Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00065

PHÂN LOẠI VĂN BẢN TIẾNG VIỆT DỰA TRÊN MÔ HÌNH CHỦ ĐỀ

Bùi Khánh Linh [1], Nguyễn Thị Thu Hà[1], Nguyễn Thị Ngọc Tú[1], Đào Thanh Tĩnh[2]
[1] Khoa CNTT – Trường Đại học Điện lực – Hà Nội
[2] Khoa CNTT – Trường Đại học Lê Quý Đôn – Hà Nội
linbk@epu.edu.vn, hantt@epu.edu.vn, tunn@epu.edu.vn, tinhdt@mta.edu.vn

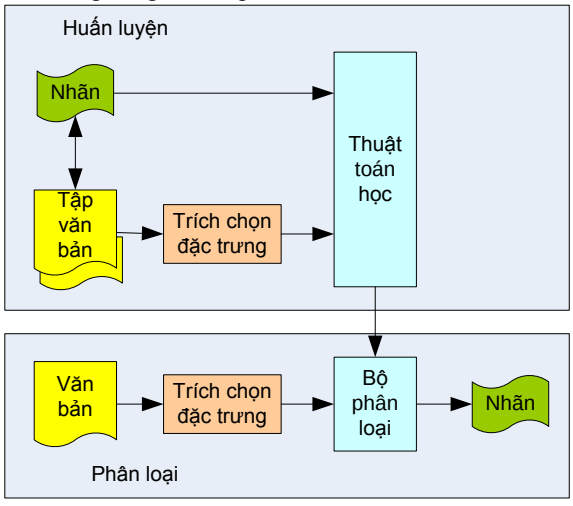
TÓM TẮT— *Trong bài báo này, chúng tôi đề xuất một giải pháp mới trong xử lý tiếng Việt bằng cách xây dựng mô hình chủ đề tiếng Việt. Phương pháp này sử dụng cách thức tìm một từ lõi và phát triển để tự sinh ra các từ khác trong chủ đề dựa trên Naive Bayes. Dựa trên tập dữ liệu huấn luyện, chúng tôi tính toán xác suất của các từ trong mô hình chủ đề tiếng Việt. Kết quả thực nghiệm cho thấy rằng, phương pháp của chúng tôi đề xuất có hiệu quả trong việc phân loại các văn bản tiếng Việt theo nhiều lớp chủ đề nhỏ hơn. Đồng thời, có độ chính xác khá cao và thời gian xử lý phân loại nhanh hơn so với các phương pháp đã được đề xuất trước đó.*

Từ khóa— *Mô hình chủ đề, tiếng Việt, khai phá văn bản, từ lõi, Naive Bayes*.

I. ĐẶT VẤN ĐỀ

Phân loại văn bản là một trong những phần quan trọng của việc khai phá dữ liệu văn bản, khá nhiều các hệ thống phân loại văn bản sử dụng kỹ thuật dựa trên tri thức (knowledge based) hoặc dựa trên các luật được xây dựng sẵn để tạo thành một tập hợp các quy tắc logic để hiểu và phân loại văn bản. Mỗi loại (hay còn gọi là lớp – class) tương đương với một chủ đề ví dụ "thể thao", "chính trị" hay "nghệ thuật". Nhiệm vụ phân loại được bắt đầu xây dựng từ một tập các văn bản $D=\{d_1,d_2,..,d_n\}$ được gọi là tập huấn luyện, trong đó các tài liệu d_i được gán nhãn c_j - với c_j thuộc tập các chủ đề $C=\{c_1,c_2,...,c_m\}$. Nhiệm vụ tiếp theo là xác định được mô hình phân loại, trên cơ sở đó có thể gán đúng lớp để một tài liệu d_k bất kỳ có thể phân loại chính xác vào một trong những chủ đề của tập chủ đề C [1],[2],[3],[6].

Bài toán phân loại văn bản được mô phỏng thành quá trình học như sau:



Hình 1. Quá trình học phân loại văn bản

Đối với những bài toán xử lý phân loại các đối tượng, việc quan trọng là xác định đặc trưng bởi hầu hết trong những bài toán này, số chiều đặc trưng là khá lớn. Bởi vậy, các đề xuất trước đây [4], [5],[7-11], [13] sẽ gặp phải những khó khăn sau:

- Thời gian tính toán lớn (do số chiều đặc trưng nhiều)
- Độ chính xác cũng như hiệu năng của hệ thống bị hạn chế.

Một khó khăn khác nữa trong cách xử lý phân loại tự động đối với các văn bản tiếng Việt, là độ khó trong xử lý ngôn ngữ, bởi ngôn ngữ tiếng Việt thuộc lớp ngôn ngữ đơn lập (single syllable language), các từ trong tiếng Việt có thể là từ đơn hoặc từ ghép, do vậy khó khăn trong việc tách từ. Bởi thế, chúng tôi đã tiếp cận bài toán theo hai bước: xử lý giảm đặc trưng và áp dụng lý thuyết Naive Bayes trong phân loại.

Xử lý giảm số chiều của đặc trưng bằng cách xây dựng mô hình chủ đề (topic modeling), số lượng thuật ngữ (term) trong mỗi chủ đề sẽ giảm hơn nhiều so với số các từ trong một văn bản, mặt khác sẽ giải quyết bài toán tách từ tiếng Việt nhờ đó làm tăng độ chính xác của hệ thống, tiếp theo áp dụng lý thuyết Naive Bayes để phân loại các văn bản theo đúng chủ đề đã chọn [12].

Phần II của bài báo trình bày phương pháp tiếp cận và giải quyết bài toán phân loại văn bản tiếng Việt dựa trên mô hình chủ đề và lý thuyết Bayes. Phần III của bài báo trình bày cách thức thử nghiệm dựa trên phương pháp đã được đề xuất tại phần II và cuối cùng là kết luận.

II. PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN TIẾNG VIỆT DỰA TRÊN MÔ HÌNH CHỦ ĐỀ VÀ LÝ THUYẾT NAIVE BAYES

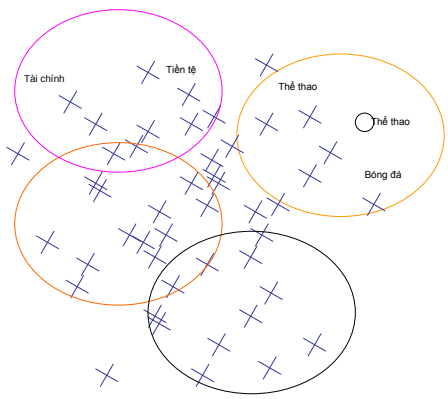
2.1. Xây dựng mô hình chủ đề

Khái niệm mô hình chủ đề được Griffiths và Steyvers đưa ra lần đầu tiên vào những năm 2002, 2003. Tiếp theo vào năm 2007, Griffiths và các cộng sự đã sinh ra một mô hình xác suất cho văn bản dựa trên mô hình phân phối ẩn Dirichlet (LDA). Nó được mô tả là một loại mô hình thống kê để phát hiện ra các "chủ đề" trừu tượng có trong một tập các tài liệu. Khi xem xét tài liệu bất kỳ, sự xuất hiện của các từ trong tài liệu đó sẽ gợi ý cho người đọc về một chủ đề liên quan, nó cũng có thể xuất hiện nhiều ở một tài liệu khác mà [12].

Bảng 1. Các từ chủ đề trong tập mô tả của Andrews năm 2009

Theatre	Music	League	Prison	Rate	Pub	Market	Railway	Air
Stage	Band	Cup	Years	Cent	Guinness	Stock	Train	Aircraft
Arts	Rock	Season	Sentence	Inflation	Beer	Exchange	Station	Flying
Play	Song	Team	Jail	Recession	Drink	Demand	Steam	Flight
Dance	Record	Game	Home	Recovery	Bar	Share	Rail	Plane
Opera	Pop	Match	Prisoner	Economy	Dringking	Group	Engine	Airport
cast	dance	division	serving	cut	alcohol	news	track	pilot

Với bảng trên, mỗi cột mô tả cho một chủ đề riêng biệt. Các nghiên cứu trước xây dựng mô hình các từ chủ đề dựa trên Bayes hay mô hình Markov ẩn. Trong bài báo này, chúng tôi lựa chọn cách thức xây dựng tập từ chủ đề dựa trên mô hình xác suất điều kiện dựa trên tập dữ liệu huấn luyện. Tập dữ liệu huấn luyện này gồm các văn bản đã được phân loại trước bởi con người và được gán nhãn vào đúng chủ đề thích hợp. Hình 2 dưới đây minh họa một số chủ đề trong tập không gian gồm n chủ đề khác nhau. Trong đó, các ký hiệu o là ký hiệu biểu diễn từ lõi (core term) và ký hiệu x là ký hiệu biểu diễn các từ chủ đề trong không gian n chủ đề.



Hình 2. Mô hình chủ đề dựa trên xác suất

Giả sử A={A1,...,Ak} là một không gian k chiều các chủ đề. Mỗi không gian Ai bao gồm tập các từ thuộc nó nếu như khả năng xuất hiện của nó trong Ai là khác 0. Các không gian Ai và Aj có thể giao nhau, do vậy, các từ thuộc Ai có thể cũng thuộc một không gian Aj khác.

Giả sử ta lấy một từ gọi là từ lõi (core term) của không gian Ai (từ này được coi là từ có trọng số cao nhất), khoảng cách của các từ còn lại trong không gian Ai chỉ cần so với từ lõi. Để tính được khoảng cách của các từ đó so với lõi, chúng tôi sử dụng cách tính xác suất có điều kiện. Trên thực tế, ta xây dựng mô hình chủ đề theo phương pháp xác suất điều kiện theo những bước sau:

Tập văn bản huấn luyện gồm n văn bản D={d1, d2, ..., dn}.

Đối với mỗi văn bản được phân vào từng chủ đề C={c1, c2, ...,cm}.

Sử dụng VnTagger [14] để tách các từ trong D và trích rút ra tập các danh từ N.

Tính tần suất xuất hiện lớn nhất của 1 danh từ đối với mỗi 1 chủ đề, gọi là từ lõi (core).

Tính xác suất có điều kiện các từ còn lại với các từ core, từ đó sẽ thuộc chủ đề nào mà có xác suất điều kiện với từ core là khác 0.

Dưới đây là thuật toán mô tả phương pháp xây dựng mô hình chủ đề.

THUẬT TOÁN XÂY DỰNG MÔ HÌNH CHỦ ĐỀ

Đầu vào:

- D: Tập văn bản huấn luyện đã được gán nhãn tương ứng với các chủ đề C;
- VnTagger: Công cụ nhận dạng, tách từ;
- C: Tập các chủ đề

Đầu ra:

- T: Tập các từ được gán nhãn tương ứng với mỗi C.

Khởi tạo:

V= ϕ ; N=0; n=0;

For each di in Ck do

Vk ← Vntagger(di); // nhận diện các danh từ trong mỗi di và đưa vào tập danh từ V

For each Ck do

If w(j) ∈ Vk then // Nếu từ wj thuộc tập danh từ V

n(j) ← n(j) +1; // đếm số lần xuất hiện w(j) trong mỗi chủ đề Ck

Nk=argmax(n(j)); // Lấy tần suất lớn nhất của từ wj trong mỗi chủ đề Ck

For each Ck do

For all w in V

if Pr(w(i)|Ni) <>0 then Vk ← w(i); // cho các từ w(i) vào tập Vk của Ck

2.2. Phân loại văn bản tiếng Việt với mô hình chủ đề và Naive Bayes

Sau khi xây dựng được tập từ chủ đề đối với mỗi một lớp chủ đề. Tiếp theo sử dụng phân loại Naive Bayes để xây dựng mô hình phân loại tự động.

Ý tưởng: Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng…

Cài đặt:

Mô tả vector đặc trưng của văn bản: Là vector có số chiều là số đặc trưng trong toàn tập dữ liệu, các đặc trưng này đôi một khác nhau. Nếu văn bản có chứa đặc trưng đó sẽ có giá trị 1, ngược lại là 0.

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

Huấn luyện: tính $P(C_i)$ và $P(x_k|C_i)$

Đầu vào:

- Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
- Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

Đầu ra:

- Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

Công thức tính $P(C_i)$ đã làm trơn Laplace

$$P(C_i) = \frac{|docs_i| + 1}{|total\ docs| + m}$$

Trong đó:

|docsi|: số văn bản của tập huấn luyện thuộc phân lớp i.

|total docs|: số văn bản trong tập huấn luyện.

m số phân lớp.

Khởi tạo mảng A, B có kích thước m.

Duyệt qua các văn bản trong tập dữ liệu, đếm số văn bản trong mỗi phân lớp lưu vào A.

Tính xác suất cho từng phân lớp theo công thức trên và lưu vào mảng B.

Công thức tính $P(x_k|C_i)$ đã làm trơn Laplace:

$$P(x_k|C_i) = \frac{|docs_{x_k i}| + 1}{|docs_i| + d_k}$$

Trong đó:

$|docs_{x_k i}|$: Số văn bản trong trong phân lớp i có đặc trưng thứ k mang giá trị xk. (hay số văn bản trong lớp i, có xuất hiện/không xuất hiện đặc trưng k).

$|docs_i|$: Số văn bản của tập huấn luyện thuộc phân lớp i.

d_k: Số giá trị có thể có của đặc trưng thứ k.

Với vector đặc trưng như mô tả bên trên, d_k ở đây mang giá trị là 2, tương ứng với xuất hiện và không xuất hiện. Do chỉ có 2 giá trị, ta có thể tính nhanh xác suất không xuất hiện theo công thức $P(\bar{x}) = 1 - P(x)$.

Khởi tạo mảng 3 chiều C, chiều 1 có kích thước là m (số phân lớp), chiều 2 có kích thước là N (số đặc trưng), chiều 3 có kích là 2 (dk) để lưu các giá trị $P(x_k|C_i)$.

Duyệt qua các văn bản trong tập dữ liệu, tiến hành thống kê các chỉ số cần thiết để tính xác suất $P(x_k|C_i)$ theo công thức trên và lưu vào mảng C.

Phân lớp:

Đầu vào:

Vector đặc trưng của văn bản cần phân lớp.

Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

Đầu ra:

Nhãn/lớp của văn bản cần phân loại.

Công thức tính xác suất thuộc phân lớp i khi biết trước mẫu X

$$P(C_i|X) = P(C_i)\prod_{k=1}^{n} P(x_k|C_i)$$

Dựa vào vector đặc trưng của văn bản cần phân lớp, áp dụng công thức trên tính xác suất thuộc từng phân lớp cho văn bản, và chọn ra lớp có xác suất cao nhất.

III. KẾT QUẢ THỰC NGHIỆM, THẢO LUẬN

3.1. Số liệu đầu vào

3.1.1. Xây dựng tập ngữ liệu

Chúng tôi thực nghiệm trên tập văn bản tiếng Việt. Tài liệu sử dụng để xây dựng kho từ chủ đề là các văn bản đã được gán nhãn theo chủ đề. Cho đến thời điểm này, kho ngữ liệu chuẩn phục vụ cho xây dựng kho từ chủ đề cho tiếng Việt vẫn chưa có. Do đó, ta phải xây dựng kho dữ liệu này một cách thủ công bằng cách tìm kiếm các văn bản trên các nguồn thông tin như: http://vnexpress.net, http://vietnamnet.vn,... Tập văn bản đầu vào là văn bản dạng thô, để đơn giản cho việc xử lý dữ liệu, với mỗi văn bản đầu vào, ta sẽ thực hiệc qua bước tiền xử lý ký tự để đưa văn bản về dạng xâu chuẩn. Để đảm bảo tính chính xác cao, các văn bản được xử lý rất thủ công và tỉ mỉ sau đó lưu lại vào 1 file dữ liệu txt và được gán nhãn theo chủ đề. Các file dữ liệu này được sử dụng trong quá trình huấn luyện tiếp theo.

3.1.2. Xây dựng mô hình chủ đề

Trong các văn bản huấn luyện, phần tách từ được sử dụng công cụ gán nhãn từ loại VnTagger, công cụ này sử dụng kho ngữ liệu với 20,000 câu đã được gán nhãn từ loại do nhóm xử lý ngôn ngữ tự nhiên tiếng Việt phát triển nằm trong nhánh đề tài KC01.01/06-10 [14].

Dữ liệu trong mô hình chủ đề bao gồm các danh từ, do vậy sử dụng công cụ VnTagger để tách ra các danh từ trong tập dữ liệu đã xây dựng, sau đó ta tiến hành xây dựng tập từ riêng đối với mỗi chủ đề khác nhau.

Để xây dựng tập các từ chủ đề đối với mỗi mục chủ đề, cần xác định 1 từ lõi đối với mỗi chủ đề. Sau đó tính xác suất có điều kiện của các từ còn lại so với các từ lõi để xác định các danh từ đó thuộc chủ đề nào.

Bảng 2 dưới đây mô tả một số chủ đề và từ chủ đề đã được xây dựng bằng phương pháp của bài báo đề xuất.

Bảng 2. Danh sách một số chủ đề đã được xây dựng

TÊN CHỦ ĐỀ					
Nghệ thuật	Thể thao	Công nghệ	Thị trường	Tài chính	Nhà đất
Dân ca	Bóng đá	Lõi tứ	Giá	Cán cân	Bất động sản
Nghệ sĩ	Bóng chày	Tablet	Thực phẩm	Ngân hàng	Nhà đất
Showbiz	Cầu thủ	Điện thoại	Chứng khoán	Lãi suất	Lãi suất
Người mẫu	Thủ môn	Smartphone	Chỉ số	Tỉ lệ	Biệt thự
Ảnh	Cup	Iphone	Lương	Cắt giảm	Chung cư
Sân khấu	Tỉ số	Samsung	Người mua	Tài chính	Chủ thầu
Ca nhạc	Chelsea	Transformer	Hàng hóa	Chứng khoán	Bất động sản

3.2. Phương pháp, công cụ mô phỏng

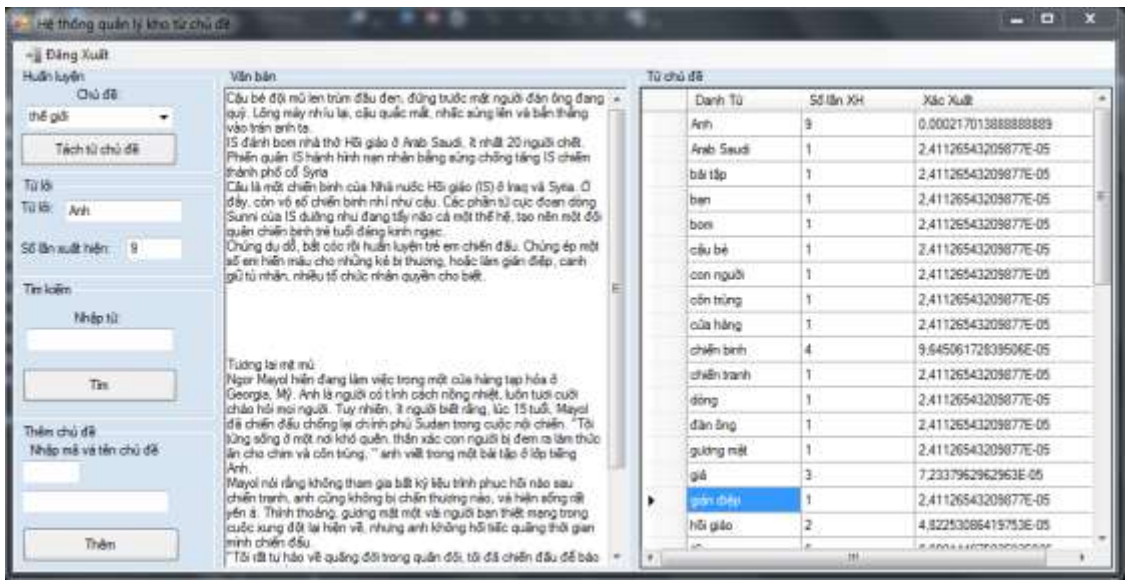
Cách đánh giá phân loại văn bản tập trung vào đánh giá thực nghiệm chứ không sử dụng cách đánh giá dựa trên phân tích lý thuyết. Các đánh giá thực nghiệm của phân loại văn bản sử dụng độ đo chính xác (precision). Ngoài ra, chúng tôi còn lấy số đặc trưng thô trung bình của n văn bản thử nghiệm so sánh với phương pháp đã được giảm bớt số đặc trưng thông qua việc xây dựng mô hình chủ đề.

Số đặc trưng trung bình được tính bằng tổng số các đặc trưng trong các văn bản thử nghiệm trên tổng số văn bản thử nghiệm.

Trong đánh giá thực nghiệm chúng tôi so sánh với phương pháp truyền thống là cách thức chỉ loại đi các từ dừng, các từ vô nghĩa trong văn bản và phương pháp dựa trên mô hình chủ đề.

3.3. Kết quả mô phỏng và bình luận

Để thử nghiệm chúng tôi sử dụng 220 văn bản với 6 chủ đề khác nhau: Nghệ thuật, thể thao, công nghệ, thị trường, tài chính, nhà đất. Trong đó có chủ đề thị trường và tài chính là lĩnh vực tương đối giống nhau.



Hình 3: Bộ từ chủ đề sau khi huấn luyện

Bảng 3. Kết quả thực nghiệm

Chủ đề	*Số văn bản thử nghiệm*	*Phương pháp truyền thống*		*Phương pháp dựa trên mô hình chủ đề*	
		Số đặc trưng TB	*Độ chính xác*	*Số đặc trưng TB*	*Độ chính xác*
Nghệ thuật	50	1120	86%	435	91.6%
Thể thao	30	835	88%	251	96%
Công nghệ	40	456	85.4%	216	97%
Thị trường	25	727	78%	304	93%
Tài chính	30	883	80.33%	378	94.8%
Nhà đất	45	954	82%	452	92%

Dựa trên cách đánh giá sử dụng độ đo chính xác và cách so sánh với phương pháp truyền thống thấy có sự giảm chiều rõ rệt các đặc trưng, số lượng các đặc trưng sau khi xây dựng mô hình chủ đề giảm còn 40.9% so với số lượng đặc trưng ban đầu trên tổng số 220 văn bản thực nghiệm (6 chủ đề khác nhau). Độ chính xác trung bình trên 6 chủ đề cũng tăng từ 83% lên tới 94.07%.

IV. KẾT LUẬN

Mô hình chủ đề được áp dụng vào nhiều các bài toán xử lý ngôn ngữ tự nhiên trên thế giới, dựa trên mô hình chủ đề này, các công cụ khai phá văn bản được xây dựng và đảm bảo tính ổn định, độ chính xác cao cũng như giảm thiểu chi phí về mặt thời gian xử lý so với những dữ liệu thô. Tuy nhiên, với cách thức xây dựng sử dụng các mô hình học xác suất như HMM hay Naive Bayes mang lại sự tốn kém về mặt chi phí cũng như thời gian khi xây dựng.

Trong bài báo này, chúng tôi sử dụng một cách tiếp cận khác để xây dựng mô hình chủ đề, giảm bớt được thời gian cũng như chi phí, đặc biệt đối với ngôn ngữ tiếng Việt hiện nay chưa xây dựng được mô hình chủ đề, là một trong những giải pháp giúp giải quyết những bài toán xây dựng các công cụ khai phá trên văn bản tiếng Việt.

Với mô hình chủ đề chúng tôi đã xây dựng, chúng tôi đã tiến hành thử nghiệm với công cụ phân loại văn bản, các kết quả thực nghiệm đã cho thấy sự hiệu quả của phương pháp này, các lớp văn bản được phân loại thành lớp nhỏ hơn, và số chiều của đặc trưng giảm tới hơn 50% so với số đặc trưng lúc ban đầu chưa xử lý.

Lời cảm ơn: *Nhóm tác giả trân trọng cảm ơn sự giúp đỡ về ý tưởng của TS. Nguyễn Lê Minh, TS. Nguyễn Phương Thái, TS. Nguyễn Văn Vinh – Đại học Công nghệ, Đại học Quốc gia Hà Nội đã hỗ trợ và đóng góp giúp chúng tôi những cách tiếp cận trong vấn đề giảm chiều đặc trưng trong xử lý những bài toán dữ liệu lớn.*

TÀI LIỆU THAM KHẢO

[1] C. Apte, F. Damerau, S. Weiss. Automated Learning of Decision Rules for Text Categorization,ACM Transactions on Information Systems, 12(3), pp. 233–251, 1994.

[2] S. Bhagat, G. Cormode, S. Muthukrishnan. Node Classification in Social Networks, Book Chapter inSocial Network Data Analytics, Ed. Charu Aggarwal, Springer, 2011.

[3] Ana Cardoso-Cachopo, Arlindo L. Oliveira, An Empirical Comparison of Text Categorization Methods, Lecture Notes in Computer Science, Volume 2857, Jan 2003, Pages 183 - 196.

[4] Han X., Zu G., Ohyama W., Wakabayashi T., Kimura F., Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination, LNCS, Volume 3309, Jan 2004, pp. 463-468.

[5] Novovicova J., Malik A., and Pudil P., "Feature Selection Using Improved Mutual Information for Text Classification", SSPR&SPR 2004, LNCS 3138, pp. 1010–1017, 2004.

[6] Sebastiani F., "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34 (1),2002, pp. 1-47. [26] Shanahan J. and Roma N., Improving SVM Text Classification Performance through Threshold Adjustment, LNAI 2837, 2003, 361-372.

[7] Soucy P. and Mineau G., "Feature Selection Strategies for Text Categorization", AI 2003, LNAI 2671, 2003, pp. 505-509.

[8] Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F., "Feature Selection Algorithms to Improve Documents Classification Performance", LNAI 2663, 2003, pp. 288-296.

[9] Torkkola K., "Discriminative Features for Text Document Classification", Proc. International Conference on Pattern Recognition, Canada, 2002.

[10] Vinciarelli A., "Noisy Text Categorization, Pattern Recognition", 17th International Conference on (ICPR'04) , 2004, pp. 554-557.

[11] Zu G., Ohyama W., Wakabayashi T., Kimura F., "Accuracy improvement of automatic text classification based on feature transformation": Proc: the 2003 ACM Symposium on Document Engineering, November 20-22, 2003, pp.118-120.

[12] Mark Steyvers, Tom Griffiths, Probabilistic Topic Models, In: In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds),Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum.

[13] Ha Nguyen Thi Thu ; Quynh Nguyen Huu ; Khanh Nguyen Thi Hong ; Hung Le Manh, Optimization for Vietnamese text classification problem by reducing features set, Information Science and Service Science and Data Mining (ISSDM), 2012 6th IEEE International Conference on New Trends in , Page(s): 209 – 212.

[14] http://vlsp.vietlp.org:8080

VIETNAMESE TEXT CLASSIFICATION BASED ON TOPIC MODELING

Bui Khanh Linh, Nguyen Thi Thu Ha, Nguyen Thi Ngoc Tu, Dao Thanh Tinh

ABSTRACT*— In this paper, we present a method for Vietnamese text classification based on topic modeling. This method is used to find a way from the core and development to other words in the subject based on Naïve Bayes theory. The experimental results, our method really effectively, high accuracy and can reduce complex of calculating. This method process faster than proposed methods.*

Keywords*— Topic modeling, Vietnamese text, data mining, core term, naïve bayes.*