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

**TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ 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

**TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN**

**Hà Nội - 2016**

# MỤC LỤC

# MỞ ĐẦU

Ngày nay, với sự phát triển cơ sở hạ tầng công nghệ đặc biệt là công nghệ mạng đã thúc đẩy nhu cầu được trao đổi, chia sẻ dữ liệu của con người, làm cho 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, những 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 tìm kiếm nhưng câu trả lời tiềm năng. Cuối cùng, trả về cho người dùng một kết quả ngắn gọn, súc tích và chính xác nhất. Để có thể đưa ra những tiêu chí trong 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 xác định được lớp câu hỏi, 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 nhiệm vụ quan trọng nhất.

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 cũng nảy sinh là việc xử lý ngôn ngữ tự nhiên. Một số ngôn ngữ có hệ thống từ loại rất đa dạng và phức tạp. Trong những ngôn ngữ như tiếng Việt, xác định các đặc trưng ngữ nghĩa và đưa ra chiến lược xử lý là công việc tương đối vất vả, mất nhiều thời gian. Do đó 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 này.

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 thứ 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 hiện tại 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 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 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.

Một câu hỏi trong ngôn ngữ tự nhiên 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 ràng buộc về loại câu trả lời, cung cấp thông tin xử lý để đưa ra một hoặc nhiều chiến lược phân lớp nhằm làm giảm không gian tìm kiếm các câu trả lời tiềm năng trong kho ngữ liệu khổng lồ. Bên cạnh đó, 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. Công việc xác định chính xác ngữ nghĩa cho câu hỏi là một thách thức không hề nhỏ.

## 1.2 Tìm hiểu các loại câu hỏi

Xác định loại câu hỏi mang một ý nghĩa to lớn trong phân tích các câu hỏi bởi đối với mỗi loại câu hỏi sẽ có những đặc trưng và cách tiếp cận khác nhau. Mỗi loại câu hỏi thì cần có chiến lược xử lý phù hợp.

## 1.3 Taxonomy câu hỏi

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 các nút con của các nút cha là 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 sẽ xác định được các nút con của nó. 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.

## 1.4 Mục tiêu của luận văn

Ban đầu, phân lớp câu hỏi chỉ tập trung vào phân lớp phẳng nhưng dần dần có nhiều vấn đề nảy sinh cần phải được đáp ứng nên phân lớp phẳng không còn phù hợp mà thay vào đó là các mô hình phân lớp cục bộ (Local Classifier), phân lớp toàn cục (Global Classifier hay Big-Bang), phân lớp phân cấp (Hierarchical Classifier)…

Sau một số tìm hiểu, nghiên cứu về các miền câu hỏi cụ thể và thấy rằng kết quả phân lớp của một số lớp có tỉ lệ chính xác rất cao còn một số khác thì lại kém 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 kém hơn. Kết quả phân lớp ở các lớp có độ chính xác thấp hơn làm tăng độ chính xác chung trong nhiệm vụ phân lớp.

# Chương 2: Các phương pháp tiếp cận cho 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 chỉ sử dụng bộ phân lớp phẳng, các mối quan hệ bên trong của nhãn lớp bị bỏ qua, đặc biệt là sử dụng toàn bộ lớp nhãn trong một thời điểm với 1 dữ liệu câu hỏi.

### 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, ..., được xem như sự bổ sung và cải tiến của một số phương pháp phân lớp khác. Ý tưởng cơ bản trong mô hình này là giảm số lượng các lớp nhãn trong tập để cử cho mỗi câu hỏi theo từng bước. Đầu ra của một phân lớp là một tập nhãn lớp được sử dụng làm bộ phân lớp trong lần phân lớp tiếp theo. Khi ở phân lớp cấp 1 câu hỏi đã được phân vào lớp tổng thể, lớp này đã được bao quát hơn rất nhiều so với các lớp con.

## 2.2 Giải thuật phân lớp câu hỏi

Về cơ bản, phân lớp câu hỏi thường sử dụng 2 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.

### 2.1.3 Giải thuật học máy có giám sát

Trong học máy có giám sát, chương trình học sẽ được cung cấp 2 bộ 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 mà nhận biết dữ liệu chưa được gán nhãn trong tập dữ liệu kiểm tra với độ chính xác cao nhất có thể..

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) và Spare Network of Winnows (SNoW).

### 2.1.4 Giải thuật học máy bán giám sát

Trong lịch sử của học máy bán giám sát, có lẽ ý tưởng đầu tiên về việc tận dụng các đặc trưng có trong dữ liệu chưa được gán nhãn chính là việc tự học hay còn gọi là tự huấn luyện, tự gán nhãn. Bên cạnh đó, để gán nhãn cho dữ liệu trong bộ huấn luyện cần nhiều thời gian, công sức và còn có thể có sai sót. Với bộ dữ liệu huấn luyện, những lỗi đó có thể gây ảnh hưởng tới hiệu suất phân lớp. Vì vậy việc sử dụng dữ liệu chưa gán nhãn kết hợp cùng dữ liệu đã gán nhãn trong học máy bán giám sát giúp khắc phục được những hạn chế phát sinh đó.

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

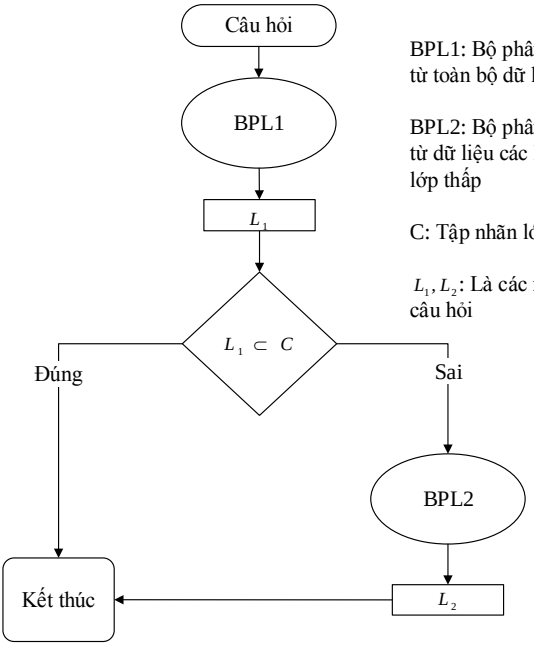
# Chương 3: Đề xuất cải tiến mô hình phân lớp

## 3.1 Thực trạng

Trên thực tế, phân lớp đạt kết quả tốt cần phải dựa vào nhiều yếu tố khác nhau. Và một yếu tố quan trọng trong đó là chất lượng và số lượng của các nhãn lớp khác nhau.

## 3.2 Mô hình đề xuất

Dựa trên 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. Với mô hình này, các lớp có độ chính xác cao sẽ được tách ra làm 1 cấp và các nhãn lớp còn lại sẽ được đưa vào 1 cấp.



*Hình 3.1: Mô hình phân lớp đề xuất*

Việc xây dựng bộ phân lớp thứ 2 theo như mô hình đề xuất được thực hiện như sau: Loại bỏ toàn bộ câu hỏi trong bộ dữ liệu huấn luyện đã được gán nhãn lớp mà nhãn lớp đó nằm trong tập nhãn lớp có độ chính xác cao đề xuất.

## 3.3 Mô hình xử lý dữ liệu

### 3.3.1 Thu thập dữ liệu

Dữ liệu cần sử dụng để xây dựng bộ huấn luyện có thể được khai thác và tận dụng từ nhiều nguồn khác nhau. Điều này phụ thuộc vào nhu cầu cũng như mục đích cần đạt tới của hệ thống sử dụng. Ngoài ra, số lượng câu hỏi cần thu thập từ các nguồn cũng nên đạt tới một ngưỡng nào đó. Nếu số lượng câu hỏi quá ít và chất lượng kém sẽ không đủ để tạo ra một bộ dữ liệu huấn luyện tốt.

### 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 các nguồn cung cấp. Vì trong các nguồn dữ liệu ta không thể biết trước được có bao nhiêu dữ liệu bị trùng lặp, vô nghĩa cần phải loại bỏ.

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

Từ tập dữ liệu gồm những câu hỏi có ích sẽ giúp tạo ra 1 bộ phân lớp câu hỏi theo những đặc trưng của các câu hỏi đó. Công việc này mất khá nhiều thời gian vì phải duyệt nhiều lần qua toàn bộ các câu hỏi một cách cẩn thận để tìm ra những đặc trưng riêng biệt. Sau khi đã có được bộ phân lớp với một

số lượng lớp nhất định rồi, bước tiếp theo là gán nhãn lớp dựa theo bộ phân lớp cho mỗi câu hỏi.

#### 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 có hệ thống từ loại rất đa dạng và phức tạp. Có thể kể đến như danh từ, động từ, tính từ, số từ, lượng từ, phó từ, thán từ… Trong một số từ loại này lại có các nhóm từ loại nhỏ liên quan. Bên cạnh đó chúng ta cũng cần chú ý về cấu trúc của từ như từ đơn, từ ghép, từ láy.... Các từ đứng cạnh nhau nhưng có thể có nghĩa riêng và khi ghép vào thì lại mang một nghĩa khác.

#### 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. 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. Các lớp nhãn, từ loại được lưu trong các tập hợp phải đảm bảo không trùng nhau, mỗi giá trị chỉ đại diện cho duy nhất 1 lớp nhãn, từ loại.

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

Trong bước xử lý này, trước hết cần phải xác định được nhãn lớp nào có độ phân lớp chính xác cao. Ta áp dụng giải thuật tham lam (Greedy Algorithms) trong việc tìm kiếm các nhãn lớp yêu cầu. Đây là giải thuật có thiết kế đơn giản và được sử dụng để lựa chọn tối ưu cục bộ với hy vọng sẽ chọn được tối ưu toàn cục.

# Chương 4: Ứng dụng vào hệ thống giải đáp thắc mắc tại Trung tâm đào tạo trực tuyến Elearning

## 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 đã 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 hơn số lượng lớn học viên đăng ký theo học tại nhiều đơn vị liên kết trên cả nước.

## 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. 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.

## 4.3 Chuẩn bị dữ liệu thực nghiệm

### 4.3.1 Thu thập dữ liệu

Sau khi áp dụng các phương pháp chạy crawler thì kết quả đưa ra được là một tập gồm hơn 4000 câu hỏi ở dạng thô

chưa xử lý. Ở giai đoạn tiếp theo, các câu hỏi sẽ được xử lý bằng một số công cụ đã có sẵn và một số công cụ tự viết theo mục đích sử dụng.

### 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, số lượng còn lại chính xác là 1509 câu hỏi.

#### 4.3.2.2 Xây dựng bộ phân lớp và gán nhãn lớp câu hỏi

Từ tập dữ liệu đã xử lý thô, ta tiến hành 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 thì 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.

Cuối cùng ta 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 đã được gán nhãn, tiếp theo ta tiến hành chuẩn hóa các đặc trưng trong câu hỏi bằng công cụ VnTagger.

#### 4.3.2.4 Định dạng dữ liệu theo chuẩn 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.

#### 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 huấn luyện tại trung tâm E-Learning, có 9 nhãn lớp có độ phân lớp chính xác cao. Các lớp này sẽ bị loại bỏ ra khỏi tập dữ liệu huấn luyện để xây dựng bộ phân lớp thứ 2. Số lượng câu hỏi trong tập dữ liệu huấn luyện sau khi loại bỏ câu hỏi của 9 nhãn lớp còn 842 câu hỏi.

### 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.

## 4.4 Kết quả thực nghiệm

### 4.4.1 Cài đặt môi trường thực nghiệm

Thực nghiệm được tiến hành trên máy chủ Linux có cấu hình được trình bày trong Bảng 4.2

*Bảng 4.2. 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

Với bộ dữ liệu của Li và Roth, thực nghiệm với 5952 câu hỏi, trong đó có 5452 câu hỏi huấn luyện và 500 câu hỏi kiểm tra. Áp dụng mô hình phân lớp đề xuất, loại bỏ lớp có độ chính xác cao đề xuất đã nêu ở các chương trước. Kết quả chi tiết được trình bày trong bảng dưới đây.

*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** |

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

Kết quả này thực hiện dựa trên bộ phân lớp gồm 22 lớp có liên quan đến các vấn đề trong trung tâm đào tạo trực tuyến. Chi tiết được thể hiện trong bảng dưới đây.

*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** |

**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.