PHƯƠNG PHÁP PHÂN CỤM TỪ TIẾNG VIỆT DỰA TRÊN PHƯƠNG PHÁP DENDROGRAMVÀ WIKIPEDIA

VIETNAMESE WORDS CLUSTERING METHOD BASED ON DENDROGRAM AND WIKIPEDIA

Nguyễn Thị Lệ Quyên, Phạm Minh Tuấn

Trường Đại học Bách khoa, Đại học Đà Nẵng; Email: quyen09t1@gmail.com, pmtuan@dut.udn.vn

Tóm tắt - Ngày nay, cùng với phát triển thông tin một cách nhanh chóng, việc phân loại văn bản tự động đang là một vấn đề cấp thiết. Nhiều phương pháp học máy như cây quyết định, mạng nơron nhân tạo hay máy vector hỗ trợ được áp dụng cho tiếng Anh và mang lại hiệu quả cao.Tuy nhiên các phương pháp này lại gặp khó khăn khi áp dụng cho phân loại tiếng Việt vì tiếng Việt có rất nhiều từ đồng nghĩa nhưng cách biểu diễn khác nhau. Báo cáo này đề xuất phương pháp phân cụm các từ tiếng Việt dựa vào tần số xuất hiện cùng nhau trên một trang Wikipedia tiếng Việt nhằm rút gọn vector thuộc tính của văn bản. Báo cáo này đồng thời đề xuất sử dụng phương pháp phân tích nhóm (Cluster Analysis) sử dụng đồ thị dendrogram trong việc phân cụm các từ Tiếng Việt. Kết quả thực nghiệm cho thấy phương pháp đề xuất đã phân cụm đúng các từ đồng nghĩa và các từ có chung một chủ đề.

Từ khóa - Văn bản tiếng Việt, Phân cụm từ, Phân tích nhóm, dendrogram, wikipedia

Abstract - Nowadays, within the development of quick information technology, the automatic document classification is an urgent issue. Many machine learning methods such as decision trees, artificial neural networks and support vector machines are applied to classify English documents and bring high efficiency. However, these methods are difficult to apply to classify Vietnamese documents because Vietnamese has many synonyms but performing different ways. This paper proposed a Vietnamese word clustering methods based on frequency appearing together on a Vietnamese Wikipedia page to shortened the length of feature vector of the document. This paper also proposed methods using cluster analysis based on graph clustering dendrogram. The experimental results show that the proposed method has the correct clustering of the synonyms and the words with a common theme.

Key words - Vietnamese documents; words clustering; cluster analysis; dendrogram; wikipedia

1. Đặt vấn đề

Ngày nay, việc trao đổi thông tin hầu hết đều dưới dạng văn bản như: thời sự, tư liệu, tài liệu, kết quả nghiên cứu khoa học… Cùng với việc phát triển tri thức cũng như toàn cầu hóa về internet, số lượng văn bản này ngày càng được gia tăng và lan truyền rộng rãi một cách nhanh chóng. Tuy nhiên, trong quá trình lan truyền và cập nhật thông tin một cách nhanh chóng này, các thông tin được lưu trữ (dưới dạng tài liệu số) cũng ngày càng tăng và rất khó khăn trong việc sắp xếp hay truy vấn tài liệu nếu không được phân loại một cách hợp lý.

Phân loại văn bản là một vấn đề quan trọng trong lĩnh vực xử lý ngôn ngữ. Nhiệm vụ của bài toán là phân loại các tài liệu vào các nhóm chủ đề cho trước. Đây là bài toán thường gặp trong thực tế như phân loại các tài liệu theo từng chủ đề (pháp luật, trính trị, giáo dục, thể thao,…) khác nhau. Việc tìm kiếm thông tin dễ dàng và nhanh chóng hơn khi các văn bản đã được phân loại. Tuy nhiên quá trình phân loại tiêu tốn thiều thời gian và chi phí nếu làm một cách thủ công. Vì vậy, thực hiện việc phân loại tự động văn bản số hiện nay là một vấn đề cấp thiết.

Để giải quyết vấn đề trên, có nhiều phương pháp học máy như cây quyết định [1], mạng nơron nhân tạo hay máy vector hỗ trợ đã được áp dụng vào bài toán phân loại văn bản tự động [1][2][3][4][5] một cách khá hiệu quả. Các phương pháp phân loại này thông thường sử dụng mô hình không gian vector (Vector space model - VSM) [2][6][7][8] nhằm trích chọn đặc tính cho văn bản huấn luyện cũng như văn bản cần phân loại. Đặc trưng của phương pháp này chính là tìm mối tương quan giữa hai văn bản hay giữa văn bản và câu truy vấn dựa trên các vector thuộc tính. Ví dụ, mỗi thuộc tính trong vector có thể được tính bằng tần số suất hiện của một từ trong văn bản. Phương pháp sử dụng hàm Cosine hay TF-IDF (term frequency – inverse document frequency) [1] là một trong số các phương pháp VSM thông dụng có thể kể tới. Từ kết quả phương pháp VSM này, các mô hình xác suất được xây dựng thông qua học máy (Machine Learning) nhằm mục đích phân loại văn bản một cách tự động.

Trong nghiên cứu này, tác giả chú trọng vào các vấn đề trích chọn đặc tính trong phân loại văn bản tiếng Việt [2][3][9]. Vấn đề được đặt ra là trong tiếng Việt có rất nhiều từ đồng nghĩa nhưng cách viết các ký tự lại khác nhau trên văn bản số. Ví dụ như, nghĩa các từ "khủng khiếp", "kinh khủng" và "kinh hoàng" rất tương đồng nhưng khi so sánh về mặt ký tự thì không giống nhau, dẫn tới các văn bản cùng nghĩa nhưng khác về cách viết sẽ có hệ số hàm tương quan thấp. Ngoài ra, trong tiếng Việt cũng có rất nhiều nhóm từ thường xuất hiện đi kèm cùng nhau trong một văn bản. Ví dụ như từ "nhồi máu" thường đi với từ "cơ tim" trong một văn bản. Đối với những văn bản có những nhóm từ này trong đó nó sẽ dễ có hệ số tương quan cao trong khi có thể không cùng thể loại, dẫn tới việc học và phân loại văn bản không hiệu quả.

Để tránh các tường hợp về đa dạng cách biểu diễn từ đồng nghĩa hay tồn tại các nhóm từ thường đi kèm cùng nhau trong một văn bản, tác giả đề xuất phương pháp phân cụm các từ tiếng Việt dựa vào tần số xuất hiện cùng nhau của một cặp từ trên một trang Wikipedia [10] tiếng Việt (số trang Wikipedia có chứa đồng thời cả hai từ). Các từ nằm trong một cụm có thể được coi như một thuộc tính trong văn bản. Nhờ vậy có rút gọn vector thuộc tính của văn bản hơn so với cách thức sử dụng mỗi từ cho một thuộc tính. Báo cáo này đồng thời đề xuất sử dụng phương pháp phân tích nhóm (Cluster Analysis) sử dụng đồ thị dendrogram [11][12] trong việc phân cụm các từ tiếng Việt.

2. Các phương pháp phân cụm

Có rất nhiều phương pháp phân cụm có thể kể tới như k-means hay Fuzzy c-means [13][14]. Tuy nhiên, đầu ra

của các thuật toán này phụ thuộc vào các vector đầu vào của các đối tượng cần phân cụm. Đối với bài toán phân cụm các từ tiếng Việt, việc định nghĩa các từ này thành vector chưa được nghiên cứu phổ biến, dẫn tới việc sử dụng k-means hay fuzzy c-means là không hợp lý.

Bài báo này sử dụng phương pháp phân tích nhóm (Cluster Analysis) [11][12]nhằm phân cụm các đối tượng dữ liệu giống nhau hoặc có hệ số tương quan cao. Có rất nhiều phương pháp phân tích nhóm, tuy nhiên bài báo này chỉ giới hạn sử dụng phương pháp đồ thị phân tầng dendrogram nhằm phân cụm các từ tiếng Việt. Phương pháp dendrogram là một phương pháp xây dựng sơ đồ dạng cây được sử dụng để minh họa cho sự sắp xếp các cụm đã được phân cụm theo tầng.

Thuật toán xây dựng đồ thị dendrogram tổng quát được trình bày như sau:

Bước 1. Đặt tất cả các dữ liệu thành từng nhóm riêng lẻ. Gọi mỗi dữ liệu là một nhóm.

Bước 2. Từ ma trận khoảng cách các nhóm, gom hai nhóm có khoảng cách gần nhất thành một nhóm.

Bước 3. Nếu số lượng nhóm là một thì kết thúc. Ngược lại thì thực hiện Bước 4.

Bước 4. Tính khoảng cách nhóm vừa được tạo ở Bước 2 với các nhóm còn lại và cập nhật ma trận khoảng cách.

Bước 5. Quay lại Bước 2.

Có rất nhiều phương pháp tính khoảng cách giữa hai nhóm. Dựa theo tính chất của từng dữ liệu, ta có các phương pháp tính khoảng cách sau:

1. **Nearest neighbor method:** Khoảng cách giữa hai nhóm được tính bởi khoảng cách nhỏ nhất trong tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

2. **Furthest neighbor method:** Khoảng cách giữa hai nhóm được tính bởi khoảng cách lớn nhất trong tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

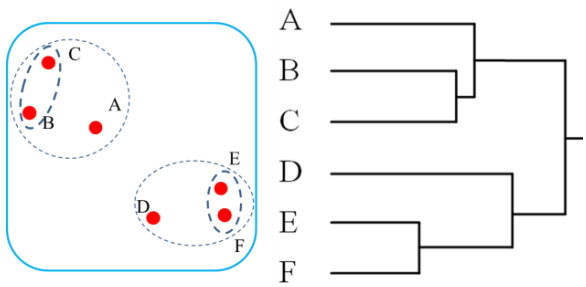
3. **Group average method:** Khoảng cách giữa hai nhóm được tính bởi khoảng cách trung bình của tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

4. **Centroid method:** Khoảng cách giữa hai nhóm được tính bởi khoảng cách của trọng tâm của hai nhóm.

5. **Wards method:** Khoảng cách giữa hai nhóm được tính bởi tổng bình phương khoảng cách của tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

Khoảng cách ở đây có thể được tính bằng nhiều cách khác nhau. Nếu các dữ liệu được thể hiện bằng các vector hay các điểm trong không gian Euclide thì ta có thể sử dụng khoảng cách Euclide hay khoảng cách Minkowski để tính. Tuy nhiên tùy theo tính chất của bài toán hay dữ liệu mà chúng ta có thể định nghĩa khoảng cách bằng các phương pháp khác như sử dụng khoảng cách Manhattan, khoảng cách Mahalanobis, xác suất, hệ số tương quan, v.v… Đối với văn bản thì ta còn có thể tính khoảng cách dựa theo hệ số tương quan về từ, về cấu trúc câu, về ngữ nghĩa của hai văn bản. Bài báo này sử dụng xác suất xuất hiện cùng nhau trên một văn bản để tính khoảng cách giữa hai từ trong tiếng Việt.

Hình 1 là một ví dụ cách xây dựng đồ thị dendrogram dựa trên phương pháp Nearest neighbor method với khoảng cách Euclide. Hình 1, bên trái là các vector “A”, ”B”, “C”, ”F”, “E”, ”F”, trong trong không gian 2 chiều. Ta thấy “E” và “F” có khoảng cách nhỏ nhất nên được gom thành một nhóm gồm hai phần tử. Tương tự ta cũng có “B” và “C” cũng được gom thành một nhóm. Từ các nhóm nhỏ, ta lại có được các nhóm lớn hơn nhờ việc gom các nhóm nhỏ lại với nhau. Ta được các nhóm “A,B,C” và nhóm “D,E,F”. Kết quả là cuối cùng tất cả các đối tượng được gom lại thành một nhóm.



Hình 1. *Ví dụ về đồ thị dendrogram*

3. Phương pháp đề xuất

Trong báo cáo này, nhóm tác giả đề xuất kết hợp Wikipedia và phương pháp phân tích nhóm dựa trên đồ thị dendrogram nhằm phân cụm tự động cho từ tiếng Việt. Wikipedia là một bách khoa toàn thư mở với nhiều ngôn ngữ thể hiện dưới một website trên internet [10]. Bài báo này sử dụng 1.184.476 trang Wikipedia tiếng Việt, được Wikipedia lưu trữ và cập nhật tại thời điểm ngày 01 tháng 01 năm 2014. Tất cả dữ liệu này được lưu theo định dạng file xml và có kích thước là 91.8GByte.

Phương pháp đề xuất sử dụng một bộ từ điển tiếng Việt và tiến hành phân cụm các từ có tần số xuất hiện chung trên cùng một trang của Wikipedia. Phương pháp đề xuất được trình bày như sau:

- Đầu tiên, phương pháp đề xuất loại bỏ những từ loại liên kết câu có thể gây nghiễu trong quá trình tính toán như: “và”, “thì”, “là”, “những”, “cho nên”, “do đó”, “bởi vì”…
- Tiếp theo, loại bỏ các từ có tần số xuất hiện rất thấp hoặc xuất hiện quá cao. Việc loại bỏ các từ có tần số xuất hiện thấp là vì những từ này khó có thể mang lại kết quả thống kê chính xác. Việc loại bỏ những từ có tần số xuất hiện quá cao là vì các từ này chủ yếu là các từ khóa của các trang Wikipedia, chẳng hạn như bách khoa, toàn thư, mục lục, phân loại, tham khảo, chú thích, phân bố, liên kết ngoài.
- Sau đó, phương pháp đề xuất tính toán ma trận *P*tần số xuất hiện chungtrên cùng một trang Wikipedia của các cặp từ trong từ điển.
- Cuối cùng, xây dựng đồ thị dendrogram dựa ma trận đã tính toán. Thuật toán xây dựng đồ thị dendrogram được trình bày như sau:

Bước 1 : Khởi tạo ma trận w thể hiện xác suất xuất hiện cùng nhau của các cặp từ thứ i và j trên cùng một trang Wikipedia.

$$w[i,j] = \frac{P_{ij}}{P_{ii}+P_{jj}-P_{ij}}$$

Bước 2 : Xây dựng đồ thị dendrogram bằng cách lặp đi lặp lại việc dưới đây đến khi tất cả các từ đã được đánh dấu:

+ Tìm phần tử lớn nhất trong w thể hiện tần số xuất hiện cao nhất của cặp từ x và y.

+ Cập nhật lại ma trận w với mọi i

$$w[x,i] = \min(w[x,i], w[y,i])$$
$$w[i,y] = \min(w[i,x], w[i,y])$$

Với tần số xuất hiện chungP_{ij} là tổng số trang Wikipedia xuất hiện cảhai từ thứ i và $j$trong từ điển. Ta có, $P_{ii}+P_{jj}-P_{ij}$ là tổng số trang có ít nhất một trong hai từ thứ i và j. Suy ra $w[i,j]$ là xác suất xuất hiện cùng nhau trong tập chứa tất cả các trang có ít nhất một trong hai từ thứ i và j.

Việc phân cụm được thực hiện bằng cách cắt đồ thị theo một chiều dài nhất định thì ta sẽ được các cụm từ hay đi với nhau. Nhưng qua thực nghiệm nhận thấy rằng việc cắt theo chiều cao đôi khi dẫn tới kết quả không mong muốn do ghép quá nhiều từ vào một cụm. Bài báo đề xuất việc cắt theo số bậc (tầng) của cây đồ thị kết hợp với chiều cao. Như vậy, chúng ta sẽ nhóm được một cụm có số từ theo như số bậc cho trước và cũng đúng theo chiều cao như mong muốn ban đầu.

4. Kết quả nghiên cứu

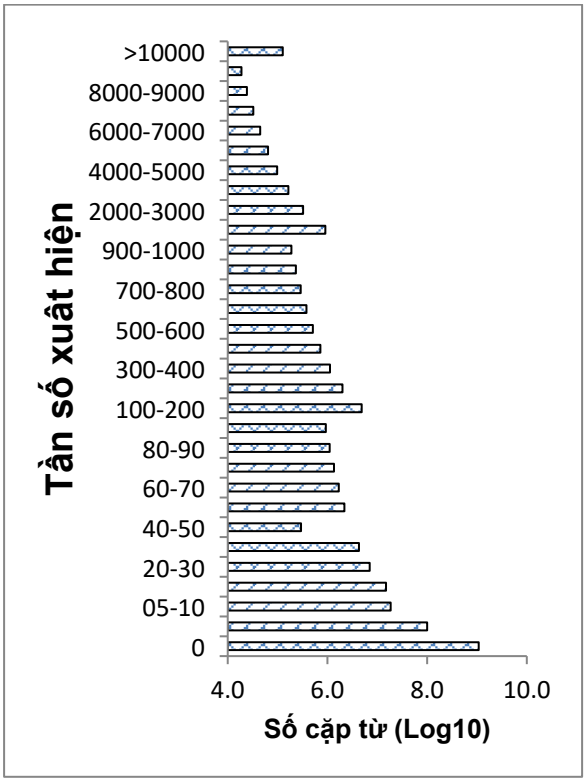
Bài báo này tiến hành thực nghiệm với bộ từ điển gồm các từ tiếng Việt khoảng 39000 từ. Bộ từ điển này được tạo ra từ bộ từ điển Việt-Pháp [15] bằng cách lấy danh sách của tất cả các từ tiếng Việt có trong từ điển Việt-Pháp. Sau khi lược bỏ các từ liên kết từ như "là", "và", "hoặc",…từ điển còn lại 34520 từ. Thông qua việc phân tích tần số xuất hiện trên Wikipedia, các từ có tần số thấp sẽ được loại bỏ vì khả năng gom thành của các từ này là rất thấp. Qua quá trình này, từ điển tiếp tục được rút gọn còn 14015 từ.

Hình 2 biểu diễn số lượng cặp từ theo tần số xuất hiện chung. Dễ dàng thấy rằng số cặp từ không xuất hiện chung trên một trang bất kỳ có số lượng lớn nhất (1.1×10^9cặp từ). Số lượng cặp từ tỉ lệ nghịch với tần số xuất hiện chung.

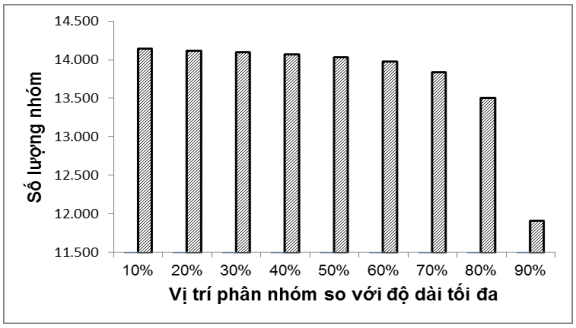
Hình 3 biểu diễn kết quả của việc phân cụm sử dụng phương pháp phân tích nhóm dựa trên đồ thị dendrogram.

Tại vị trí cắt là 40% so với độ dài tối đa, nghiên cứu đã tìm được các nhóm từ có liên quan hoặc gần nghĩa thể hiện ở hình 4, và hình 5.

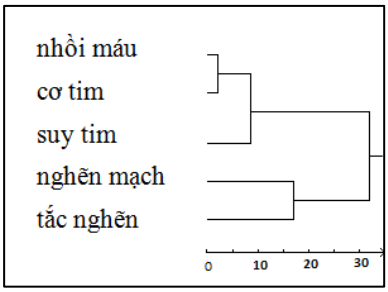
Theo hình 4 ta có khoảng cách của 2 từ "nhồi máu" và "cơ tim" rất thấp, có thể thấy được 2 từ này thường xuyên đi chung với nhau theo cụm từ "nhồi máu cơ tim". Từ "suy tim" có quan hệ gần với "nhồi máu | cơ tim" và "tắc nghẽn | nghẽn mạch" có quan hệ xa hơn so với "nhồi máu | cơ tim | suy tim". Tuy nhiên các từ này được gom đúng thành một nhóm chứng tỏ phương pháp đề xuất đã phân cụm thành công các cụm từ có liên quan chặt chẽ với nhau.



Hình 2. *Số lượng cặp từ theo tần số xuất hiện chung*



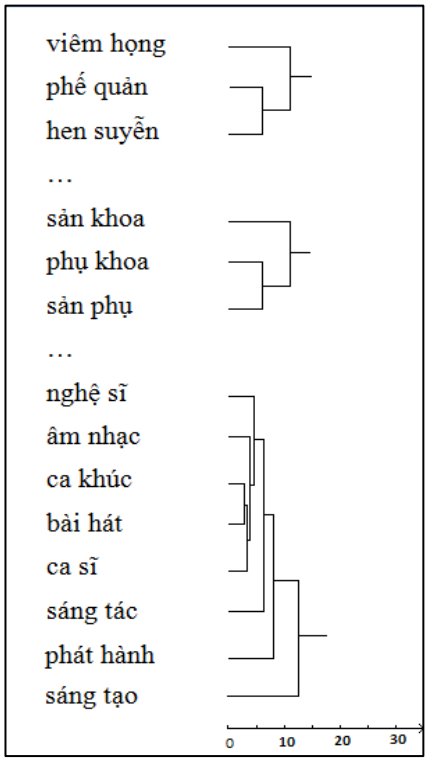
Hình 3. *Số lượng nhóm phụ thuộc vào vị trí phân cụm trên đồ thị dendrogram*



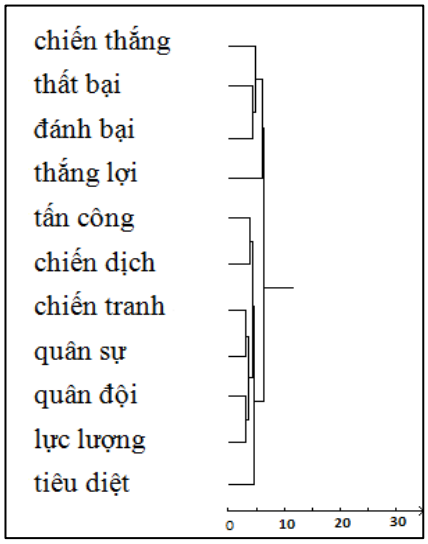
Hình 4. *Kết quả phân cụm với dendrogram*

Hình 5 là một số kết quả phân cụm đúng sử dụng phương pháp đề xuất. Ta dễ dàng nhận thấy được rằng các nhóm từ được phân cụm thành các chủ đề.

Trong kết quả thực nghiệm, tác giả đã tiến hành chọn ngẫu nhiên 1000 nhóm từ và tiến hành đếm thủ công số lượng nhóm đồng nghĩa đúng. Kết quả thu được là có 56% nhóm bao gồm hai từ đồng nghĩa. Ngoài ra còn phát hiện một số cụm từ bao gồm cả danh từ, động từ và tính từ cho một chủ đề.Ví dụ như hình 6.



Hình 5. *Một vài kết quả khác*



Hình 6. *Ví dụ đồ thị dendrogram cho các từ (chiến thắng, thắng lợi, đánh bại, thất bại, tấn công, chiến dịch, chiến tranh, quân sự, quân đội, lực lượng, tiêu diệt)*

Tuy nhiên, vẫn còn có một số từ không mang cùng một ý nghĩa nhưng có chung một nhóm từ như, "lòng tham" và "vui lòng tham khảo". Những từ này thông thường một từ là chuỗi con của từ kia, dẫn tới việc hay xuất hiện cùng nhau nên kết quả phân cụm chưa được chính xác.Ngoài ra, trong tiếng Việt còn có rất nhiều từ, cụm từ không có trong từ điển mà tác giả đã sử dụng như "cà chớn", "cà cháo". Hơn nữa báo cáo này chỉ giới hạn trên các trang Wikipedia nên chưa thể phát hiện hết tất cả các từ, cụm từ liên quan với nhau trong tiếng Việt.

5. Kết luận

Bài báo đã đề xuất phương pháp kết hợp Wikipedia và phương pháp phân tích nhóm dựa trên đồ thị dendrogram nhằm phân cụm cho từ tiếng Việt.Kết quả thực nghiệm cho thấy, phương pháp đề xuất đã phân cụm đúng các cụm từ đồng nghĩa cũng như các từ có cùng chủ đề.Tuy nhiên báo cáo chỉ dừng lại ở việc đánh giá tính hợp lí của việc áp dụng đồ thị dendrogram trong việc đánh giá mối quan hệ giữa các từtrong từ điển Tiếng Việt.Trong những nghiên cứu tới, tác giả sẽ tiến hành sử dụng kết quả phân cụm đã trình bày vào việc trích chọn đặc tính trong phân loại văn bản tự động.

TÀI LIỆU THAM KHẢO

[1] Trần Cao Đệ và Phạm Nguyên Khang, Phân loại với máy học vector hỗ trợ và cây quyết định, *Tạp chí khoa học Trường Đại học Cần Thơ*, p. 52-63, 21a, 2012.

[2] Vo Duy Thanh, Vo Trung Hung, Pham Minh Tuan, Doan Van Ban, "Text classification based on semi-supervised learning", *Proceeding of the SoCPaR 2013, IEEE catalog number CFP1395H-ART*, ISBN 978-1-4799-3400-3, 2013

[3] H. Q. Thắng and Đ. T. T. Phương, "Tiếp cận phương pháp học không giám sát trong học có giám sát với bài toán phân lớp văn bản tiếng Việt và đề xuất cải tiến công thức tính độ liên quan giữa hai văn bản trong mô hình vectơ," *Kỷ yếu Hội thảo ICT.rda'04*, pp. 251-261, 2005.

[4] Nguyễn Ngọc Bình, Dùng lý thuyết tập thô và các kỹ thuật khác để phân loại, phân cụm văn bản tiếng Việt, *Kỷ yếu hội thảo ICT.rda'04*. Hà nội 2004.

[5] Chih-Hao Tsai, MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm. http://technology.chtsai.org/MMSEG/, 2000.

[6] Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen and Hung Quoc Ngo, A Comparative Study on Vietnamese Text Classification Methods, Research, Innovation and Vision for the Future, *2007 IEEE International Conference* on, p. 267-273, 1-4244-0694-3, 2007.

[7] Hung Nguyen, Ha Nguyen, Thuc Vu, Nghia Tran, and Kiem Hoang. 2005. Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese. *Proceedings of 4th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future 2006 (RIVF'06)*. Ho Chi Minh City, Vietnam, Feb 12-16, 2006.

[8] Giang Son Nguyen, Xiaoying Gao and Peter Andreae, Vietnamese Document Representation and Classification, AI 2009: Advances in Artificial Intelligence Lecture Notes in Computer Science, *Springer*, Volume 5866, p. 577-586, 2009.

[9] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML)*, 1998.

[10] Trang Wikipedia – bách khoa toàn thư mở: http://vi.wikipedia.org/

[11] Jin Chen, Alan M. Mac Eachren, and Donna J. Peuquet. Constructing Overview + Detail Dendrogram – Matrix Views. *IEEE Trans Vis Comput Graph*. 2009 Nov-Dec; 15(6): 889-89

[12] Greenacre, M. J. *Correspondence Analysis inPractice*. London: Academic Press, 1993

[13] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, p. 281–297, 1967.

[14] J. Bezdek, R. Ehrlich and W. Full, "FCM: the fuzzyc-means clustering algorithm," *Computers and Geosciences*, vol. 10, p. 191–203, 1984.

[15] Ho Ngoc Duc, The Free Vietnamese Dictionary Project, http://www.informatik.uni-leipzig.de/~duc/Dict/install.html.

(BBT nhận bài: 31/03/2014, phản biện xong: 05/05/2014)