TẠP CHÍ KHOA HỌC
Khoa học Tự nhiên và Công nghệ, Số 10 (9/2017) tr 104 - 113

TÌM KIẾM CÂU HỎI CÓ Ý NGHĨA TỪ CÁC TRANG WEB HỎI ĐÁP CỘNG ĐỒNG

Nguyễn Văn Tú, Hoàng Thị Lam, Nguyễn Thị Thanh Hà
Trường Đại học Tây Bắc

Tóm tắt: *Các trang web hỏi đáp cộng đồng có chứa một lượng lớn thông tin hỏi - đáp có giá trị sinh ra bởi những người sử dụng. Trong các trang web hỏi đáp cộng đồng, người dùng có thể gửi các câu hỏi, trả lời các câu hỏi của người khác và cung cấp thông tin phản hồi cho những câu hỏi/câu trả lời. Trong nghiên cứu này chúng tôi sử dụng tiếp cận học máy nhằm tìm kiếm những câu hỏi có ý nghĩa từ các trang web hỏi đáp cộng đồng. Những câu hỏi này cùng với các câu trả lời tốt nhất của chúng sẽ là nguồn dữ liệu quan trọng cho xây dựng cơ sở tri thức, truy hồi thông tin và hỏi đáp non - factiod. Nghiên cứu thực hiện trích rút những đặc trưng quan trọng từ mỗi luồng hỏi - đáp cũng như thông tin của người sử dụng và xây dựng mô hình phân loại để xác định được những câu hỏi có ý nghĩa. Các thực nghiệm được xây dựng trên tập dữ liệu trích rút từ trang web hỏi đáp cộng đồng Yahoo! Answer đã chỉ ra tính hiệu quả của phương pháp đề xuất của chúng tôi.*

Từ khóa: *Hỏi đáp cộng đồng, hỏi đáp tự động, phân loại, support vector machine.*

1. Tổng quan

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và truy xuất thông tin, vấn đề hỏi - đáp đã thu hút nhiều sự chú ý trong những năm qua. Tuy nhiên, các nghiên cứu về hỏi - đáp chủ yếu tập trung vào việc tìm câu trả lời chính xác cho câu hỏi *factoid* được trích rút từ các tài liệu liên quan. Đánh giá nổi tiếng nhất về nhiệm vụ hỏi - đáp *factoid* là hội nghị truy hồi văn bản (Text REtrieval Conference-TREC(1)). Các câu hỏi và câu trả lời được phát hành bởi TREC đã trở thành nguồn dữ liệu quan trọng cho các nhà nghiên cứu trong việc nghiên cứu xây dựng các hệ thống hỏi đáp tự động [10]. Tuy nhiên, khi phải đối mặt với các câu hỏi *non - factoid* như các câu hỏi về lý do tại sao, như thế nào,… hầu như không có hệ thống hỏi đáp tự động nào làm việc hiệu quả. Đồng thời chi phí của việc xây dựng dữ liệu huấn luyện với các câu hỏi *non - factoid* là quá cao. Trong các hệ thống này, chi phí của việc xây dựng dữ liệu lớn nhưng sau đó nó chỉ làm việc trong một miền cụ thể. Do đó việc trích rút các tri thức từ nguồn tài nguyên mở (thường là từ Internet) là một hướng tiếp cận thích hợp cho xây dựng các hệ thống hỏi đáp tự động miền mở.

Các trang web hỏi đáp cộng đồng (community Question Answering - cQA) có chứa một lượng lớn nguồn tài nguyên phong phú là thông tin hỏi - đáp sinh ra bởi những người sử dụng. Trích rút các cặp câu hỏi-câu trả lời có ý nghĩa là nhiệm vụ quan trọng cho xây dựng cơ sở tri thức, truy hồi thông tin và hỏi đáp *non-factiod*. Tuy nhiên các cặp câu hỏi-câu trả lời trong các cQA này thường rất nhiễu, cả trong câu hỏi và các câu trả lời.

Do đó trước khi sử dụng, các cặp câu hỏi - câu trả lời này cần phải được phân loại. Việc phân loại này giúp người dùng xây dựng được kho dữ liệu là các cặp câu hỏi - câu trả lời có ý nghĩa. Nhóm tác giả đã tìm hiểu về các nghiên cứu liên quan và thấy rằng các nghiên cứu

Ngày nhận bài: 22/01/2017. Ngày nhận kết quả phản biện: 10/03/2017 Ngày nhận đăng: 20/9/2017

Liên lạc: Nguyễn Văn Tú, e - mail: tuspttb@gmail.com

(1)http://trec.nist.gov/

trước đây thường chỉ tập trung vào việc tìm kiếm câu trả lời tốt nhất cho mỗi câu hỏi [2, 6, 9]. Tuy nhiện, nghiên cứu này sử dụng tiếp cận học máy nhằm tìm kiếm các câu hỏi có ý nghĩa từ các dữ liệu hỏi đáp thu thập từ các cQA. Để tìm kiếm được các câu hỏi có ý nghĩa từ các cQA, bài báo này đề xuất sử dụng sự kết hợp của nhiều loại đặc trưng quan trọng trích rút từ mỗi luồng hỏi đáp cũng như thông tin của người gửi câu hỏi/câu trả lời và xây dựng mô hình phân loại để xác định được các câu hỏi có ý nghĩa.

Bài báo này được tổ chức như sau: Phần 2 trình bày về tiếp cận của nhóm tác giả trong việc giải quyết bài toán; Phần 3 trình bày về các đặc trưng được sử dụng; Phần 4 là các thực nghiệm và các kết quả nghiên cứu; Kết luận và hướng phát triển được trình bày trong Phần 5.

2. Cách tiếp cận của chúng tôi về tìm kiếm câu hỏi có ý nghĩa

2.1. Bài toán tìm kiếm câu hỏi có ý nghĩa

Việc tìm kiếm những câu hỏi có ý nghĩa từ các cQA là nhằm xây dựng được các câu hỏi có ý nghĩa trong một tập rất lớn các luồng hỏi - đáp. Vì vậy, nghiên cứu này coi vấn đề tìm kiếm các câu hỏi có ý nghĩa từ các cQA như là một vấn đề phân loại nhị phân và được phát biểu như sau:

Cho một tập Q = {q_1, q_2, …, q_n} các câu hỏi trích rút từ các luồng hỏi - đáp trên các cQA. Việc phân loại các câu hỏi trong Q chính là gán nhãn cho các câu hỏi {q_1, q_2, …, q_n} các nhãn tương ứng là {l_1, l_2, …, l_n} trong đó l_i = "Good" nếu q_i là câu hỏi có ý nghĩa, l_i = "Bad" nếu q_i không phải là một câu hỏi có ý nghĩa.

2.2. Thuật toán phân loại câu hỏi

Có nhiều bộ phân loại khác nhau đã được sử dụng để phân loại các dữ liệu văn bản như: *Support Vector Machine, Naive Bayes, Maximum Entropy Models, Sparse Network of Winnows,…* Tuy nhiên trong các bộ phân loại đó thì Support Vector Machine được xem là hiệu quả hơn cả [1, 3, 4]. Trong vấn đề phân loại câu hỏi, mỗi câu hỏi được coi như là một văn bản ngắn và được biểu diễn trong mô hình không gian vecto có số chiều rất lớn, điều này có thể được phân loại tốt bởi *Support Vector Machine*.

Tuy nhiên, dữ liệu trong nghiên cứu này là các dữ liệu thu thập từ các cQA, các dữ liệu chưa được gán nhãn. Vì vậy, thuật toán phân loại bán giám sát SVM (*semi-supervised Support Vector Machines*) đã được sử dụng, đây là thuật toán phân loại bán giám sát được xây dựng từ thuật toán phân loại có giám sát *Support Vector Machine*.

Trong *semi-supervised Support Vector Machines*, tập dữ liệu huấn luyện gồm hai phần, *x* là phần dữ liệu huấn luyện gồm *n* dữ liệu đã được gán nhãn là y_i $(y_i \in \{+1, -1\})$ và *x** là phần dữ liệu huấn luyện gồm *k* dữ liệu chưa được gán nhãn y_j^* $(y_j^* \in \{+1, -1\})$. Mục đích của thuật toán *semi-supervised Support Vector Machines* là gán nhãn cho *k* dữ liệu chưa gán nhãn này. Mỗi dữ liệu được xem như một điểm trong không gian vecto. Mỗi điểm i thuộc phần dữ liệu đã gán nhãn có một sai số là ξ_i và mỗi điểm j thuộc phần dữ liệu chưa gán nhãn sẽ có sai số là

ξ_j^*. Thuật toán *semi-supervised Support Vector Machines* sẽ giải bài toán tối ưu sau thay cho bài toán tối ưu ở thuật toán SVM.

$$\underset{w,\xi,\xi^*}{argmin} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=n+1}^{n+k}\xi_j^* \quad (1)$$

Thỏa mãn các điều kiện:

$$y_i(w.x_i - b) \geq 1 - \xi_i$$
$$y_j^*(w.x_i - b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, \forall i \in \{1,..., n\}$$
$$\xi_j^* \geq 0, \forall j \in \{n+1,...,n+k\}$$

Trong đó C và C^* là các tham số phạt của các dữ liệu đã gán nhãn và các dữ liệu chưa gán nhãn.

Nghiên cứu này sử dụng thuật toán *semi-supervised Support Vector Machines* đã được cài đặt bởi Vikas Sindhwani và các tác giả trình bày trong [7, 8].(2)

2.3. Hiệu suất trong phân loại câu hỏi

Để đánh giá hiệu suất của việc phân loại các câu hỏi, chúng tôi sử dụng các độ đo *precision, recall, F_1-measure, accuracy* được định nghĩa như dưới đây. Để ước lượng các độ đo này có thể dựa vào Bảng 1.

Bảng 1. Các kết quả dự đoán của phân loại

	Label $y^* = +1$	Label $y^* = -1$
Prediction $f(x^*) = +1$	*TP*	*FP*
Prediction $f(x^*) = -1$	*FN*	*TN*

Mỗi ô trong bảng đại diện cho một trong bốn kết quả đầu ra có thể của một dự đoán f(x*). Trong đó:

TP (True Positive): Số lượng các câu hỏi *positive* được phân loại đúng.

TN (True Negative): Số lượng các câu hỏi *negative* được phân loại đúng.

FP (False Positive): Số lượng các câu hỏi *positive* bị phân loại sai.

FN (False Negative): Số lượng các câu hỏi *negative* bị phân loại sai.

Precision được định nghĩa như là xác suất mà một dữ liệu phân loại là $f(x^*) = +1$ là một phân loại đúng. Nó được ước lượng như sau:

(2)http://vikas.sindhwani.org/svmlin.html

$$Precision\ \ p = \frac{TP}{Tp + FP} \tag{2}$$

Recall được định nghĩa như là xác suất mà một dữ liệu với nhãn là $y^* = +1$ đã được phân loại đúng, được ước lượng như sau:

$$Recall\ \ r = \frac{TP}{TP + FN} \tag{3}$$

$$\underset{F\ \ measure}{\underset{1p + r}{-}} = \underline{2^*}\ \underline{p^*r} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + Tn + FN} \tag{5}$$

3. Các đặc trạng trong phân loại câu hỏi

Trong phân loại câu hỏi trích rút từ các cQA, tập các đặc trưng khác nhau có thể được sử dụng. Các đặc trưng trong phân loại câu hỏi có thể được phân thành 3 loại khác nhau: Các đặc trưng về từ vựng, các đặc trưng về cú pháp và các đặc trưng dựa trên thông tin người dùng (người gửi câu hỏi/câu trả lời, người đánh giá câu trả lời). Mỗi câu hỏi trong các cQA gồm 2 phần: Chủ đề hỏi (QSubject) và phần mô tả ngắn nội dung câu hỏi (QContent). Phần chủ đề hỏi và phần mô tả nội dung câu hỏi được coi như là một câu hỏi duy nhất gồm nhiều câu [5]. Bảng 2 là một ví dụ về câu hỏi trong cQA Yahoo!Answer.

Bảng 2. Ví dụ về câu hỏi trong trang web cQA Yahoo!Answer

Chủ đề hỏi hỏi	Nội dung câu hỏi
Body massage	Hi there, i can see a lot of massage center here, but i dont know which one is better. can someone help me which massage center is good... and how much will it cost me? thanks

3.1. Các đặc trưng từ vựng

Đặc trưng n-gram: Các đặc trưng *n-gram* của một câu hỏi thường được trích rút dựa trên ngữ cảnh của các từ của câu hỏi, nghĩa là, các từ đó xuất hiện trong một câu hỏi. Trong phân loại câu hỏi, một câu hỏi được biểu diễn giống như sự biểu diễn tài liệu trong mô hình không gian vecto. Một câu hỏi x có thể được biểu diễn như sau:

$$x = x_1, x_2, \ldots, x_N \tag{6}$$

trong đó x_i là tần số xuất hiện của từ i trong câu hỏi x và N là tổng số các từ trong x. Do tính thưa thớt của các đặc trưng, chỉ các đặc trưng có giá trị khác không mới được giữ lại trong véc tơ đặc trưng. Bởi vậy các câu hỏi cũng được biểu diễn dưới hình thức sau:

$$x = \{(t_1, f_1), \ldots, (t_p, f_p)\} \tag{7}$$

trong đó t_i là từ thứ i trong câu hỏi x và f_i là tần số xuất hiện của t_i trong câu hỏi x. Không gian đặc trưng này được gọi là các đặc trưng *bag-of-words* hoặc là *unigrams*. *Unigrams* là một trường hợp đặc biệt của cách gọi các đặc trưng *n-gram*. Để trích rút các đặc trưng *n-gram*, bất

kì n từ nào liên tiếp trong một câu hỏi đều được coi là một đặc trưng. Bảng 3 là danh sách các đặc trưng *n-gram* của câu hỏi "How many Grammys did Michael Jackson win in 1983?".

Bảng 3. Ví dụ về các đặc trọng từ vựng

Tên đặc trưng	Đặc trưng
Unigram	{(How, 1) (many, 1) (Grammys, 1) (did, 1) (Michael, 1) (Jackson, 1) (win, 1) (in, 1) (1983, 1) (?, 1)}
Bigram	{(How-many, 1) (many-Grammys, 1) (Grammys-did, 1) (did-Michael, 1) (Michael-Jackson, 1) …(1983-?, 1)}
Trigram	{(How-many-Grammys, 1) (many-Grammys-did, 1) …(in-1983-?, 1)}

Số từ trong câu hỏi: Trong phân loại câu hỏi, nghiên cứu cũng sử dụng số từ của mỗi câu hỏi như là một đặc trưng để phân loại. Dữ liệu thực tế trong các cQA cho thấy rằng những câu hỏi chứa ít hơn 5 từ thường là những câu hỏi nhiễu, không mang ý nghĩa.

Số câu (sentence) trong câu hỏi: Mỗi câu hỏi trong cQA thường là một đoạn văn bản ngắn gồm nhiều câu trình bày những thông tin mà người gửi câu hỏi muốn hỏi. Nghiên cứu này sử dụng số câu trong mỗi câu hỏi như là một đặc trưng dùng để phân loại. Những câu hỏi gồm nhiều câu thường mang thông tin đầy đủ của câu hỏi, tuy nhiên nó lại chứa nhiều thông tin nhiễu khó phân loại.

Số lượng stop word trong câu hỏi: Stop word là những từ quá phổ biến, quá chung chung (ví dụ một số *stop word* trong tiếng Anh: The, a, and, …). Mỗi câu hỏi của người dùng thường chứa các từ *stop word*. Số lượng các từ stop word trong mỗi câu hỏi được sử dụng là một đặc trưng dùng để phân loại câu hỏi.

Từ để hỏi: Nghiên cứu sử dụng từ để hỏi (từ để hỏi là một trong các từ: What, Who, When, Why, Where, How) như là một đặc trưng dùng để phân loại câu hỏi.

3.2. Các đặc trưng cú pháp

Các đặc trưng cú pháp có thể được trích rút dựa vào việc phân tích cấu trúc cú pháp của câu hỏi.

Tagged Unigrams: Tagged Unigrams cho biết thẻ từ loại của mỗi từ trong câu hỏi như NN (Noun - danh từ), NP (Noun Phrase - cụm danh từ), VP (Verb Phrase - cụm động từ), JJ (adjective - tính từ), … Ví dụ câu hỏi sau "How many Grammys did Michael Jackson win in 1983?" với *Tagged Unigrams* của nó:

How_WRB many_JJ Grammys_NNPS did_VBD Michael_NNP Jackson_NNP win_VBP in_IN 1983_CD?_.

Số thực thể trong câu hỏi: Số lượng thực thể trong mỗi câu hỏi được sử dụng như là một đặc trưng cú pháp để phân loại câu hỏi.

Số danh từ trong câu hỏi: Nghiên cứu sử dụng số lượng các danh từ trong mỗi câu hỏi như là một đặc trưng cú pháp dùng để phân loại câu hỏi.

3.3. Các đặc trưng dựa trên thông tin người dùng

Số câu hỏi của người hỏi: Trên mỗi cQA, mỗi người có thể gửi nhiều câu hỏi mà họ cần sự giải đáp của cộng đồng người sử dụng cQA. Những người thường xuyên gửi câu hỏi thì các câu hỏi của họ thường mang nhiều ý nghĩa hơn cho cộng đồng. Do đó, nghiên cứu này sử dụng số lượng câu hỏi của người hỏi như là một đặc trưng trong việc giải quyết bài toán phân loại câu hỏi.

Số câu trả lời của người hỏi: Những người thường xuyên tham gia hỏi đáp trên cQA thì thường những câu hỏi cũng như câu trả lời của họ sẽ mang nhiều ý nghĩa hơn. Vì vậy, nghiên cứu này đề xuất sử dụng số lượng câu trả lời của người hỏi như là một đặc trưng dùng để phân loại câu hỏi.

Số câu trả lời của câu hỏi: Mỗi câu hỏi khi được đăng lên cQA có thể nhận được nhiều câu trả lời của những người quan tâm. Số lượng câu trả lời cho câu hỏi này thể hiện sự quan tâm của người dùng đối với các vấn đề mà câu hỏi đang đề cập tới. Vì vậy, nghiên cứu sử dụng số lượng câu trả lời của mỗi câu hỏi như là một đặc trưng nhằm đánh giá tầm quan trọng của câu hỏi.

Số lượng đánh giá cho câu trả lời tốt nhất của câu hỏi: Khi một câu hỏi được đăng lên có thể nhận được nhiều câu trả lời của người dùng. Người dùng cũng có quyền đánh giá câu trả lời nào đó là câu trả lời tốt nhất. Một câu hỏi mà câu trả lời của nó nhận được nhiều đánh giá thì thường là một câu hỏi có nhiều ý nghĩa. Vì vậy, nghiên cứu sử dụng số lượng đánh giá của câu trả lời như là một đặc trưng để phân loại câu hỏi của nó.

4. Các thực nghiệm và kết quả

4.1. Tập dữ liệu

Để thực hiện các thực nghiệm của mình, nghiên cứu đã thu thập dữ liệu từ trang web hỏi đáp cộng đồng Yahoo!Answer[(3)]. Tập dữ liệu này chứa 3.000 câu hỏi cùng với 9.895 câu trả lời và các thông tin liên quan đến câu hỏi/câu trả lời và thông tin về người gửi câu hỏi - câu trả lời. Tất cả các cặp câu hỏi-câu trả lời đều được trình bày bằng ngôn ngữ tiếng Anh.

4.2. Các thực nghiệm và kết quả

Để cài đặt các thực nghiệm của mình, nghiên cứu chia tập dữ liệu thành hai tập con *test* và *training*. Tập *test* gồm 300 câu hỏi được sử dụng làm tập dữ liệu kiểm tra, tập *training* gồm 2700 câu hỏi được sử dụng làm tập dữ liệu huấn luyện.

4.2.1. Thực nghiệm 1

Thực nghiệm 1 chỉ sử dụng các đặc trưng từ vựng trích rút từ các câu hỏi trong tập dữ liệu. Các đặc trưng này gồm: Đặc trưng *Unigram, Bigram*, số từ trong câu hỏi, số *sentence* trong câu hỏi, số lượng *stop word* trong câu hỏi, từ để hỏi. Các kết quả của thực nghiệm 1 được trình bày trong Bảng 4.

(3)https://answers.yahoo.com/

Bảng 4. Kết quả phân loại khi sử dụng các đặc trưng từ vựng

Tên đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F_1-measure (%)
Unigram	54,67	57,05	56,33	56,69
Bigram	46,00	48,75	49,37	49,06
Số từ trong câu hỏi	50,00	52,25	58,86	55,36
Số câu trong câu hỏi	46,57	49,45	56,96	52,94
Số lượng *stop word*	49,67	52,02	56,96	54,38
Từ để hỏi	47,33	22,95	50,00	34,17
Tất cả đặc trưng trên	55,33	61,39	57,06	59,14

4.2.2. Thực nghiệm 2

Thực nghiệm 2 chỉ sử dụng các đặc trưng cú pháp trích rút từ các câu hỏi trong tập dữ liệu. Các đặc trưng này gồm: Đặc trưng *Tagged Unigrams*, số thực thể trong câu hỏi, số danh từ trong câu hỏi. Các kết quả của thực nghiệm 2 được trình bày trong Bảng 5.

Bảng 5. Kết quả phân loại khi sử dụng các đặc trưng cú pháp

Tên đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F_1-measure (%)
Tagged Unigrams	53,00	55,63	53,16	54,37
Số thực thể trong câu hỏi	50,33	52,26	65,82	58,26
Số danh từ trong câu hỏi	50,00	52,33	56,96	54,55
Tất cả đặc trưng trên	55,67	58,74	53,16	55,81

4.2.3. Thực nghiệm 3

Thực nghiệm 3 chúng tôi muốn kiểm tra tính hiệu quả của các đặc trưng tính toán dựa vào thông tin của người dùng. Các đặc trưng này gồm: số câu hỏi của người hỏi, số câu trả lời của người hỏi, số câu trả lời của câu hỏi, số lượng đánh giá cho câu trả lời tốt nhất của câu hỏi. Các kết quả của thực nghiệm 3 được cho trong Bảng 6.

Bảng 6. Kết quả phân loại khi sử dụng các đặc trưng dựa trên thông tin người dùng

Tên đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F_1-measure (%)
Số câu hỏi của người hỏi	71,00	72,33	72,78	72,55
~~Số câu trả lời của người hỏi~~	~~49,67~~	~~68,42~~	~~8,23~~	~~14,69~~
~~Số câu trả lời của câu hỏi~~	~~79,00~~	~~96,12~~	~~62,66~~	~~75,86~~
Số lượng đánh giá cho câu trả lời tốt nhất của câu hỏi	89,67	91,13	89,44	90,28
Tất cả đặc trưng trên	90,67	92,41	90,12	91,25

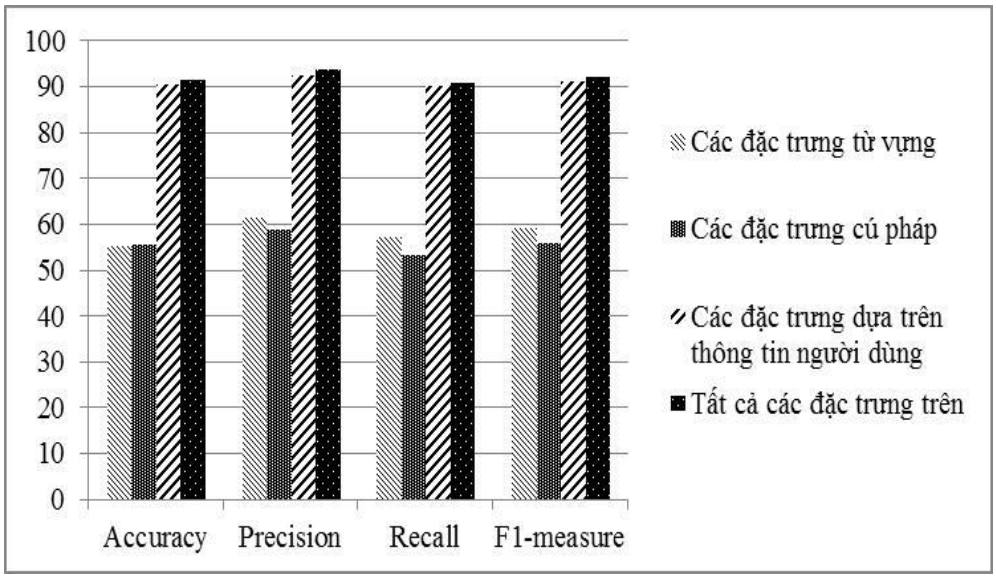
4.2.4. Thực nghiệm 4

Thực nghiệm 4 kiểm tra sự đóng góp của tất cả các loại đặc trưng: Đặc trưng từ vựng, đặc trưng cú pháp và đặc trưng dựa trên thông tin người dùng. Chính vì vậy nghiên cứu kết hợp tất cả các loại đặc trưng này trong việc phân loại câu hỏi. Các kết quả của thực nghiệm 4 được trình bày trong Bảng 7.

Bảng 7. Kết quả phân loại khi sử dụng sự kết hợp của nhiều loại đặc trọng

Tên đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F_1-measure (%)
Các đặc trưng từ vựng	55,33	61,39	57,06	59,14
Các đặc trưng cú pháp	55,67	58,74	53,16	55,81
Các đặc trưng dựa trên thông tin người dùng	90,67	92,41	90,12	91,25
Tất cả đặc trưng trên	91,67	93,67	90,79	92,21

Từ các kết quả của các thực nghiệm trên đã cho thấy vai trò của việc sử dụng mỗi loại đặc trưng khác nhau trong phân loại câu hỏi. Kết quả của các thực nghiệm 1, 2, 3 cho thấy việc phân loại câu hỏi cần sự kết hợp của nhiều loại đặc trưng khác nhau. Đồng thời trong các thực nghiệm này cũng cho thấy các đặc trưng trích rút từ thông tin người sử dụng sẽ cho kết quả phân loại cao nhất. Thực nghiệm 4 là sự tổng hợp của các thực nghiệm 1, 2, 3 khi sử dụng kết hợp tất cả các đặc trưng từ vựng, cú pháp và đặc trưng trích rút từ thông tin người dùng. Kết quả của thực nghiệm này cho thấy sự kết hợp của các loại đặc trưng khác nhau mà nghiên cứu sử dụng sẽ cho kết quả phân loại cao nhất. Biểu đồ 1 so sánh tầm quan trọng của các loại đặc trưng khác nhau trong phân loại câu hỏi trong các cQA.



Biểu đồ 1. So sánh vai trò của các loại đặc trọng trong phân loại câu hỏi

5. Kết luận và hướng phát triển

Nghiên cứu đã trình bày một tiếp cận dựa trên học máy để phân loại câu hỏi trong các cQA. Để huấn luyện một thuật toán học, nhóm tác giả đã trích rút một tập các đặc trưng tốt từ từ vựng, cú pháp và đặc trưng dựa trên thông tin người dùng. Các kết quả thực nghiệm đã chứng minh rằng các đề xuất sẽ mạng lại hiệu quả cao trong vấn đề tìm kiếm các câu hỏi có ý nghĩa trong các cQA. Việc kết hợp của tất cả các loại đặc trưng từ vựng, cú pháp và đặc trưng dựa trên thông tin người dùng sẽ cho hiệu suất phân loại cao nhất. Nghiên cứu tiếp theo sẽ tìm ra phương pháp để kết hợp các đặc trưng tốt hơn nhằm đạt được sự phân loại có độ chính xác cao hơn.

TÀI LIỆU THAM KHẢO

[1] Zhiheng Huang, Marcus Thint, and Zengchang Qin (2008). Question classification using head words and their hypernyms, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08), pp. 927-936.

[2] H. Hu, B. Liu, B. Wang, M. Liu, X. Wang (2013). Multimodal DBN for predicting high-quality answers in cQA portals, In Proceedings of ACL, pp. 843-847.

[3] Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti (2005). Enhanced answer type inference from questions using sequential models, In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pp. 315-322.

[4] Babak Loni, Gijs van Tulder, Pascal Wiggers, David M. J. Tax, and Marco Loog (2011). Question classification with weighted combination of lexical, syntactical and semantic features, In Proceedings of the 15th international conference of Text, Dialog and Speech, pp. 243-250.

[5] Vinay Pande, Tanmoy Mukherjee, Vasudeva Varma (2013). Summarizing Answers For Community Question Answer Services, The International Conference of the German Society for Computational Linguistics and Language Technology, pp. 151-161.

[6] C. Shah, J. Pomerantz (2010). Evaluating and predicting answer quality in community QA, In Proceedings of SIGIR.

[7] Vikas Sindhwani, S. Sathiya Keerthi (2007). Newton Methods for FastSolution of Semisupervised Linear SVMs, Large Scale Kernel Machines, MIT Press, pp. 155-174.

[8] Vikas Sindhwani, S. Sathiya Keerthi (2006). Large Scale Semi-supervised Linear SVMs, SIGIR'06, pp. 477-484.

[9] H. Toba, Z. Y. Ming, M. Adriani, T. Chua (2014). Discovering high quality answers in community question answering archives using a hierarchy of classifiers, Information Sciences, 261: 101-115.

[10] Zeyi Wen, Rui Zhang, Kotagiri Ramamohanarao (2014). Enabling Precision/Recall Preferences for Semi-supervised SVM Training, CIKM'14, pp. 421-430.

FINDING MEANINGFUL QUESTIONS FROM COMMUNITY QUESTION ANSWERING WEBSITES

Nguyen Van Tu, Hoang Thi Lam, Nguyen Thi Thanh Ha ***Tay Bac University***

Abstract*: The community question answering websites contains a large amount of question - answer information valuable generated by the users. In the community question answering websites, users can post questions, answer other people's questions, and provide feedbacks to the questions / answers. In this study, we use machine learning approach in order to find meaningful questions from the community question answering websites. These questions, along with their best answers will be an important data source for building the knowledge base, information retrieval and non-factiod question answering. The extracted important features from each question research - answer thread as well as information of the user and built a classification model to identify meaningful questions. The experiments are established on dataset extracted from the community question answering websites Yahoo! The answers have shown the effectiveness of the method proposed by us.*

Keywords*: Automated question answering, community question answering, classification, support vector machine.*