PHÂN LOẠI VĂN BẢN DỰA TRÊN RÚT TRÍCH TỰ ĐỘNG TÓM TẮT CỦA VĂN BẢN

Trương Quốc Định [1]

[1] Khoa Công nghệ thông tin & Truyền thông, Trường Đại học Cần Thơ

tqdinh@cit.ctu.edu.vn

TÓM TẮT— *Phân loại văn bản là tiến trình đưa các văn bản chưa biết chủ đề vào các lớp văn bản đã biết chủ đề. Các chủ đề này được xác định bởi một tập các tài liệu mẫu. Để thực hiện quá trình phân loại, một giải thuật máy học được sử dụng để xây dựng bộ phân loại từ tập huấn luyện bao gồm nhiều văn bản (thường là toàn bộ nội dung văn bản), sau đó dùng bộ phân loại này để dự đoán lớp của những tài liệu mới. Nhiều kỹ thuật máy học và khai phá dữ liệu đã được áp dụng vào bài toán phân loại văn bản như: Bayes ngây thơ, cây quyết định, k-láng giềng gần nhất, mạng nơron, máy học vectơ hỗ trợ … Trong bài báo này, chúng tôi đề xuất mô hình phân loại văn bản dựa vào tóm tắt của văn bản được rút trích một cách tự động và sử dụng giải pháp máy học để thực hiện phân loại. Mô hình đề xuất được chúng tôi thực nghiệm trên một tập 2000 các tài liệu văn bản tiếng Việt là các bài viết được tải xuống từ các trang báo điện tử vnexpress.net, vietnamnet.vn. Các kết quả thực nghiệm bước đầu đã khẳng định tính khả thi của mô hình mà chúng tôi đề xuất đồng thời gợi mở một hướng nghiên cứu khả thi cho bài toán phân loại văn bản.*

Từ khóa— *Phân loại văn bản, tóm tắt tự động văn bản, máy học.*

I. GIỚI THIỆU

Phân loại văn bản (text classification) là bài toán cơ bản nhất của lĩnh vực khai phá dữ liệu văn bản. Phân loại văn bản chính là gán nhãn (lớp/chủ đề) một cách tự động dựa vào nội dung của văn bản. Phân loại văn bản được ứng dụng trong nhiều lĩnh vực như tìm kiếm thông tin, lọc văn bản, tổng hợp tin tức tự động, thư viện điện tử. Phân lớp văn bản có thể thực hiện thủ công hoặc tự động bằng cách sử dụng các kỹ thuật như máy học vectơ hỗ trợ (support vector machines - SVM) [1], cách tiếp cận sử dụng lý thuyết tập thô [2], cách tiếp cận theo luật kết hợp [3]. Dù là tiếp cận theo hướng nào đi nữa thì các kỹ thuật vừa nêu cũng sử dụng toàn văn nội dung của văn bản để thực hiện phân lớp, điều này đồng nghĩa với việc các bài toán phân lớp luôn phải đối phó với một lượng lớn các đặc trưng. Khi mà trong một số lĩnh vực như tìm kiếm thông tin, thư viện điện tử, tổng hợp tin tức tự động có số lượng văn bản gia tăng đáng kể hàng ngày thì việc phải thực hiện phân loại từng văn bản với nội dung cực lớn (sách trong thư viện chẳng hạn) cũng là một thách thức không nhỏ.

Chúng tôi cảm thấy rằng theo thói quen, trước khi quyết định đọc một quyển sách, một bài báo khoa học … độc giả luôn đọc tóm tắt của tài liệu này để đi đến quyết định có mua hoặc đọc tiếp hay không. Điều này chứng tỏ rằng tóm tắt thể hiện được nội dung cốt lõi, mấu chốt của tài liệu và có thể là chủ đề của tài liệu. Một ý tưởng rằng nếu có được một tóm tắt tốt thì việc xác định chủ đề của văn bản, hay nói cách khác là việc phân lớp văn bản có thể dựa vào tóm tắt.

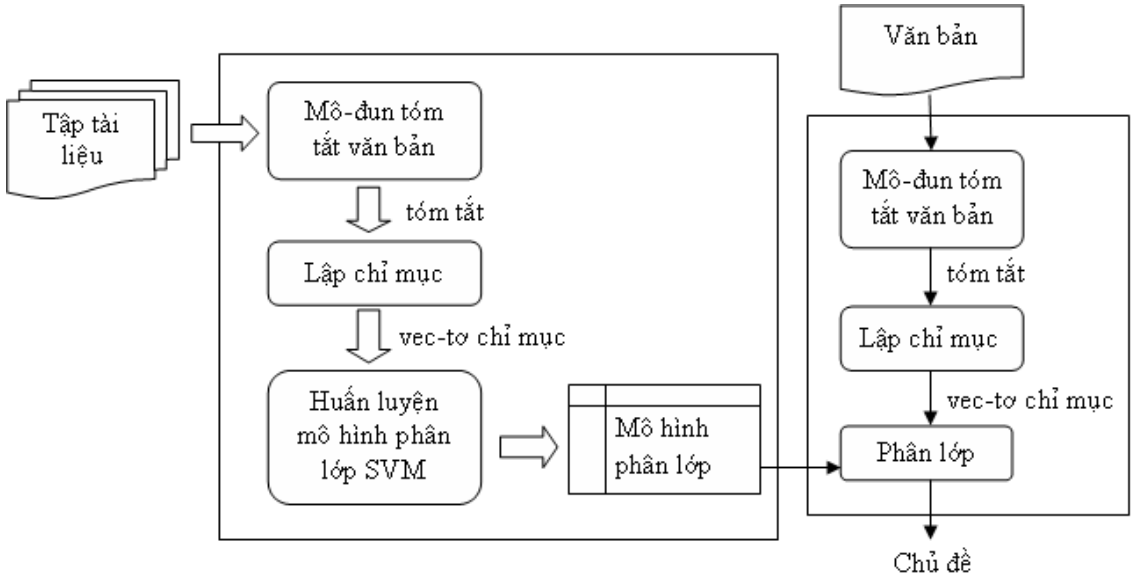
Ngày nay, các giải pháp tóm tắt tự động văn bản đã đạt được một số thành công bước đầu, có thể liệt kê ra đây các công trình tiêu biểu như: tóm tắt văn bản tự động dựa trên việc phân cụm từ khóa và hàm xếp hạng [4], tóm tắt đa văn bản dựa vào trích xuất câu [5], tóm tắt đa văn bản dựa trên chiến lược xếp hạng câu [6]. Các giải pháp trên cũng đã được nhiều nhà nghiên cứu trong nước ứng dụng vào lĩnh vực tóm tắt tự động văn bản tiếng Việt.

Bắt nguồn từ thành công của lĩnh vực tóm tắt văn bản tự động cùng với nhận định rằng phân loại văn bản có thể chỉ cần dựa vào nội dung tóm tắt của văn bản đó thay vì toàn bộ nội dung văn bản chúng tôi đề xuất mô hình phân loại văn bản dựa trên việc rút trích tự động tóm tắt của văn bản đó. Mục tiêu nghiên cứu của chúng tôi là kiểm chứng tính khả thi của giải pháp mà chúng tôi đề xuất chứ không nhắm đến việc đề xuất một giải pháp tốt hơn các giải pháp hiện có dựa trên nội dung toàn văn.

Trong lĩnh vực khai phá dữ liệu, bài toán phân lớp văn bản phần lớn dựa trên kỹ thuật máy học như: cây quyết định (decision tree) [7], k-láng giềng gần nhất (k-nearest neighbor) [8], mạng nơron (nerual network) [9], máy học vectơ hỗ trợ [1] … trong đó giải pháp ứng dụng máy học vectơ hỗ trợ như bộ phân lớp chiếm đa số. Vì thế trong nội dung của bài báo này, chúng tôi cũng sẽ sử dụng máy học vectơ hỗ trợ để phân lớp văn bản và thực hiện đối chiếu, so sánh giải pháp mà chúng tối đề xuất với phương pháp sử dụng toàn văn nội dung văn bản để phân lớp. Chúng tôi cũng thực hiện kiểm chứng giải pháp đề xuất với bộ phân lớp sử dụng kỹ thuật cây quyết định. Kết quả thực nghiệm này một lần nữa khẳng định độ tin cậy của giải pháp mà chúng tôi đề xuất. Nói một cách khác, dùng tóm tắt của văn bản để phân loại không chỉ đúng trong trường hợp bộ phân lớp dựa trên SVM mà còn đúng cho các kỹ thuật khác. Bên cạnh đó, phân hệ rút trích tự động tóm tắt của văn bản được xây dựng dựa trên mô hình chúng tôi đã đề xuất trong [10].

II. MÔ HÌNH ĐỀ XUẤT

Mô hình tổng quan của hệ thống phân loại văn bản dựa vào tóm tắt được minh họa trong hình 1. Hệ thống đề xuất bao gồm hai thành phần chính: thành phần huấn luyện và thành phần phân lớp. Văn bản đầu vào được đưa qua mô-đun tạo tóm tắt trước khi đưa vào thành phần huấn luyện mô hình phân lớp. Thành phần phân lớp cũng có cách xử lý tương tự với kết quả là chủ đề của văn bản cần phân lớp.



Hình 1. Mô hình hệ thống phân loại văn bản dựa vào tóm tắt

A. Biểu diễn văn bản

Văn bản đầu vào cho việc huấn luyện và phân lớp có cấu trúc plain text. Để có thể thực hiện rút trích tự động tóm tắt cũng như phân lớp văn bản với máy học vectơ hỗ trợ thì văn bản cần được biểu diễn dưới dạng thích hợp. Chúng tôi sử dụng mô hình túi từ (bag of words) để biểu diễn văn bản. Mô hình này chỉ quan tâm đến tần suất một từ chỉ mục nào đó xuất hiện trong nội dung văn bản bao nhiêu lần mà không quan tâm đến vị trí xuất hiện của từ chỉ mục đó. Đối với mô hình túi từ, hai công việc cần phải giải quyết đó là tách từ và gán trọng số.

Tiếng Việt có đặc điểm là từ có thể là từ đơn hoặc từ ghép vì thế khoảng trắng không còn là dấu hiệu phân cách từ (như tiếng Anh chẳng hạn). Việc phân tách một câu thành tập hợp đúng các từ có nghĩa là hết sức quan trọng đặc biệt với phân hệ rút trích tóm tắt tự động. Chúng tôi xây dựng mô-đun tách từ bằng cách sử dụng thư viện vnTokenizer [11]. Thư viện này được viết bằng JAVA với độ chính xác tách đúng từ theo công bố của tác giả là trong khoảng từ 96% đến 98%. Ví dụ sau đây minh họa kết quả của giai đoạn tách từ:

- Văn bản nguồn: *"Để có thể thực hiện rút trích tự động tóm tắt cũng như phân lớp văn bản với máy học vectơ hỗ trợ thì văn bản cần được biểu diễn dưới dạng thích hợp"*.
- Văn bản sau giai đoạn tách từ: *"Để có_thể thực_hiện rút trích tự_động tóm_tắt cũng_như phân lớp văn_bản với máy học vectơ hỗ_trợ thì văn_bản cần được biểu_diễn dưới dạng thích_hợp"*.

Sử dụng các kết quả của lĩnh vực tìm kiếm thông tin, chúng tôi quyết định sử dụng giải pháp TF-IDF để đánh trọng số của từ. TF-IDF là giải pháp đánh trọng số kết hợp tính chất quan trọng của một từ trong tài liệu chứa nó (TF – tần suất xuất hiện của từ trong tài liệu) với tính phân biệt của từ trong tập tài liệu nguồn (IDF – nghịch đảo tần suất tài liệu). Cách đánh trọng số TF-IDF có rất nhiều biến thể, trong đó cách được sử dụng nhiều nhất đó là

- TF: tần suất xuất hiện của một từ khóa chỉ mục (indexing term) trong một tài liệu
- IDF = $\log(1 + \frac{N}{n})$, trong đó N là tổng số tài liệu, n là số tài liệu có chứa từ khóa chỉ mục tương ứng.

Như vậy văn bản sẽ được biểu diễn trong không gian các từ chỉ mục với thể hiện là một vectơ các trọng số. Không gian các từ chỉ mục này chính là không gian các đặc trưng sử dụng ở giai đoạn rút trích tóm tắt tự động.

B. Tóm tắt văn bản tiếng Việt tự động

Hệ thống tóm tắt văn bản tiếng Việt tự động [10] do chúng tôi đề xuất đạt kết quả khả quan. Chúng tôi đã đề xuất phương pháp tóm tắt văn bản tiếng Việt dựa trên khái niệm độ tương tự giữa các câu. Sau đó hệ thống tính điểm xếp hạng câu bằng thuật toán PageRank cải tiến. Tóm tắt của văn bản sẽ là các câu có thứ hạng cao nhất.

Các bước chính của quá trình tóm tắt như sau:

- Tách nội dung văn bản thành các câu riêng biệt. Chuyển đổi các câu thành các vectơ trong không gian các từ chỉ mục.
- Xây dựng đồ thị biểu diễn văn bản trong đó mỗi đỉnh của đồ thị tương ứng với một câu của văn bản. Cung nối giữa hai đỉnh có trọng số là độ tương tự giữa hai câu.
- Thuật toán PageRank [12] được biến đổi để phù hợp hơn với ngữ cảnh mới: đồ thị vô hướng và có trọng số. Điểm số PageRank của mỗi đỉnh là căn cứ để lựa chọn câu đưa vào tóm tắt.

Thuật toán tóm tắt tự động mà chúng tôi đề xuất thuộc nhóm giải pháp tóm tắt nội dung văn bản bằng cách rút trích các câu quan trọng của văn bản. Ưu điểm của nhóm giải pháp này đó là chúng ta có thể quyết định được tỷ lệ câu sẽ đưa vào tóm tắt hoặc là thực hiện tóm tắt một cách đệ quy.

Ở bước đầu tiên, sau khi thực hiện tách câu dựa vào các dấu câu, chúng tôi thực hiện giai đoạn biểu diễn văn bản như đã nêu ở phần A – Biểu diễn văn bản. Sau giai đoạn này mỗi câu sẽ được biểu diễn bởi một vectơ trong không gian các từ chỉ mục. Chúng tôi cũng sử dụng cách đánh trọng số TF-IDF để tính trọng số của từ chỉ mục trong mỗi câu.

Ở bước tiếp theo, một đồ thị vô hướng có trọng số sẽ được xây dựng trong đó đỉnh của đồ thị biểu diễn cho các câu của văn bản. Cung nối hai đỉnh sẽ có trọng số là độ tương tự giữa hai câu được định nghĩa theo hệ số Jaccard [13]. Việc lựa chọn hệ số Jaccard để tính độ tương đồng cho câu đã được kiểm chứng trong [10] là đạt kết quả tốt cho bài toán tóm tắt tự động.

Ở bước cuối cùng, thuật toán PageRank [12] đã được điều chỉnh để phù hợp với ngữ cảnh đồ thị vô hướng có trọng số, thay vì là độ thị có hướng và không trọng số như đồ thị WEB. Thuật toán PageRank thông qua một số lần lặp sẽ cập nhật độ quan trọng của các đỉnh một đồ thị. Quá trình lặp sẽ dừng khi lỗi hội tụ thấp hơn một ngưỡng xác định trước hoặc quá trình lặp đã lặp đủ số lượt quy định. Điểm số quan trọng của đỉnh A tại mỗi bước lặp được cập nhật bởi công thức dưới đây

$PR(A) = \frac{1-d}{N} + d(W_{AB}\frac{PR(B)}{L(B)} + W_{AC}\frac{PR(C)}{L(C)} + \cdots)$ trong đó d thường được chọn là 0.85, W_{ij} là trọng số cung nối hai đỉnh i và j, L(i) là số cung xuất phát từ đỉnh i. Trong ngữ cảnh của bài toán tóm tắt, L(i) luôn có giá trị là N-1 với N là số câu của văn bản.

Các câu được sắp xếp theo thứ tự giảm dần của giá trị độ quan trọng. Một tỷ lệ nhất định các câu có giá trị quan trọng cao nhất sẽ được đưa vào tóm tắt. Trong công trình này, chúng tôi sử dụng tỷ lệ câu đưa vào tóm tắt là 15% hoặc tối thiểu là 2 câu.

Ví dụ dưới đây minh họa kết quả của quá trình tóm tắt:

- **Văn bản**: *Windows XP ngừng hỗ trợ vào ngày 8/4 năm sau. Nhiều nhân viên bán hàng bảo hiểm tại Nhật Bản sẽ được chuyển từ máy tính cũ lên tablet chạy Windows 8 để tương tác tốt hơn với khách hàng. Microsoft tại Nhật Bản hôm nay thông báo đang giúp một công ty bảo hiểm lớn của Nhật Bản là Meiji Yasuda nhằm nâng cấp hàng loạt máy tính chạy hệ điều hành sắp tròn 12 tuổi. Các thiết bị mới sẽ chạy Windows 8 do Fujitsu sản xuất cùng nhiều phần mềm và tiện ích cài đặt sẵn. "Trước đây, đội ngũ bán hàng sẽ chuẩn bị các đề xuất trên máy tính chạy Windows XP và sau đó in ra để chia sẻ với các khách hàng. Tuy nhiên, hệ thống thiết bị mới sẽ giúp chấm dứt các bước làm phiền toái này", thông báo của Microsoft có đoạn. Ngoài trang bị phần cứng mới, hãng phần mềm Mỹ cũng sẽ tổ chức khóa đào tạo và hướng dẫn sử dụng thao tác trên phần mềm mới. Các khách hàng cũng sẽ thuận tiện hơn trong việc sử dụng như đăng ký thông tin, tìm hiểu trực tiếp các gói bảo hiểm mà không phải ngập trong đống giấy tờ, văn bản như trước đây. Meiji Yasuda cũng sẽ là công ty bảo hiểm nhân thọ đầu tiên của Nhật thông qua việc sử dụng hoàn toàn hệ điều hành Windows 8 Pro. Microsoft dự kiến sẽ chấm dứt hỗ trợ hệ điều hành Windows XP từ ngày 8/4/2014. Tuy nhiên, đây vẫn là hệ điều hành có số lượng người dùng khổng lồ, kém không nhiều so với vị trí dẫn đầu thuộc về Windows 7. Hỗ trợ công ty bảo hiểm Nhật Bản là một trong những động thái "mạnh tay" của Microsoft giúp Windows XP sớm "nghỉ hưu" và nhường sự phát triển cho các hệ điều hành mới hơn.*
- **Tóm tắt**: *Nhiều nhân viên bán hàng bảo hiểm tại Nhật Bản sẽ được chuyển từ máy tính cũ lên tablet chạy Windows 8 để tương tác tốt hơn với khách hàng. Microsoft tại Nhật Bản hôm nay thông báo đang giúp một công ty bảo hiểm lớn của Nhật Bản là Meiji Yasuda nhằm nâng cấp hàng loạt máy tính chạy hệ điều hành sắp tròn 12 tuổi. "Trước đây, đội ngũ bán hàng sẽ chuẩn bị các đề xuất trên máy tính chạy Windows XP và sau đó in ra để chia sẻ với các khách hàng. Tuy nhiên, hệ thống thiết bị mới sẽ giúp chấm dứt các bước làm phiền toái này", thông báo của Microsoft có đoạn. Microsoft dự kiến sẽ chấm dứt hỗ trợ hệ điều hành Windows XP từ ngày 8/4/2014.*

C. Phân loại văn bản

Để thực hiện quá trình phân lớp, bộ phân lớp cần được huấn luyện từ các tài liệu mẫu đã gán nhãn chủ đề, sau đó dùng bộ phân lớp này để dự đoán lớp của những tài liệu mới (chưa biết chủ đề). Mục tiêu nghiên cứu của chúng tôi là kiểm chứng khả năng sử dụng tóm tắt của văn bản để thực hiện phân loại văn bản thay vì là sử dụng toàn văn nội dung của văn bản. Trong phạm vi nghiên cứu của bài báo này, chúng tôi trích xuất tự động tóm tắt của văn bản để thực hiện kiểm tra. Nếu khả thi, có nghĩa là độ chính xác của phân lớp trong trường hợp sử dụng tóm tắt không quá chênh lệch với trường hợp sử dụng toàn văn thì tương lai hoặc có thể dùng tóm tắt tự động của văn bản hoặc tóm tắt sẵn có của văn bản được tác giả tóm tắt để phân lớp. Chúng tôi thực hiện kiểm chứng với 2 phương pháp SVM và cây quyết định. Chúng tôi sử dụng công cụ WEKA [14] có tích hợp thư viện LIBSVM 1.6 [15] và cây quyết định J48.

Để có thể thực hiện huấn luyện bộ phân lớp với WEKA chúng tôi đã chuẩn bị dữ liệu theo định dạng ARFF (Attribute Relation File Format). Tuy nhiên đối với bài toán phân loại văn bản thì mỗi văn bản được biểu diễn bởi một

vectơ rất thưa, có nghĩa là số lượng đặc trưng có giá trị là 0 rất cao vì thế thay vì sử dụng cấu trúc ARFF chuẩn chúng tôi đã sử dụng cấu trúc sparse ARFF. Cấu trúc sparse ARFF như sau

% <chú thích>

% …

@RELATION <Tên tập dữ liệu>

@ATTRIBUTE <Tên thuộc tính> <Kiểu dữ liệu thuộc tính>

…

@ATTRIBUTE class {Lớp 1, Lớp 2, …}

@DATA

{<$index_i$> <giá trị>, …, <$index_n$> "nhãn lớp"}

D. Đánh giá một giải thuật máy học

Để đánh giá một giải thuật máy học một số chỉ số thông dụng được sử dụng. Giả sử như bộ phân lớp có 2 lớp là lớp âm và lớp dương thì các chỉ số được định nghĩa như sau:

- Số đúng dương (TP- True positive): số phần tử dương được phân loại dương
- Số sai âm (FN - False negative): số phần tử dương được phân loại âm
- Số đúng âm (TN- True negative): số phần tử âm được phân loại âm
- Số sai dương (FP - False positive): số phần tử âm được phân loại dương
- Độ chính xác (Precision) = TP/(TP + FP)
- Độ bao phủ (Recall) = TP/(TP + FN)
- Độ đo F1= 2*Precision*Recall/(Precision + Recall)

Các chỉ số này sẽ được sử dụng trong quá trình đánh giá kết quả thực nghiệm.

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Mục tiêu nghiên cứu của chúng tôi là đánh giá tính khả thi của mô hình phân lớp văn bản dựa vào tóm tắt vì thế chúng tôi sẽ thực hiện so sánh kết quả phân lớp của trường hợp sử dụng tóm tắt của văn bản với trường hợp sử dụng nội dung toàn văn của văn bản. Chúng tôi thực nghiệm trên máy tính cá nhân Asus X202E, CORE i3, 4GB RAM, WINDOWS 8.1.

A. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm dùng đánh giá hệ thống là tập hợp 2000 bài viết thuộc 10 chủ đề khác nhau được thu thập từ trang báo điện tử vnexpress.net và vietnamnet.vn. Chủ đề và số lượng văn bản cho từng chủ đề được cho trong bảng 1.

Bảng 1. Dữ liệu thực nghiệm

Chủ đề	Số lượng văn bản
Vi tính	200
Kinh doanh	200
Pháp luật	200
Giáo dục	200
Sức khỏe	200
Thể thao	200
Khoa học	200
Du lịch	200
Gia đình	200
Ẩm thực	200

Từ 2000 tài liệu toàn văn này, chúng tôi thực hiện tạo tóm tắt tự động để thu được tập dữ liệu văn bản tóm tắt với số lớp chủ đề và số văn bản như nhau. Thời gian tạo tóm tắt trung bình cho mỗi văn bản là khoảng 1 giây khi thực hiện trên máy tính cá nhân có cấu hình như đã nêu ở phân trên (chi tiết xem bảng 2).

Bảng 2. Thời gian tạo tóm tắt tự động cho bộ dữ liệu 2000 văn bản

Chủ đề	Kích thước tập dữ liệu (MB)	Thời gian thực hiện (giây)
Vi tính	5.91	201
Kinh doanh	6.84	280
Pháp luật	6.01	229

Giáo dục	6.59	273
Sức khỏe	6.21	230
Thể thao	6.28	229
Khoa học	6.94	229
Du lịch	6.46	186
Gia đình	6.89	202
Ẩm thực	6.28	242

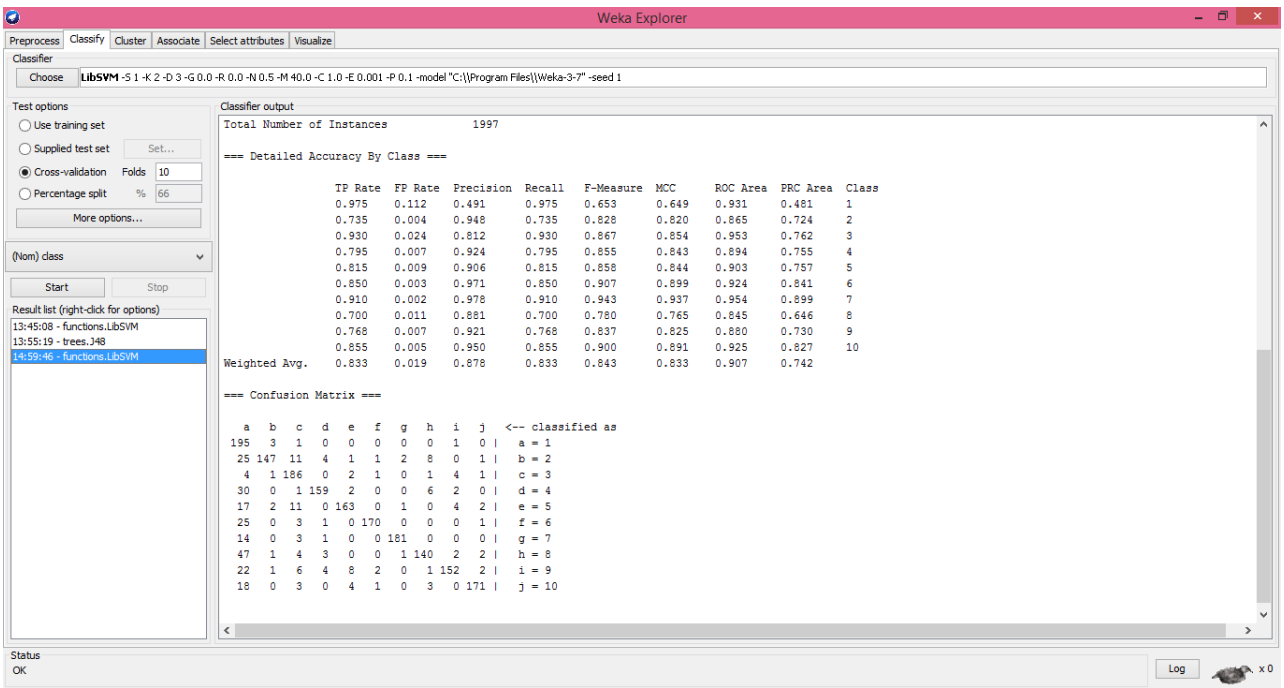
Trong quá trình tạo tập tin sparse ARFF đã tồn tại lỗi nên số lượng mẫu dùng cho quá trình kiểm thử phân lớp trong trường hợp toàn văn là 1997 và số lượng mẫu dùng cho quá trình kiểm thử phân lớp trong trường hợp tóm tắt là 1998 mẫu.

B. Đánh giá kết quả

Trong cả hai trường hợp phân lớp với libSVM và J48, chúng tôi đều sử dụng nghi thức k-fold cross-validation với k = 10.

1. Phân lớp với libSVM

Trong trường hợp phân lớp với libSVM, kết quả phân lớp trên tập dữ liệu toàn văn với công cụ Weka có kết quả như hình 2.



Hình 2. Kết quả thực nghiệm tập dữ liệu toàn văn với bộ phân lớp libSVM sử dụng công cụ Weka

Bảng 3 trình bày chi tiết các chỉ số đánh giá cho tập dữ liệu toàn văn này

Bảng 3. Kết quả phân lớp dữ liệu toàn văn với libSVM

Lớp	TP Rate	FP Rate	Precision	Recall	F-measure
Vi tính	0.975	0.112	0.491	0.975	0.653
Kinh doanh	0.735	0.004	0.948	0.735	0.828
Pháp luật	0.930	0.024	0.812	0.930	0.867
Giáo dục	0.795	0.007	0.924	0.795	0.855
Sức khỏe	0.815	0.009	0.906	0.815	0.858
Thể thao	0.850	0.003	0.971	0.850	0.907
Khoa học	0.910	0.002	0.978	0.910	0.943
Du lịch	0.700	0.011	0.881	0.700	0.780

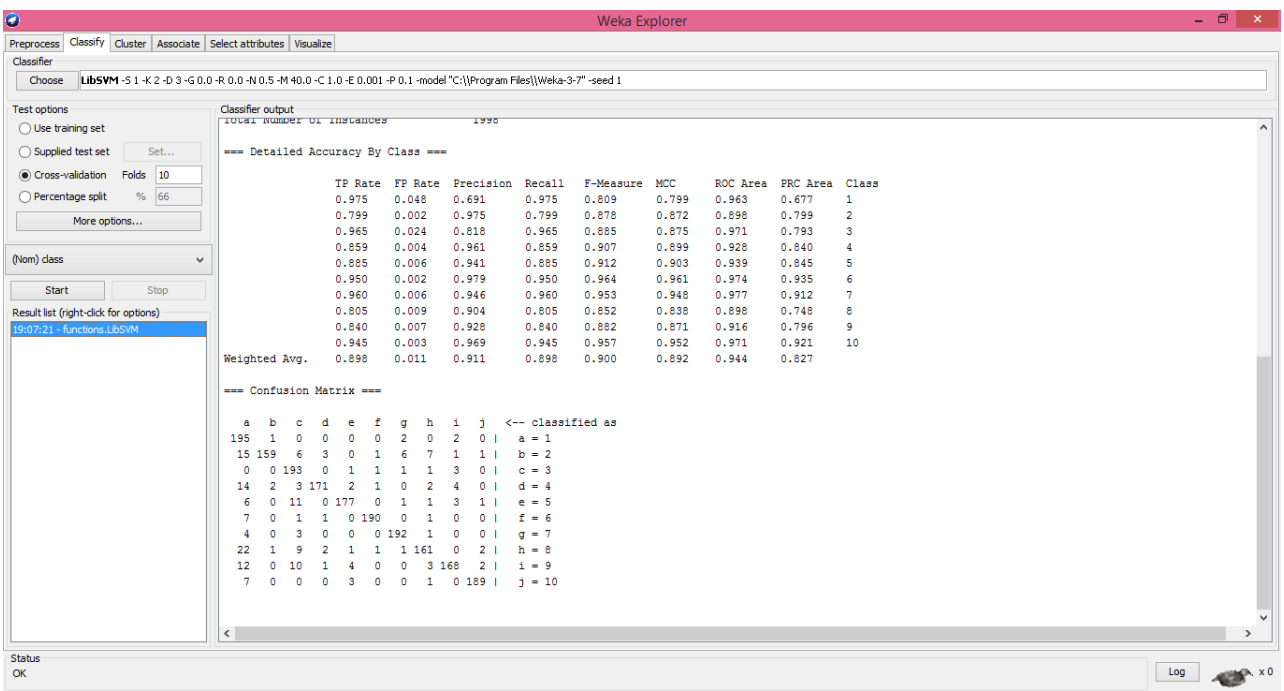
Gia đình	0.768	0.007	0.921	0.768	0.837
Ẩm thực	0.855	0.005	0.950	0.855	0.900
Trung bình	0.833	0.019	0.878	0.833	0.843

Bảng 4 trình bày chi tiết ma trận sai số (confusion matrix) trong trường hợp dữ liệu toàn văn

Bảng 4. Ma trận sai số cho trường hợp dữ liệu toàn văn

Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10
Vi tính	1	**195**	3	1	0	0	0	0	0	1	0
Kinh doanh	2	25	**147**	11	4	1	1	2	8	0	1
Pháp luật	3	4	1	**186**	0	2	1	0	1	4	1
Giáo dục	4	30	0	1	**159**	2	0	0	6	2	0
Sức khỏe	5	17	2	11	0	**163**	0	1	0	4	2
Thể thao	6	25	0	3	1	0	**170**	0	0	0	1
Khoa học	7	14	0	3	1	0	0	**181**	0	0	0
Du lịch	8	47	1	4	3	0	0	1	**140**	2	2
Gia đình	9	22	1	6	4	8	2	0	1	**152**	2
Ẩm thực	10	18	0	3	0	4	1	0	3	0	**171**

Trong trường hợp phân lớp với libSVM, kết quả phân lớp trên tập dữ liệu tóm tắt với công cụ Weka có kết quả như hình 3.



Hình 3. Kết quả thực nghiệm tập dữ liệu tóm tắt với bộ phân lớp libSVM sử dụng công cụ Weka

Bảng 5 trình bày chi tiết các chỉ số đánh giá cho tập dữ tóm tắt văn này

Bảng 5. Kết quả phân lớp dữ liệu tóm tắt với libSVM

Lớp	TP Rate	FP Rate	Precision	Recall	F-measure
Vi tính	0.975	0.048	0.691	0.975	0.809
Kinh doanh	0.799	0.002	0.975	0.799	0.878
Pháp luật	0.965	0.024	0.818	0.965	0.885

Giáo dục	0.859	0.004	0.961	0.859	0.907
Sức khỏe	0.885	0.006	0.941	0.885	0.912
Thể thao	0.950	0.002	0.979	0.950	0.964
Khoa học	0.960	0.006	0.946	0.960	0.953
Du lịch	0.805	0.009	0.904	0.805	0.852
Gia đình	0.840	0.007	0.928	0.840	0.882
Ẩm thực	0.945	0.003	0.969	0.945	0.957
Trung bình	0.898	0.011	0.911	0.898	0.900

Bảng 6 trình bày chi tiết ma trận sai số (confusion matrix) trong trường hợp dữ liệu tóm tắt.

Bảng 6. Ma trận sai số cho trường hợp dữ liệu tóm tắt

Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10
Vi tính	1	**195**	1	0	0	0	0	2	0	2	0
Kinh doanh	2	15	**159**	6	3	0	1	6	7	1	1
Pháp luật	3	0	0	**193**	0	1	1	1	1	3	0
Giáo dục	4	14	2	3	**171**	2	1	0	2	4	0
Sức khỏe	5	6	0	11	0	**177**	0	1	1	3	1
Thể thao	6	7	0	1	1	0	**190**	0	1	0	0
Khoa học	7	4	0	3	0	0	0	**192**	1	0	0
Du lịch	8	22	1	9	2	1	1	1	**161**	0	2
Gia đình	9	12	0	10	1	4	0	0	3	**168**	2
Ẩm thực	10	7	0	0	0	3	0	0	1	0	**189**

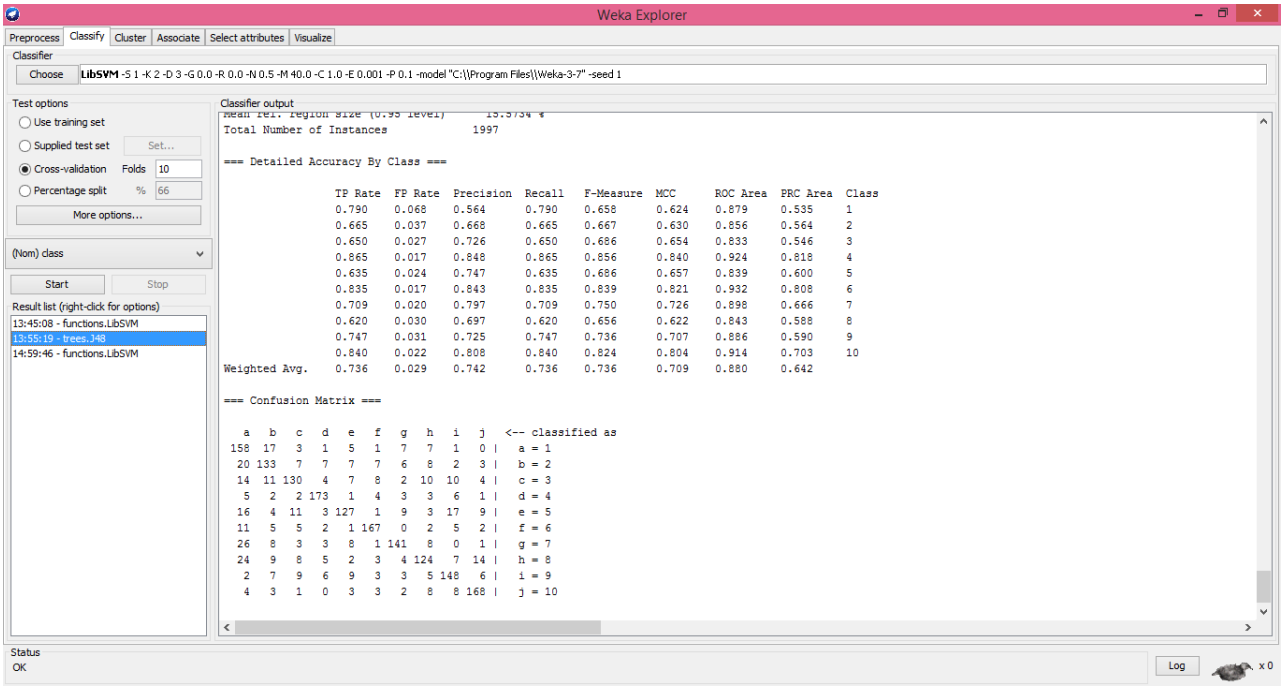
Với kết quả thực nghiệm như trên chúng ta có thể thấy rằng kết quả phân loại với tập dữ liệu tóm tắt là tốt hơn tập dữ liệu toàn văn trên tất tả các chỉ số và hầu như trên tất cả các lớp. Nếu xét riêng chỉ số TP rate, chỉ số phân lớp đúng, thì tập dữ liệu tóm tắt cho kết quả tốt hơn tập toàn văn ở cả 9/10 lớp và ở giá trị trung bình thì vượt hơn được gần **7%**, **89.8%** so với **83.3%**.

2. Phân lớp với cây quyết định – J48

Trong trường hợp phân lớp với J48, kết quả phân lớp trên tập dữ liệu toàn văn với công cụ Weka có kết quả như hình 4. Bảng 7 trình bày chi tiết các chỉ số đánh giá cho tập dữ liệu toàn văn này.

Bảng 7. Kết quả phân lớp dữ liệu toàn văn với J48

Lớp	TP Rate	FP Rate	Precision	Recall	F-measure
Vi tính	0.790	0.068	0.564	0.790	0.658
Kinh doanh	0.665	0.037	0.668	0.665	0.667
Pháp luật	0.650	0.027	0.726	0.650	0.686
Giáo dục	0.865	0.017	0.848	0.865	0.856
Sức khỏe	0.635	0.024	0.747	0.635	0.686
Thể thao	0.835	0.017	0.843	0.835	0.839
Khoa học	0.709	0.020	0.797	0.709	0.750
Du lịch	0.620	0.030	0.697	0.620	0.656
Gia đình	0.747	0.031	0.725	0.747	0.736
Ẩm thực	0.840	0.022	0.808	0.840	0.824
Trung bình	0.736	0.029	0.742	0.736	0.736



Hình 4. Kết quả thực nghiệm tập dữ liệu toàn văn với bộ phân lớp J48 sử dụng công cụ Weka

Bảng 8 trình bày chi tiết ma trận sai số (confusion matrix) trong trường hợp dữ liệu toàn văn.

Bảng 8. Ma trận sai số cho trường hợp dữ liệu toàn văn

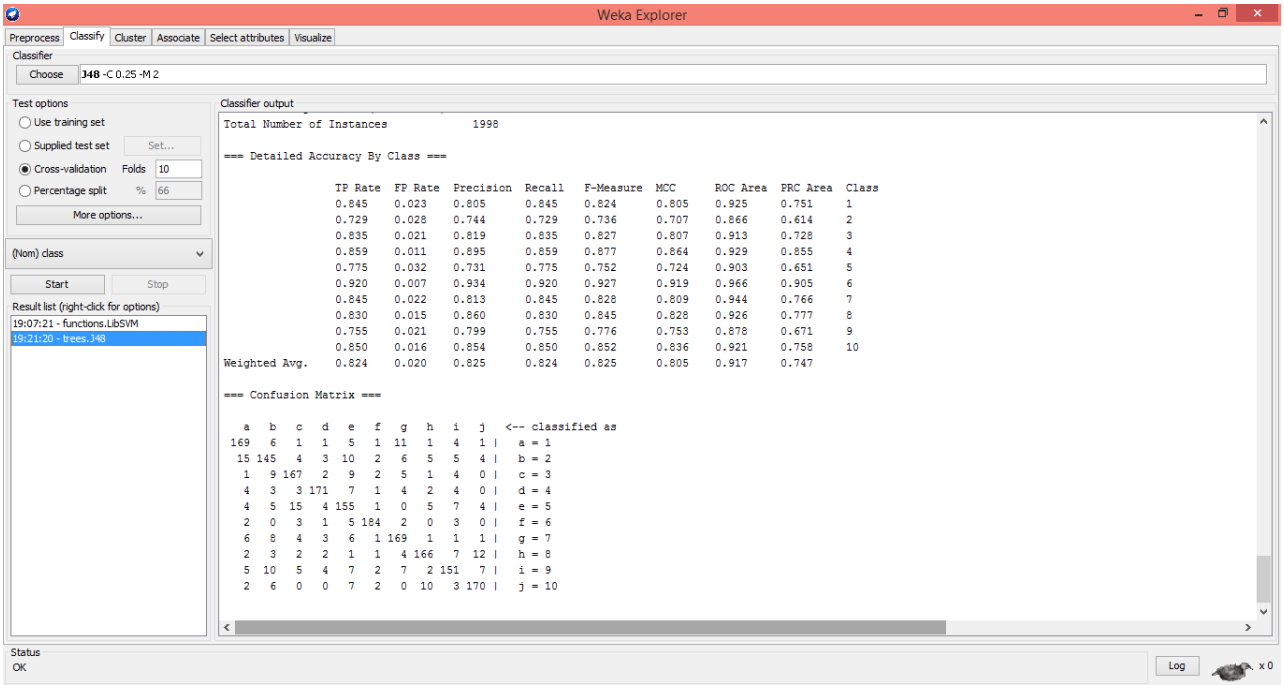
Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10
Vi tính	1	**158**	17	3	1	5	1	7	7	1	0
Kinh doanh	2	20	**133**	7	7	7	7	6	8	2	3
Pháp luật	3	14	11	**130**	4	7	8	2	10	10	4
Giáo dục	4	5	2	2	**173**	1	4	3	3	6	1
Sức khỏe	5	16	4	11	3	**127**	1	9	3	17	9
Thể thao	6	11	5	5	2	1	**167**	0	2	5	2
Khoa học	7	26	8	3	3	8	1	**141**	8	0	1
Du lịch	8	24	9	8	5	2	3	4	**124**	7	14
Gia đình	9	2	7	9	6	9	3	3	5	**148**	6
Ẩm thực	10	4	3	1	0	3	3	2	8	8	**168**

Trong trường hợp phân lớp với J48, kết quả phân lớp trên tập dữ liệu tóm tắt với công cụ Weka có kết quả như hình 5. Bảng 9 trình bày chi tiết các chỉ số đánh giá cho tập dữ liệu tóm tắt này.

Bảng 9. Kết quả phân lớp dữ liệu tóm tắt với J48

Lớp	TP Rate	FP Rate	Precision	Recall	F-measure
Vi tính	0.845	0.023	0.805	0.845	0.824
Kinh doanh	0.729	0.028	0.744	0.729	0.736
Pháp luật	0.835	0.021	0.819	0.835	0.827
Giáo dục	0.859	0.011	0.895	0.859	0.877
Sức khỏe	0.775	0.032	0.731	0.775	0.752
Thể thao	0.920	0.007	0.934	0.920	0.927
Khoa học	0.845	0.022	0.813	0.845	0.828
Du lịch	0.830	0.015	0.860	0.830	0.845

Gia đình	0.755	0.021	0.799	0.755	0.776
Ẩm thực	0.850	0.016	0.854	0.850	0.852
Trung bình	0.824	0.020	0.825	0.824	0.825



Hình 5. Kết quả thực nghiệm tập dữ liệu tóm tắt với bộ phân lớp J48 sử dụng công cụ Weka

Bảng 10 trình bày chi tiết ma trận sai số (confusion matrix) trong trường hợp dữ liệu tóm tắt.

Bảng 10. Ma trận sai số cho trường hợp dữ liệu tóm tắt

Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10
Vi tính	1	**169**	6	1	1	5	1	11	1	4	1
Kinh doanh	2	15	**145**	4	3	10	2	6	5	5	4
Pháp luật	3	1	9	**167**	2	9	2	5	1	4	0
Giáo dục	4	4	3	3	**171**	7	1	4	2	4	0
Sức khỏe	5	4	5	15	4	**155**	1	0	5	7	4
Thể thao	6	2	0	3	1	5	**184**	2	0	3	0
Khoa học	7	6	8	4	3	6	1	**169**	1	1	1
Du lịch	8	2	3	2	2	1	1	4	**166**	7	12
Gia đình	9	5	10	5	4	7	2	7	2	**151**	7
Ẩm thực	10	2	6	0	0	7	2	0	10	3	**170**

Với kết quả thực nghiệm như trên chúng ta có thể thấy rằng kết quả phân loại với tập dữ liệu tóm tắt trong trường hợp sử dụng cây quyết định J48 cũng vẫn tốt hơn tập dữ liệu toàn văn trên tất tả các chỉ số và trên tất cả các lớp. Nếu xét riêng chỉ số TP rate, chỉ số phân lớp đúng, thì tập dữ liệu tóm tắt cho kết quả tốt hơn tập toàn văn ở cả 10/10 lớp và ở giá trị trung bình thì vượt hơn được gần **10%**, **82,4%** so với **73,6%**.

IV. KẾT LUẬN

Trong bài báo này chúng tôi giới thiệu mô hình phân lớp văn bản dựa trên tóm tắt tự động của văn bản. Đây là một hướng tiếp cận rất mới và chưa có nhiều nghiên cứu trên thế giới gũng như ở Việt Nam vì đại bộ phận đều cho rằng khi thực hiện tóm tắt văn bản thì thông tin dùng cho phân lớp đã mất đi khá nhiều. Kết quả mà chúng tôi thu được từ nghiên cứu này là hết sức khả quan và thiết nghĩ là hoàn toàn khả thi khi ứng dụng vào thực tế. Thật vậy giải pháp chúng tôi đề xuất luôn chiếm ưu thế ở tất các tiêu chí đánh giá mà đặc biệt quan trọng là vượt trội ở các tiêu chí tỷ lệ phân lớp đúng trên tất cả các lớp.

Kết quả khả quan của mô hình đề xuất có thể được lý giải bởi nhiều nguyên nhân: 1- Tóm tắt của một văn bản về lý thuyết sẽ tóm lược được nội dung cốt lõi truyền tải bởi văn bản. Một khi đã tóm lược được nội dung chính thì chủ đề của văn bản hoàn toàn có thể xác định được. 2- Cách thức biểu diễn văn bản đã thể hiện tốt nội dung, ngữ nghĩa của văn bản. Thật vậy, trong nghiên cứu của mình, chúng tôi dựa trên "mô hình túi từ - bag of words" để biểu diễn nội dung văn bản, phương pháp này có ưu điểm là cài đặt đơn giản nhưng có hạn chế lớn là làm mất đi ngữ nghĩa của văn bản vì không quan tâm đến vị trí của từ mà chỉ quan tâm đến tần suất xuất hiện của từ. Việc sử dụng thư viện vnTokenizer có khả năng nhận biết chính xác từ đơn và từ ghép đồng thời việc tạo tóm tắt được thực hiện trên mức câu nên đã giúp giữ lại phần nào ngữ nghĩa của văn bản; 3- Mô hình tóm tắt tự động văn bản mà chúng tôi đề xuất trong nghiên cứu trước đây thật sự là khả thi. Điểm mấu chốt của bài toán tóm tắt là tính độ tương tự giữa các câu và tính điểm xếp hạng các câu dựa trên mô hình đồ thị. Độ tương tự giữa các câu được tính thông qua độ đo Jaccard có chú trọng đến mối tương quan về độ dài của các câu. Thuật toán PageRank dùng để tính điểm xếp hạng các câu đưa vào tóm tắt là thuật toán xếp hạng các trang web và đã chứng tỏ được tính khả thi khi được ứng dụng thành công trong các bộ máy tìm kiếm thông tin web. Một ưu điểm khác của mô hình tóm tắt tự động đó là quá trình tóm tắt không cần tập ngữ liệu huấn luyện, cũng như không cần xem xét tính ngữ nghĩa và cấu trúc ngữ pháp của câu và việc tóm tắt được áp dụng trên từng văn bản đơn.

Mặc dù kết quả nghiên cứu bước đầu đã khẳng định mô hình đề xuất phân lớp văn bản dựa vào tóm tắt là hoàn toàn khả thi và hoàn toàn có thể áp dụng vào thực tế, tuy nhiên kết quả ấy cũng chỉ được thực nghiệm trên một tập chưa đủ lớn các tài liệu và cũng chỉ mới chứng tỏ tính khả thi khi phương pháp phân lớp sử dụng máy học vectơ hỗ trợ và cây quyết định. Để kết quả nghiên cứu có tính thuyết phục hơn thì tập dữ liệu thực nghiệm cần có kích thước lớn hơn nữa đồng thời để không làm mất tính tổng quát thì mô hình phân lớp dựa vào tóm tắt phải được kiểm chứng trên các kỹ thuật máy học khác.

V. TÀI LIỆU THAM KHẢO

[1] Thorsten Joachims, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", Proceedings of the 10th European Conference on Machine Learning, p.137-142, April 21-23, 1998.

[2] Ho, T.B, Nguyen, N.B, "Nonhierachical Document Clustering by a Tolerance Rough Set Model", International Journal of Fuzzy Logic and Inteligent Systems, Vol. 17, No.2, 199-212, 2012.

[3] Osmar R. Zaïane , Maria-Luiza Antonie, "Classifying text documents by associating terms with text categories", Proceedings of the 13th Australasian database conference, p.215-222, January 01, 2002, Melbourne, Victoria, Australia.

[4] Massih R. Amini , Nicolas Usunier , Patrick Gallinari, "Automatic text summarization based on word-clusters and ranking algorithms", Proceedings of the 27th European conference on Advances in Information Retrieval Research, March 21-23, 2005, Santiago de Compostela, Spain [doi>10.1007/978-3-540-31865-1_11]

[5] Jade Goldstein , Vibhu Mittal , Jaime Carbonell , Mark Kantrowitz, "Multi-document summarization by sentence extraction", Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization, p.40-48, April 30-30, 2000, Seattle, Washington [doi>10.3115/1117575.1117580].

[6] R. Barzilay, N. Elhadad, and K. McKeown, "Inferring strategies for sentence ordering in multidocument news summarization," Journal of Artificial Intelligence Research, vol. 17, pp. 35–55, 2002.

[7] D. Johnson, F. Oles, T. Zhang, T. Goetz, "A Decision Tree-based Symbolic Rule Induction System for Text Categorization", IBM Systems Journal, 41(3), pp. 428–437, 2002.

[8] E.-H. Han, G. Karypis, V. Kumar, "Text Categorization using Weighted-Adjusted k-nearest neighbor classification", PAKDD Conference, 2001.

[9] M. Ruiz, P. Srinivasan, "Hierarchical neural networks for text categorization", ACM SIGIR Conference, 1999.

[10] Trương Quốc Định, Nguyễn Quang Dũng, "Một giải pháp tóm tắt văn bản tiếng Việt tự động", Kỷ yếu hội thảo khoa học quốc gia lần thứ XV, trang 233-238, Nhà xuất bản Khoa học và Kỹ thuật, Hà Nội, 2012.

[11] Lê Hông Phuong , Nguyên Thi Minh Huyên , Azim Roussanaly , Hô Tuòng Vinh, "A Hybrid Approach to Word Segmentation of Vietnamese Texts", Language and Automata Theory and Applications: Second International Conference, LATA 2008, Tarragona, Spain, March 13-19, 2008. Revised Papers, Springer-Verlag, Berlin, Heidelberg, 2008 [doi>10.1007/978-3-540-88282-4_23].

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", 1999.

[13] Jaccard P., "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547–579.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009), "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.

[15] C. C. Chang and C. J. Lin, "LIBSVM: Alibrary for support vector machines", ACM Trans, on Intelligent System and Technology, 2011.

TEXT CLASSIFICATION BASED ON AUTOMATIC TEXT SUMMARIZATION

Truong Quoc Dinh

ABSTRACT*— to be done.*