PHÂN LOẠI NỘI DUNG TÀI LIỆU WEB TIẾNG VIỆT

Trần Ngọc Phúc[1, *], Phạm Trần Vũ[2], Phạm Công Xuyên[1], Nguyễn Vũ Duy Quang[1]

[1]*Khoa Công nghệ Thông tin, Trường Đại học Lạc Hồng*

[2]*Khoa Khoa học và Kỹ thuật máy tính, Trường Đại học Bách khoa TP. HCM*

Email: *tnphuc@lhu.edu.vn, t.v.pham@cse.hcmut.edu.vn, xuyen@lhu.edu.vn, quang@lhu.edu.vn*

Đến Toà soạn: 21/8/2013; Chấp nhận đăng: 10/11/2013

TÓM TẮT

Bài báo trình bày một số kết quả nghiên cứu, ứng dụng thuật toán Latent Dirichlet Allocation (LDA) phân tích chủ đề ẩn, để tìm tập đặc trưng cho các chủ đề áp dụng cho bài toán phân loại nội dung tài liệu web. Trong bài báo này các cụm danh từ được sử dụng để làm đặc trưng văn bản trong mô hình vector. Các bước thực hiện bao gồm thuật toán tách từ, gán nhãn từ loại để rút trích ra các cụm danh từ. Sử dụng phương pháp đếm tần suất từ và độ đo sự tương đồng cosine để tiến hành phân loại. Thuật toán Latent Dirichlet Allocation được sử dụng để tìm tập đặc trưng cho các chủ đề mà không cần quan tâm đến tần số xuất hiện, độ quan trọng của từ mà vẫn đưa ra bộ dữ liệu đầy đủ và chính xác. Kết quả đã cài đặt thử nghiệm vào bài toán phân lớp các tin tức phổ biến trên các trang báo tiếng Việt với độ chính xác khoảng 90% đáp ứng được mục tiêu phân loại đề ra.

Từ khóa: phân loại văn bản, xử lí ngôn ngữ tiếng Việt.

1. GIỚI THIỆU

Sự phát triển của Công nghệ thông tin làm tăng thêm sự tiện nghi và linh động của Internet từ đó tạo ra một số lượng lớn tài liệu trên web. Trong số đó, tài liệu văn bản là phổ biến mà con người thường gặp nhất (chiếm trên 80 %). Bài toán phân loại tài liệu văn bản được đặt ra với mục đích giúp con người tiết kiệm thời gian trong việc tìm kiếm, tổng hợp thông tin, và quản lí dữ liệu. Có nhiều phương pháp phân loại tài liệu văn bản như cây quyết định, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, … Các công trình liên quan đến vấn đề phân loại dữ liệu đã được công bố như luận án tiến sĩ "*Active Learning for Text Classification"*[1] của tác giả Rong Hu năm 2011 (School of Computing, Dublin Institute of Technology).Trong luận án này, các thông tin được đưa vào học máy dùng các thuật toán gom cụm để tạo ra bộ dữ liệu mẫu. Từ đó tập trung vào việc tối ưu cho việc học máy tích cực. Bài báo *"Text Categorization with Support Vector Machines: Learning with Many Relevant Features"*[2] của tác giả Thorsten Joachims (Đại học Dortmund, Đức). Bài báo trình bày về việc sử dụng và cải tiến kĩ thuật Support Vector Machines (SVM) cho việc học máy có hiệu quả trong việc phân loại văn bản. Bài báo *"Text Categorization"*[3] của tác giả Fabrizio Sebastiani (Đại học Padova, Ý). Bài báo trình bày ba giai đoạn trong một hệ thống phân loại văn bản: lập chỉ mục tài liệu văn bản dùng

LSI (Latent Semantic Index), học tập phân loại văn bản dùng SVM và Boosting, và đánh giá phân loại văn bản. Bài báo *"Text Categorization Based on Regulazired Linear Classification Methods"*[4] của nhóm tác giả Tong Zhang và Franks J. Oles (Mathematical Sciences Department, IBM T.J. Watson Research Center, NewYork). Bài báo này trình bày phương pháp phân loại văn bản tuyến tính dựa vào các kĩ thuật Linear Least Squares Fit, Logistic Regression, SVM.

Hầu hết các công trình trên đều dựa trên tiếp cận máy học, mô hình xác suất và thống kê, thường tập trung vào bài toán phân làm 2 lớp và gặp khó khăn với dữ liệu lớn. Mặt khác, các công trình này dành cho xử lí ngôn ngữ tiếng nước ngoài, cụ thể là tiếng Anh. Để áp dụng cho các tài liệu văn bản bằng tiếng Việt thì không có được độ chính xác như mong muốn.

Ở Việt Nam có một số công trình như luận văn thạc sĩ "*Phương pháp phân cụm tài liệu Web và áp dụng vào bộ máy tìm kiếm"* [5] của tác giả Nguyễn Thị Thu Hằng (Trường Đại học Công nghệ Hà Nội). Luận văn này dựa vào các cụm từ đặc trưng, phân cụm cây hậu tố để giải quyết vấn đề phân cụm tài liệu web. Bài báo *"Phân loại văn bản dựa trên cụm từ phổ biến"* [6] của Hoàng Kiếm và Đỗ Phúc (Trường Đại học Công nghệ thông tin TP.HCM) trình bày một phương pháp phân loại dựa trên các cụm từ phổ biến. Luận văn thạc sĩ*"Phân loại trang Web dựa trên phương pháp đồng huấn luyện"*[7] của tác giả Đặng Vũ Tùng (Học viện Công nghệ Bưu chính Viễn thông Hà Nội). Luận văn này sử dụng và cải tiến một số biến thể của thuật toán đồng huấn luyện sử dụng nhiều thuật toán phân lớp. Luận văn thạc sĩ "*Phát triển bộ công cụ hỗ trợ xây dựng kho ngữ liệu cho phân tích văn bản tiếng Việt"*[8] của Lưu Văn Tăng (Trường Đại học Khoa học Tự nhiên Hà Nội). Luận văn này áp dụng việc phân loại tài liệu văn bản tiếng Việt, xây dựng kho ngữ liệu với nhiều thể loại và chủ đề.

Các công trình trên đều có những ưu điểm nhất định của nó, tuy nhiên phạm vi xử lí văn bản của nó quá rộng, hầu như không xác định cụ thể cho một loại văn bản nào. Do đó, kết quả cho ra độ chính xác không được đồng nhất và khó để đánh giá.

Ngoài ra, phân loại tài liệu văn bản còn được áp dụng rất nhiều cho một số lĩnh vực như hệ thống tìm kiếm online, quản lí thư mục trang web, lọc mail, hệ thống tư vấn online, …

Trong bài báo này, chúng tôi đề xuất một phương pháp mới, dùng thuật toán Latent Dirichlet Allocation trong việc tìm tập đặc trưng cho chủ đề văn bản để áp dụng vào bài toán phân loại đáp ứng độ chính xác cao ngay cả số lượng dữ liệu lớn.

Những vấn đề nêu ra của bài báo được trình bày trong 4 phần: phần thứ nhất giới thiệu về bài toán và những công trình liên quan. Trong phần thứ hai, trình bày về phương pháp thực hiện. Phần thứ ba, đưa ra kết quả thực nghiệm và kết luận ở phần thứ tư.

2. PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN

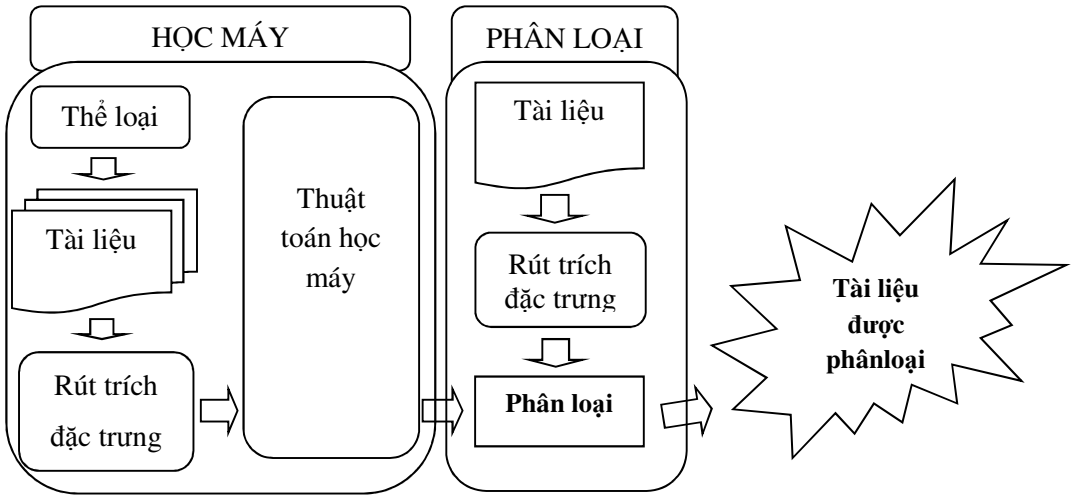
Quy trình phân loại chung cho cácphương pháp phân loại:

Các bước trong hình 1:

- Bước 1: Xây dựng bộ dữ liệu chủ quan dựa vào tài liệu văn bản đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lí và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng chủ đề.
- Bước 2: Dữ liệu cần phân loại được xử lí, rút ra đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và đưa ra kết quả.

Dữ liệu đầu vào cho quá trình học máy hay dữ liệu đầu vào để phân loại đều là dạng văn bản đã qua công đoạn tiền xử lí. Công đoạn tiền xử lí này rất quan trọng và cần thiết, nó làm tối

ưu hóa dữ liệu trong việc lưu trữ và xử lí. Các công đoạn trong quá trình tiền xử lí văn bản bao gồm: tách từ tiếng Việt, loại bỏ các từ dừng, từ tầm thường lấy các danh từ. Sau đó, rút trích đặc trưng và biểu diễn văn bản.



Hình 1. Quy trình phân loại.

2.1. Tách từ tiếng Việt

Đối với tiếng Anh, các từ được phân cách nhau bằng các khoảng trắng hoặc dấu chấm câu. Đối với tiếng Việt có thể có các từ ghép, ví dụ: "công nghệ". Bài báo sử dụngkĩ thuật tách từ Maximum Matching với công cụ tách từ vnTokenizer [9, 10]. Ở phương pháp này, việc duyệt một văn bản được thực hiện từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển, sau đó lặp lại cho đến hết câu. Công cụ vnTokenizer sử dụng kết hợp từ điển và n-gram, trong đó mô hình n-gram được huấn luyện sử dụng VietTreebank (70.000 câu đã được tách từ) có thể lọc các đơn vị từ đặc biệt (xâu dạng số, ngày tháng,…) và các file chứa các thống kê unigram và bigram trên kho văn bản tách từ mẫu đạt độ chính xác trên 97 %.

2.2. Chọn danh từ kết hợp loại bỏ từ dừng, từ thầm thường

Sau khi loại bỏ từ dừng (stop-words), giữ lại các danh từ thì câu vẫn giữ lại ý nghĩa đầy đủ của văn bản. Do đó, chỉ giữ lại những danh từ. Từ đó cho thấy vẫn giữ nguyên ý nghĩa của văn bản, đồng thời giảm chi phí cho việc lưu trữ, cũng như việc tính toán.

Các từ dừng ở đây dùng để chỉ các từ mà xuất hiện quá nhiều trong các câu văn bản của toàn tập kết quả, thường thì không giúp ích gì trong việc phân biệt nội dung của các tài liệu văn bản, ví dụ: "và", "hoặc", "cũng", "là", "mỗi", "bởi", …

Ở quá trình này, từ văn bản đã được tách từ, sử dụng kết hợp từ điển từ dừng và công cụ gán nhãn từ loại [11] để thu về các danh từ. Việc gán nhãn từ loại được thực hiện bằng cách sử dụng công cụ JvnTagger [9, 10], với độ chính xác trên 93 % với dữ liệu huấn luyện khoảng 10.000 câu và 20.000 câu của VietTreebank. Văn bản sau khi gán nhãn từ loại từ vựng là một chuỗi các danh từ.

2.3. Xây dựng bộ dữ liệu tập đặc trưng

Để thực hiện phân loại, đòi hỏi phải có bộ dữ liệu chuẩn và chính xác đáp ứng được yêu cầu. Quá trình này sử dụng mô hình phân tích chủ đề ẩn để thực hiện. Từ các tập văn bản được phân loại chủ quan theo từng thể loại riêng biệt, qua các bước tiền xử lí, thực hiện phân tích chủ đề ẩn, rút ra được các tập đặc trưng của từng thể loại.

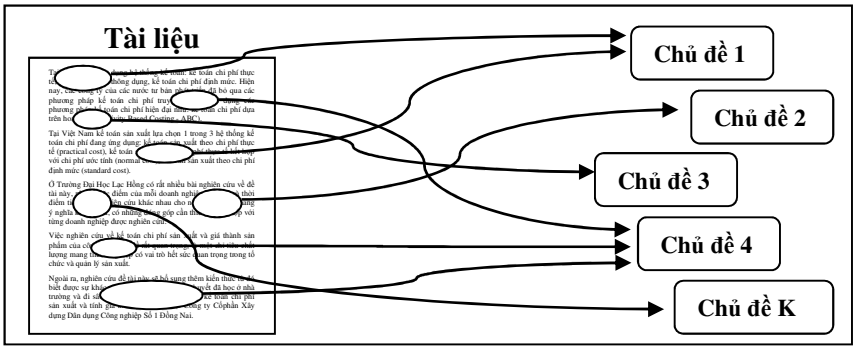
- **Mô hình phân tích chủ để ẩn**

Ý tưởng của các mô hình chủ đề ẩn là xây dựng những tài liệu mới dựa theo phân phối xác suất. Trước hết, để tạo ra một tài liệu mới, cần chọn ra một phân phối những chủ đề cho tài liệu đó. Điều này có nghĩa tài liệu được tạo nên từ những chủ đề khác nhau, với những phân phối khác nhau. Tiếp đó, để sinh các từ cho tài liệu, có thể lựa chọn ngẫu nhiên các từ dựa vào phân phối xác suất của các từ trên các chủ đề. Ngược lại, cho một tập các tài liệu, có thể xác định một tập các chủ đề ẩn cho mỗi tài liệu và phân phối xác suất của các từ trên từng chủ đề.

Hai ví dụ về phân tích chủ đề sử dụng mô hình ẩn là Probabilistic Latent Semantic Analysis (pLSA) vàLDA. pLSA là một kĩ thuật thống kê nhằm phân tích những dữ liệu xuất hiện đồng thời [12]. Nó được phát triển dựa trên Latent Semantic Analysis kết hợp với một mô hình xác suất. Tuy nhiên, theo phân tích của Blei và các cộng sự (xem [13]), mặc dù pLSA là một bước quan trọng trong việc mô hình hóa dữ liệu văn bản, nhưng nó vẫn còn chưa hoàn thiện ở chỗ chưa xây dựng được một mô hình xác suất tốt ở mức độ tài liệu. Điều đó dẫn đến vấn đề gặp phải khi phân phối xác suất cho một tài liệu nằm ngoài tập dữ liệu học, ngoài ra số lượng các tham số có thể tăng lên một cách tuyến tính khi kích thước của tập dữ liệu tăng. Trong khi đó, LDA là một mô hình hoàn thiện hơn, nó có thể khắc phục được những nhược điểm ở trên. Mô hình chủ đề ẩn LDA này được sử dụng trong việc xây dựng dữ liệu cho hệ thống.

- **Mô hình Latent Dirichlet Allocation**

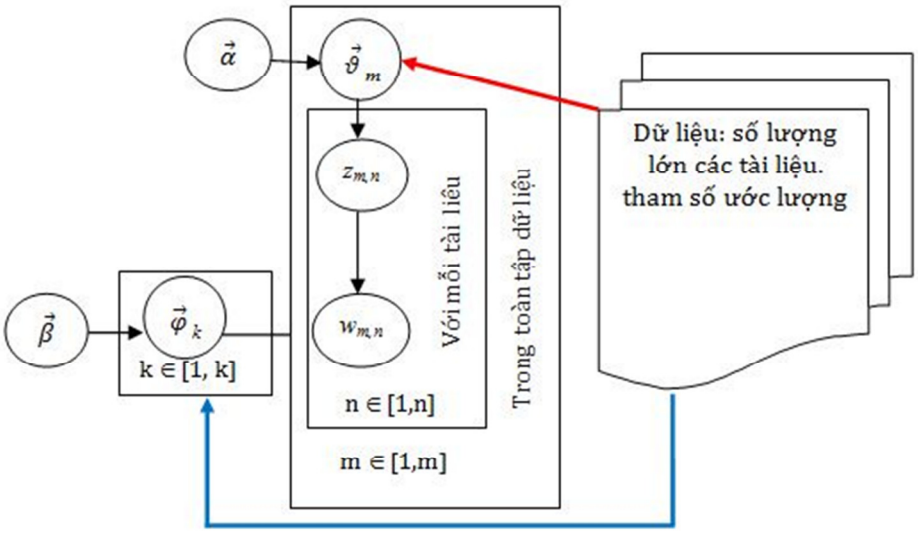
LDA [13, 14, 15, 16] là một mô hình sinh xác suất cho tập dữ liệu rời rạc như text corpora. LDA dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủ đề (topic). Về bản chất, LDA là một mô hình Bayesian 3 cấp (three-level hierarchical Bayes model: corpus level, document level, word level) trong đó mỗi phần của mô hình được coi như một mô hình trộn hữu hạn trên cơ sở tập các xác suất chủ đề.



Hình 2. Tài liệu với K chủ đề ẩn.

Ước lượng tham số cho mô hình LDA: Cho một corpus của M tài liệu biểu diễn bởi $D=\{d_1,d_2, \ldots, d_M\}$, trong đó, mỗi tài liệu m trong corpus bao gồm $n \times m$, từ w_i rút từ một tập từ vựng của các mục từ $\{t_1, \ldots, t_v\}$, V là số lượng các mục từ t trong tập từ vựng. LDA cung cấp

một mô hình sinh đầy đủ chỉ ra kết quả tốt hơn các phương pháp trước. Quá trình sinh ra văn bản như hình 3.



Hình 3 Ước lượng tham số cho tập dữ liệu.

Trong đó: Các khối vuông biểu diễn quá trình lặp.

Tham số đầu vào gồm α và β (corpus-level parameter); α: Dirichlet prior on $\vec{\vartheta}_m$ (theta); β: Dirichlet prior on $\vec{\varphi}_k$; $\vec{\vartheta}_m$ (theta): phân phối của topic trong document thứ m (document-level parameter); $z_{m,n}$: topic index (từ n của văn bản m); $w_{m,n}$: từ n của văn bản m chỉ bởi $z_{m,n}$ (word-level varible, observe word); $\vec{\varphi}_k$: phân phối của các từ được sinh từ topic $z_{m,n}$; M: số lượng các tài liệu; N_m: số lượng các từ trong tài liệu thứ m; K: số lượng các topic ẩn

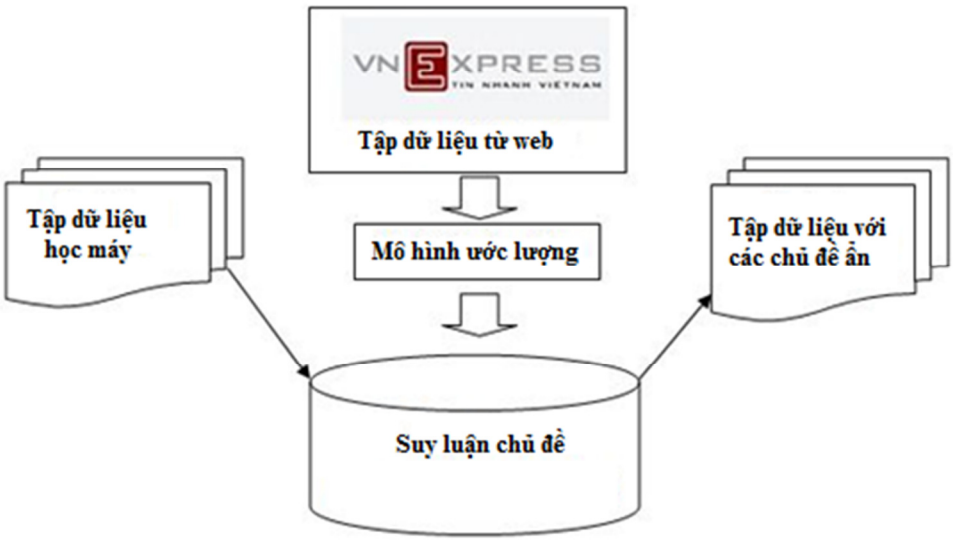
LDA sinh một tập các từ $w_{m,n}$ cho các văn bản $\vec{d}_m$ bằng cách:

- Với mỗi văn bản m, sinh ra phân phối topic $\vec{\vartheta}_m$ cho văn bản.
- Với mỗi từ $z_{m,n}$ được lấy mẫu dựa vào phân phối topic trên.
- Với mỗi topic index $z_{m,n}$ dựa vào phân phối từ $\vec{\varphi}_k$, $w_{m,n}$ được sinh ra
- $\vec{\varphi}_k$được lấy mẫu một lần cho toàn bộ corpus.

Ước lượng tham số cho mô hình LDA bằng thuật toán Gibbs Sampling, một thuật toán nhanh, đơn giản, và hiệu quả để huấn luyện LDA.

Suy luận chủ đề: Theo Nguyễn Cẩm Tú [15], với một mô hình chủ đề đã được huấn luyện tốt dựa trên tập dữ liệu toàn thể (Universial Dataset) bao phủ miền ứng dụng, ta có thể thực hiện một tiến trình quá trình suy diễn chủ đề cho các tài liệu mới tương tự như quá trình ước lượng tham số (tức là xác định được phân phối trên các chủ đề của tài liệu qua tham số theta). Trong bài báo này cũng chỉ ra rằng việc sử dụng dữ liệu từ trang vnexpress.net huấn luyện được các mô hình có ưu thế hơn trong các phân tích chủ đề trên dữ liệu tin tức. Trong khi đó, các mô hình được huấn luyện bởi dữ liệu từ Wiki tốt hơn trong phân tích chủ đề các tài liệu mang tính học thuật.

Dựa trên những nghiên cứu đó, bài báo này chọn mô hình chủ đề được huấn luyện bởi tập dữ liệu toàn thể thu thập từ trang vnexpress.net cho phân tích chủ đề. Một tiến trình phân tích chủ đề tổng quát được minh họa như sau:



Hình 4. Suy luận chủ đề cho các tin tức thu thập từ vnexpress.net.

Công cụ JGibbsLDAcủa Nguyễn Cẩm Tú [17] đã hiện thực quá trình ước lượng và suy luận chủ đề ẩn cho kết quả rất tốt, bài báo sử dụng công cụ này để xây dựng tập đặc trưng cho từng thể loại và thu được kết quả khả quan.

2.4. Phân loại văn bản

- **Phân loại văn bản sử dụng tần suất chủ đề**

Dữ liệu cần phân loại cũng phải được qua các bước tiền xử lí như dữ liệu học (tách từ, loại bỏ từ dừng, từ phổ biến) để thu được các từ đặc trưng cho văn bản cô đọng nhất mà vẫn thể hiện được đầy đủ ý nghĩa của văn bản. Lần lượt so sánh tần suất xuất hiện của từng chủ đề trên đặc trưng của văn bản vừa thu được. Tần suất của thể loại nào xuất hiện nhiều hơn thì thuộc thể loại đó.

- **Phân loại văn bản sử dụng hệ số Cosine**

Trong các phương pháp tính độ tương đồng, bài báo sử dụng phương pháp tính độ đo cosine do phương pháp này đơn giản, dễ cài đặt mà vẫn đạt quả như các phương pháp kia.

Thể hiện các từ đặc trưng đó trong mô hình không gian vector. Với văn bản d sau khi xử lí chỉ còn n từ, ta biểu diễn d dưới dạng vector $d = \{(w_1 : p_1), (w_2 : p_2), \dots, (w_n : p_n)\}$. Trong đó, w_i là từ thứ i, p_i là tần suất của từ w_i trong văn bản đó. Với p_i được tính bằng công thức:

$$p_i = \frac{N(w_i)}{\sum_{i=1}^{n} w_i}$$

với $N(w_i)$ là số lần xuất hiện của từ w_i trong văn bản; $\sum_{i=1}^{n} w_i$ là tổng số từ trong văn bản.

Các tập đặc trưng sau khi suy luận LDA cũng cho ra tập đặc trưng dạng vector $T = \{(w_{j1} : p_{j1}), (w_{j2} : p_{j2}), ..., (w_{jm} : p_{jm})\}$, trong đó w_{ji} là từ thứ i của tập đặc trưng thứ j, p_{ji} là tần suất của từ thứ w_{ji}, tần suất này tự sinh ra trong quá trình suy luận của LDA.

Với vector d và từng vector đặc trưng T_j có được, dùng công thức cosine tính độ tương đồng 2 vector, kết này dễ dàng so sánh độ tương đồng của thể loại nào lớn hơn thì thuộc thể loại đó.

Độ tương tự giữa chúng được tính theo công thức:

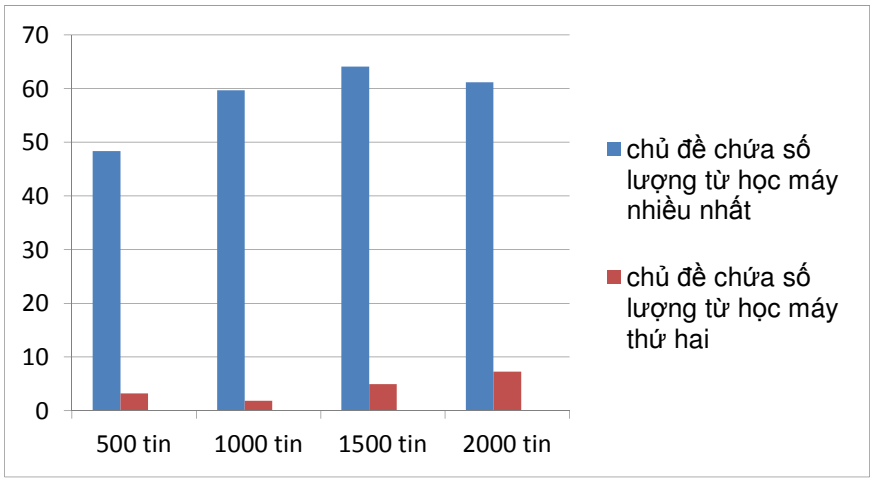
$$\text{Sim}(d, T_j) = \frac{d.T_j}{\|d\|.\|T_j\|}.$$

3. KẾT QUẢ THỰC NGHIỆM

Với dữ liệu gồm 11.185 văn bản đã được rút trích phục vụ cho việc phân tích chủ đề ẩn ước lượng và suy luận. Hơn 2.000 tin tức văn bản thu được từ trang báo điện tử vnexpress.net theo từng thể loại phục vụ cho việc học máy. 10 bộ dữ liệu (mỗi bộ 100 tin tức) thu thập từ báo điện tử vnexpress.net mới nhất thuộc 10 thể loại, phục vụ cho việc kiểm thử chương trình.

3.1. Dùng LDA tìm đặc trưng cho từng thể loại

Sau khi dùng LDA suy luận chủ đề, ứng với mỗi số lượng học máy, đều thu được một chủ đề chứa số lượng từ rất lớn trong toàn bộ số lượng từ học máy so với chủ đề chứa các từ còn lại.



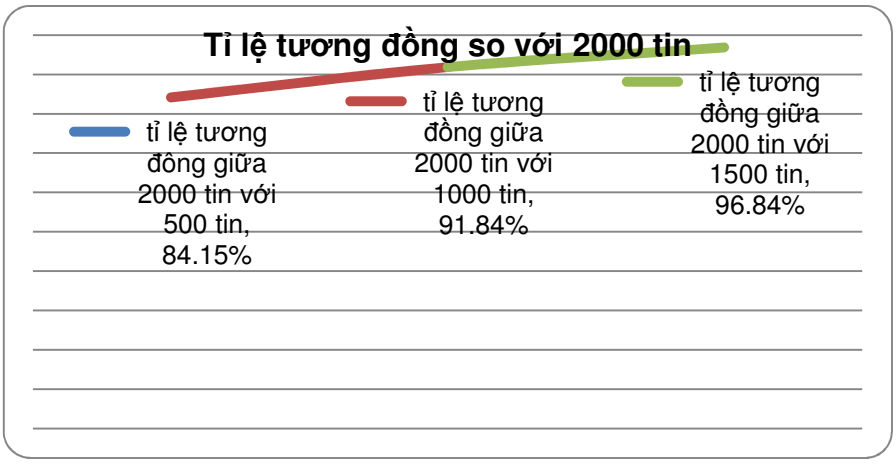
Hình 5. Biểu đồ tỉ lệ số lượng tin tức học máy thể loại kinh doanh.

Hình 5 cho thấy, có thể lấy các từ ở chủ đề cao nhất làm đặc trưng cho thể loại kinh doanh, và thu được tập đặc trưng của thể loại kinh doanh ứng với số lượng tin tức học máy.

*Bảng 1.*15/100 đặc trưng sau mỗi lần suy luận.

500 tin	1000 tin	1500 tin	2000 tin
ngân_hàng	ngân_hàng	ngân_hàng	ngân_hàng
tiền	tiền	doanh_nghiệp	doanh_nghiệp
doanh_nghiệp	doanh_nghiệp	tiền	mức
nhà_nước	công_ty	mức	tiền
chính_sách	nhà_nước	công_ty	công_ty
nợ	hàng	usd	usd
công_ty	chính_sách	thị_trường	nhà_nước
cổ_phần	nợ	nhà_nước	thị_trường
lãi_suất	mức	nợ	hàng
hàng	usd	hàng	kinh_tế
mức	thị_trường	chính_sách	lãi_suất
thẻ	lãi_suất	lãi_suất	nợ
tp_hcm	hà_nội	thuế	thuế
chính_phủ	khách_hàng	hà_nội	chính_sách
đà_nẵng	cổ_phần	lượng	nước

Lấy các đặc trưng sau khi học máy từ 2.000 tin tức làm chuẩn, lần lượt so sánh với các chủ đề của 500 tin, 1.000 tin, và 1.500 tin.



Hình 6. Biểu đồ độ tương đồng số lượng học máy của thể loại kinh doanh.

Kết quả cho thấy số lượng học máy càng cao thì tỉ lệ tương đồng càng lớn. Đến một lúc nào đó, các đặc trưng này sẽ bão hòa (ở mức bão hòa, có học thêm bao nhiêu đi nữa thì các đặc trưng này sẽ không thay đổi).

3.2. Phân loại văn bản

Với 10 bộ dữ liệu thuộc 10 thể loại thu thập từ vnexpress.net mới nhất (theo đúng thể loại do báo điện tử vnxpress.net đưa ra) đưa vào hệ thống phân loại với 2 phương pháp: sử dụng tần suất chủ đề và sử dụng hệ số Cosine. Lấy thể loại từ trang vnexpress.net làm chuẩn, kết quả phân loại so với chuẩn:

Bảng 2. Kết quả phân loại dùng tần suất chủ đề và hệ số Cosine.

Thể loại	Dùng tần suất	Hệ số Cosine
Đời sống	65	73
Khoa học	58	66
Kinh doanh	82	89
Ô tô – xe máy	89	90
Pháp luật	81	81
Thế giới	70	66
Thể thao	88	91
Văn hóa	91	89
Vi tính	80	86
Xã hội	50	40
Trung bình	**75,4%**	**77,1%**

Xét toàn hệ thống, kết quả phân loại dùng hệ số Cosine tốt hơn dùng tần suất chủ đề.

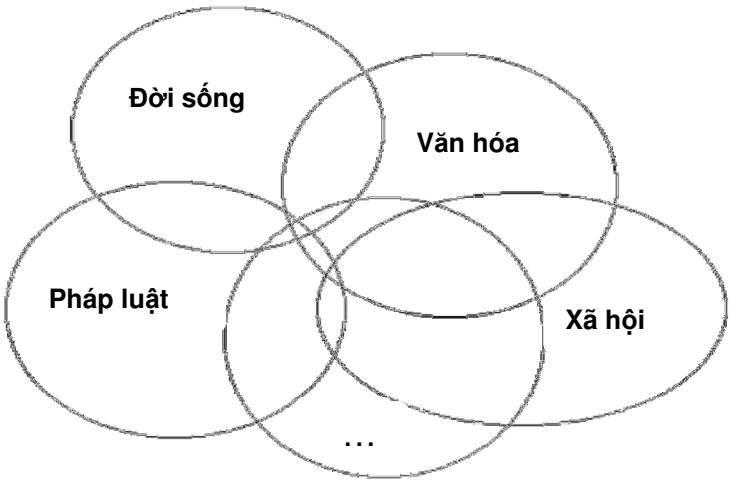
Chọn phương pháp dùng hệ số Cosine làm phương pháp chính cho hệ thống phân loại, tiếp tục xét chi tiết từng thể loại cho kết quả khác so với báo đã đưa. Thu được kết quả:

Bảng 3. Kết quả phân loại hệ thống so với báo.

Thể loại	Số tài liệu sai so với báo	Báo đưa khác mục	Hệ thống sai
Đời sống	27	18	9
Khoa học	34	21	13
Kinh doanh	11	6	5
Ô tô – xe máy	10	5	5
Pháp luật	19	8	11
Thế giới	34	22	12
Thể thao	9	4	5
Văn hóa	11	4	7
Vi tính	14	9	5
Xã hội	60	32	28
Tổng	**229**	**129**	**100**

Như vậy, khi sử dụng hệ số Cosine để tính độ tương đồng trong phân loại văn bản sẽ cho kết quả tốt hơn sử dụng tần suất chủ đề trên toàn bộ dữ liệu. Kết quả phân loại đạt độ chính xác so với dữ liệu mẫu là 77,1 %, trong số 22,9 % còn lại thì dữ liệu mẫu đưa sai là 12,9 % và hệ thống phân loại sai là 10 %. Suy ra, tỉ lệ trung bình độ chính xác của hệ thống đạt 90 %.

Kết quả 90 % là khả quan, trong khi các tập đặc trưng là các dữ liệu có dạng liên kết với nhau, vì thế có nhiều khả năng một tài liệu văn bản có thể thuộc 1 thể loại, 2 thể loại hoặc nhiều thể loại hệ thống sẽ gán tài liệu vào thể loại có hệ số cao nhất. Biểu diễn các tập đặc trưng như hình 7.



Hình 7. Các tập đặc trưng liên kết với nhau.

4.KẾT LUẬN

Bài báo trình bày các kết quả nghiên cứu về quy trình phân loại văn bản và áp dụng các thuật toán xử lí ngôn ngữ tự nhiên,sử dụng LDA để tìm tập đặc trưng, và đưa ra các độ đo để giải quyết bài toán phân loại văn bản tiếng Việt dựa vào đặc trưng. Bài toán là nền tảng cho nhiều ứng dụng quan trọng thực tế như lọc thư spam, rút trích văn bản, hệ thống khuyến cáo người dùng …

TÀI LIỆU THAM KHẢO

1. Rong Hu - Active Learning for Text Classification, Doctoral Thesis, Dublin Institute of Technology, 2011.
2. Thorsten Joachims - Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Lecture Notes in Computer Science **1398** (1998) 137-142.
3. Fabrizio Sebastiani - Text Categorization, In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109-129.
4. Tong Zhang and Frank J. Oles - Text Categorization Based on Regularized Linear Classifiacation Methods, Kluwer Academic Publishers, Manufactured in The Nethelands, April 2001.

5. Nguyễn Thị Thu Hằng - Phương pháp phân cụm tài liệu Web và áp dụng vào máy tìm kiếm, Luận văn Cao học, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2009.
6. Hoàng Kiếm, Đỗ Phúc- Phân loại văn bản dựa trên cụm từ phổ biến, Kỷ yếu Hội nghị Khoa học lần 2, Trường Đại học Khoa học Tự nhiên TP.HCM, 2002, pp. 109-113.
7. Đặng Vũ Tùng - Phân loại trang Web dựa trên phương pháp đồng huấn luyện, Luận văn Thạc sĩ kĩ thuật, Học viện Công nghệ Bưu chính Viễn thông Hà Nội, 2011.
8. Lưu Văn Tăng - Phát triển bộ công cụ hỗ trợ xây dựng kho ngữ liệu cho phân tích văn bản tiếng Việt, Luận văn Cao học, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, 2009.
9. Hệ tách từ tiếng Việt [online], http://vlsp.vietlp.org:8080/demo/?page=resources
10. Thông tin chi tiết: đề tài – dự án [online], http://vpct.gov.vn/News.aspx?ctl=projectdetail&ID=29
11. Trần Thị Oanh - Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt, Luận văn Cao học, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2008.
12. Hofmann T. - Probabilistic Latent Semantic Analysis, Uncertainity in Articial Intelligence, UAI'99, Stockholm, 1999.
13. David M. Blei, Andrew Y.Ng, Micheal I. Jordan - Latent Dirichlet Allocation, Journal of Machine Learning Research **3** (2003) 993-1022.
14. Matthew D. Hoffman, David M. Blei, Francis Bach - Online Learning for Latent Dirichlet Allocation, Advances in Neural Information Processing Systems **23** (2010), pp. 856-864.
15. Nguyễn Cẩm Tú - Hidden Topic Discovery toward Classification and Clustering in Vietnamese Web Documents, Luận văn Cao học, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2008.
16. Trần Mai Vũ - Tóm tắt văn bản dựa vào trích xuất câu, Luận văn Cao học, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 2009.
17. Công cụ phân tích chủ đề ẩn [online], http://jgibblda.sourceforge.net/

ABSTRACT

CLASSIFICATION OF VIETNAMESE WEB DOCUMENT

Tran Ngoc Phuc[1], PHAM Tran Vu[2], Pham Cong Xuyen[1], Nguyen Vu Duy Quang[1]

[1]*Department of Information Technology, Lac Hong University*

[2]*Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology*

Email: *tnphuc@lhu.edu.vn, t.v.pham@cse.hcmut.edu.vn, xuyen@lhu.edu.vn, quang@lhu.edu.vn*

This paper presents some research results on using Latent Dirichlet Allocation algorithm, which is about analyzing hidden topics exist in documents, to extract important features of web documents for classification. The features are represented as noun phrases extracted from document text using vector model. In this model, the each document is represented as a vector. The weight of each element of the vector is calculated from its occurence frequency. The classification is then measured based on the similarity of any two documents, which is calculated by the cosine of the two representing vectors. In this paper, Latent Dirichlet Allocation algorithm is used to extract hidden features of web documents for similarity calculation and gives very accurate results. A prototype application has been built, and the experiment results showed that the classification of news on Vietnamese websites had the accuracy of about 90%.

Keywords: text classification,Vietnamese language processing.