựng bộ phân lớp cấp hai, ta cần loại bỏ 517 câu hỏi đó. Kết quả phân lớp của từng bộ phân lớp áp dụng vào tập dữ liệu tại Trung tâm E-Learning được liệt kê như trong Bảng 4.6.

Bảng 4.6. Kết quả thực nghiệm với dữ liệu tại Trung tâm E-Learning

STT	**Bộ phân lớp**	**Số lượng dữ liệu huấn luyện**	**Số lượng dữ liệu kiểm tra**	**Số lượng câu đúng**	**Độ chính xác(%)**
1	Bộ phân lớp cấp một	1359	150	65/78	83.33
2	Bộ phân lớp cấp hai	842	72	59/72	81.94
	Tổng			**124/150**	**82.67**

Số lượng câu hỏi kiểm tra được gán nhãn lớp có độ phân lớp chính xác cao là 78 câu hỏi. Sau khi qua bộ phân lớp cấp một, tác giả nhận được 65/78 câu hỏi thực sự chính xác. Kết quả phân lớp với bộ phân lớp cấp hai, tổng số câu chính xác là 59/72 câu hỏi. Tổng số câu hỏi được phân lớp chính xác qua hai bộ phân lớp là 124 câu hỏi, đạt độ chính xác là **82.67%**. Như vậy, giải pháp đề xuất đã cải tiến được hiệu năng phân lớp. Cụ thể, kết quả phân lớp tăng **1.97%** độ chính xác khi so sánh với việc không sử dụng hai tầng phân lớp và tăng **3.35%** khi so sánh với phương pháp K-fold Cross Validation.

4.5 Kết luận

Kết quả của hai thực nghiệm đã cho thấy rằng việc áp dụng mô hình đề xuất làm tăng độ chính xác trong phân lớp câu hỏi. Trong thực nghiệm được thực hiện với tập dữ liệu của Li và Roth – một tập dữ liệu có số lượng lớn và mang tính khách quan, mô hình đề xuất được áp dụng cho độ chính xác tăng **1%**. Đối với tập dữ liệu tại trung tâm E-Learning, kết quả cho thấy độ chính xác đã tăng **1.97%**. Tuy còn tồn tại nhiều vấn đề chưa thực sự giải quyết hết nhưng độ chính xác phân lớp tăng lên là động lực để thúc đẩy việc hoàn thiện những thiếu sót trong quá trình chuẩn bị và xử lý dữ liệu. Mô hình đề xuất cũng chứng tỏ hiệu suất phân lớp có liên quan tới không gian nhãn lớp của câu hỏi. Không gian nhãn lớp giảm có thể làm tăng hiệu suất qua đó tăng thêm độ chính xác.

Kết luận và hướng phát triển tương lai

Phân lớp câu hỏi là nhiệm vụ quan trọng trong mỗi hệ thống hỏi đáp. Câu hỏi được phân lớp chính xác là tiền đề cho quá trình xử lý tiếp theo. Nhiều đề xuất cải tiến được thực hiện nhằm tăng độ chính xác phân lớp, qua đó làm tăng hiệu suất chung của hệ thống hỏi đáp. Trong luận văn "Nghiên cứu mô hình phân lớp câu hỏi và ứng dụng", tác giả cũng đã đề xuất cải tiến mô hình giúp tăng độ chính xác. Bên cạnh đó, luận văn còn đạt được một số kết quả như sau:

- Khái quát vấn đề phân lớp câu hỏi, nêu lên vai trò và ý nghĩa của quá trình phân lớp trong hệ thống hỏi đáp. Khảo sát và thống kê các dạng câu hỏi trong ngôn ngữ tự nhiên có thể xuất hiện trong phân lớp.
- Nghiên cứu, tìm hiểu các hướng để tiếp cận mô hình phân lớp và giải thuật áp dụng.
- Xây dựng các bước xử lý dữ liệu phân lớp và đề xuất mô hình phân lớp có khả năng làm tăng độ chính xác.
- Trong thực nghiệm, luận văn ứng dụng mô hình phân lớp đề xuất với dữ liệu câu hỏi tại trung tâm E-Learning. Xây dựng module xử lý dữ liệu câu hỏi từ nguồn dữ liệu hiện có ở trung tâm và các nguồn từ đơn vị liên kết

Tuy nhiên, luận văn cũng còn tồn tại một số hạn chế:

- Số lượng câu hỏi phục vụ cho nhiệm vụ phân lớp vẫn còn ít nên có thể độ chính xác của bộ phân lớp chưa cao.
- Việc gán nhãn lớp cho các câu hỏi vẫn chủ quan, dựa vào kiến thức cá nhân là chủ yếu nên các lớp nhãn có thể chưa phù hợp.

Trong thời gian tới, tác giả sẽ tiếp tục nghiên cứu về phân lớp câu hỏi cho việc ứng dụng vào hiện tại, mở rộng số lượng câu hỏi huấn luyện tới mức có thể chấp nhận được (3000 câu) và tiến hành làm giàu thêm các đặc trưng cho từng câu hỏi trong bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra. Việc này có ý nghĩa quan trọng trong việc nâng cao độ chính xác cho bộ phân lớp câu hỏi. Nhiều thuật toán khác nhau sẽ được sử dụng để có thể đưa ra thuật toán phù hợp hơn với ứng dụng phân lớp câu hỏi trong hệ thống hỏi đáp thắc mắc H113 tại Trung tâm E-Learning.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2009), *Giáo trình khai phá dữ liệu Web*, Nhà xuất bản Giáo dục Việt Nam.

Tiếng Anh

2. Anders Søgaard (2010), *Simple semi-supervised training of part-of-speech taggers*, The 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, Sweden.
3. Chih-Chung Chang and Chih-jen Lin (2013), *LibSVM: A library for Support Vector Machine*, Department of Computer Science National Taiwan University, Taipei, Taiwan.
4. David Tom, Claudio Giuliano (2009), *A semi-supervised approach to question classification*, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning.
5. Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, Amardeep Grewal (2002), *Probabilistic question answering on the web*, Journal of the American society for Information Science and Technology 2005.
6. Hakan Sundblad (2007), *Question Classification in Question Answering systems*, Submitted to Linköping Institute of Technology at Linköping University.
7. John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan , Bill Ogden,John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, Ralph Weishedel (2002), *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering*. Q&A Roadmap Paper
8. Oliver Chapelle, Bernhard Scholkopf, Alexander Zien (2006), *Semi supervised learning*, The MIT Press Cambridge, Massachusetts, London, England
9. Pierre Baldi, Paolo Frasconi, Padhraic Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Published by John Wiley & Sons Ltd, The Southern Gate, Chichester West Sussex PO19 8SQ, England - 2003.
10. Le Hong Phuong (2010), *An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts*. Actes du Traitement Automatique des Langues Naturelles (TALN-2010), Montreal, Canada.
11. Nguyen Tri Thanh, Nguyen Le Minh and Akira Shimazu (2008). *Using Semi-supervised Learning for Question Classification*, Journal of Natural Language Processing (15).

12. Nguyen Tri Thanh, Nguyen Le Minh and Akira Shimazu (2007), *Improving the Accuracy of Question Classification with Machine Learning*, Institute of Electrical and Electronics Engineers(IEEE).
13. Xin Li, Dan Roth (2002), *Learning question classifiers*, In Proceedings of the 19th International Conference on Compuatational Linguistics (COLING), pp.556–562.
14. Xin Li, Dan Roth (2004) . *Learning question classifiers: the role of semantic information*, Cambridge University Press.