Tạp chí Tin học và Điều khiển học, T.25, S.1 (2009), 43–52

ÁP DỤNG KỸ THUẬT KHAI THÁC ĐỒ THỊ VÀO BÀI TOÁN PHÂN LOẠI VĂN BẢN

NGUYỄN HOÀNG TÚ ANH[1], HOÀNG KIẾM[2]

[1]*Khoa CNTT - Trường Đại học Khoa học Tự nhiên, ĐHQG Tp.HCM*
[2]*Khoa KHMT - Trường Đại học Công Nghệ Thông Tin, ĐHQG Tp. HCM*

Abstract. Text classification is the problem of assigning predefined class labels to incoming, unclassified documents. In this paper, we propose a graph based mining framework for text classification. A graph based instead of traditional vector based model is used for document representation. The system employs structural patterns (subgraphs) and the Dice similarity measure to identify a topic of documents. The Manhattan measure is used to compare the performance of method. The results show that it can outperform k-NN algorithm based on vector representation and classification using the Manhattan measure.

Tóm tắt. Phân loại văn bản là quá trình gán văn bản vào một hoặc nhiều chủ đề đã xác định trước. Bài báo đề cập đến khung mô hình phân loại văn bản dựa trên phương pháp khai thác đồ thị. Các văn bản được biểu diễn bằng đồ thị thay cho mô hình không gian vectơ truyền thống. Để xác định chủ đề cho văn bản, hệ thống sử dụng các đặc trưng cấu trúc đồ thị con và độ đo tương tự Dice. Độ đo Manhattan được dùng để so sánh kết quả phân loại. Kết quả cho thấy sự vượt trội của hệ thống so với thuật toán k-NN(trên mô hình vectơ) và phương pháp phân loại dùng độ đo Manhattan.

1. MỞ ĐẦU

Với sự phát triển của công nghệ thông tin, con người dễ dàng chia sẻ dữ liệu và tri thức cho nhau trên toàn thế giới. Ta có thể truy cập tức thời vào những kho dữ liệu khổng lồ qua Internet. Điều này đòi hỏi phải có cơ chế xác định sự liên quan giữa các thông tin đang được truy cập. Một trong các cơ chế đó là phương pháp phân loại văn bản cho phép người dùng truy vấn thông tin liên quan đến các chủ đề mà họ quan tâm. Phân loại văn bản là quá trình gán văn bản vào một hoặc nhiều chủ đề đã xác định trước. Phân loại tự động văn bản là một lĩnh vực nghiên cứu lý thú, được quan tâm trong nhiều năm qua do khả năng ứng dụng rộng rãi. Rất nhiều phương pháp đã được đề xuất [11] như Nave Bayes, cây quyết định, k-láng giềng gần nhất (k-NN), mạng nơron, máy vectơ hỗ trợ (SVM) và phương pháp dựa trên luật kết hợp. Trong số đó thì cả hai phương pháp SVM và k-NN đều cho kết quả tốt hơn khi phân loại văn bản [17].

Mô hình không gian vectơ [14] là mô hình biểu diễn văn bản phổ biến. Trong mô hình này, mỗi từ trong văn bản có thể trở thành đặc trưng (hay chiều của vectơ biểu diễn văn

bản). Mặc dù mô hình đơn giản này cho kết quả phân loại khá tốt, nhưng nó cũng tồn tại các hạn chế. Mô hình không gian vectơ truyền thống chỉ tập trung vào tần suất xuất hiện của từ và không nắm bắt được các thông tin cấu trúc như thứ tự, vị trí và sự đồng hiện của từ trong văn bản.

Gần đây, mô hình biểu diễn văn bản bằng đồ thị đã được phát triển mạnh [12, 15]. Mô hình đồ thị có khả năng hạn chế nhược điểm của mô hình không gian vectơ truyền thống khi lưu lại được các thông tin cấu trúc của văn bản. Trong mô hình này, "thuật ngữ" (có thể là từ, tiếng, cụm từ...) đại diện cho đỉnh. Các cạnh liên kết đỉnh cung cấp cả nội dung lẫn thông tin cấu trúc. Trên mô hình đồ thị, có thể áp dụng thuật toán k-NN mở rộng với độ đo tương tự giữa các đồ thị [12] hoặc mô hình lai kết hợp với rút trích cấu trúc con từ đồ thị như các đặc trưng cho văn bản [6] để phân loại văn bản trên web.

Trong bài báo này, chúng tôi trình bày một khung mô hình dùng kỹ thuật khai thác đồ thị để phân loại văn bản. Chúng tôi khai thác các văn bản thuộc cùng một lớp/chủ đề để phát hiện các mẫu đại diện và phổ biến. Sau đó chúng tôi xây dựng vectơ chủ đề dựa trên tập mẫu phổ biến hay tập đồ thị con phổ biến. Văn bản mới sẽ được biểu diễn bằng đồ thị và chuyển thành vectơ với các chiều của vectơ là đồ thị con phổ biến của chủ đề. Độ đo tương tự Dice được dùng để xác định khoảng cách gần nhất giữa văn bản mới và các vectơ chủ đề. Ngoài ra độ đo Manhattan cũng được dùng để so sánh kết quả phân loại. Khung mô hình phân loại này có thể áp dụng cho mọi ngôn ngữ. Để đánh giá mô hình, chúng tôi áp dụng mô hình phân loại này vào bài toán phân loại văn bản tiếng Việt. Các kết quả ban đầu của tiếp cận này đã được trình bày trong [9].

Những khái niệm cơ bản liên quan đến kỹ thuật khai thác đồ thị được trình bày ở Mục 2. Mục 3 mô tả chi tiết khung mô hình phân lớp văn bản. Mục 4 trình bày kết quả thực nghiệm và so sánh với thuật toán k-NN. Một số vấn đề cần tiếp tục nghiên cứu thêm được nêu ở phần cuối của bài báo.

2. KHAI THÁC ĐỒ THỊ

Mục đích của quá trình khai thác đồ thị là tìm ra các mẫu hay đồ thị con lý thú và phổ biến trong dữ liệu có cấu trúc. Điều quan trọng hơn là việc biểu diễn dữ liệu dưới dạng đồ thị bảo toàn được thông tin cấu trúc của dữ liệu ban đầu, trong khi các thông tin này có thể bị mất đi khi chuyển đổi sang các hình thức biểu diễn khác. Bài toán khai thác đồ thị là lĩnh vực nghiên cứu sôi động trong vài năm gần đây. Có hai hướng tiếp cận chính đối với bài toán xác định đồ thị con phổ biến từ tập dữ liệu đồ thị [18]:

• Khai thác đồ thị con phổ biến trên một đồ thị lớn: thuật toán SUBDUE, GBI, PATH, GREW, SIGRAM.

• Khai thác đồ thị con phổ biến trên tập dữ liệu gồm nhiều đồ thị nhỏ : thuật toán AGM, FSG, gSpan, FFSM, Gaston.

Trong các phương pháp tìm đồ thị con phổ biến trên tập dữ liệu đồ thị, gSpan là thuật toán nhanh, cho kết quả ổn định và dễ dàng cải tiến cho tập đồ thị có hướng. Thuật toán gSpan nguyên thủy chỉ có thể áp dụng trên đồ thị vô hướng và đồ thị con nhỏ nhất tìm được

là đồ thị gồm hai đỉnh một cạnh. Chúng tôi đã cải tiến thuật toán gSpan để có thể áp dụng trên đồ thị có hướng và nhận biết các đỉnh đơn có tần số xuất hiện thoả ngưỡng phổ biến tối thiểu minsupp đã cho.

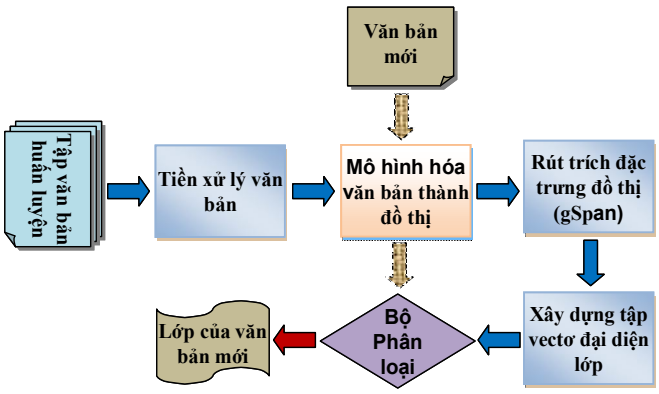
Thuật toán gSpan (graph-based Substructure pattern) [16] là một thuật toán khai thác đồ thị con phổ biến theo chiều sâu. Đầu vào của thuật toán là tập đồ thị có gán nhãn và đầu ra là tất cả các đồ thị con liên thông có tần suất xuất hiện không nhỏ hơn ngưỡng minsupp đã cho. Ý tưởng của thuật toán là phát triển các mẫu trực tiếp từ đồ thị đơn, tìm lần lượt từng đồ thị con phổ biến, từ những đồ thị có kích thước nhỏ đến đồ thị có kích thước lớn. gSpan sử dụng hệ thống biểu diễn đồ thị chính tắc để hỗ trợ quá trình tìm kiếm. Thuật toán ánh xạ mỗi mẫu vào nhãn chính tắc duy nhất và gán mỗi đồ thị một mã DFS (Depth-first search) tối tiểu. Mã DFS là thứ tự duyệt các cạnh của đồ thị theo chiều sâu. Dựa trên các nhãn này, quan hệ thứ tự đầy đủ giữa các mẫu được tạo lập. Thứ tự từ điển này cũng được dùng trong việc thiết lập cây tìm kiếm phân cấp (gọi là cây DFS). Trong quá trình duyệt cây theo chiều sâu, thuật toán gSpan chỉ mở rộng ứng viên trên các cạnh/ nhánh nằm bên phải nhất của cây DFS.

Vấn đề phức tạp nhất trong bài toán khai thác đồ thị con phổ biến là xác định đẳng cấu đồ thị con. Đây là bài toán có độ phức tạp NP khi các đỉnh trong đồ thị không được gán nhãn duy nhất (nhiều đỉnh có cùng nhãn như trong tập dữ liệu về hợp chất hóa học). Tuy nhiên, với mô hình biểu diễn văn bản thành đồ thị mà mỗi đỉnh được gán nhãn duy nhất (tên của thuật ngữ) thì độ phức tạp giảm xuống còn $O(n^2)$ (n - số lượng đồ thị).

3. MÔ HÌNH PHÂN LOẠI VĂN BẢN

Hệ thống TCG (Text Classification based on Graph) tự động thực hiện việc phân loại văn bản. Mô hình chúng tôi đề xuất sử dụng kỹ thuật khai thác đồ thị để phân loại văn bản vào chủ đề tương ứng. Hình 1 là khung mô hình phân loại TCG với các thành phần như trình bày dưới đây.

3.1. Tiền xử lý văn bản



Hình 1. Sơ đồ hệ thống TCG

Các văn bản thuộc cùng một lớp tạo nên tập dữ liệu huấn luyện để xác định các mẫu phổ biến và lý thú sau khi các hư từ bị loại. Việc loại bỏ hư từ được tiến hành khi tiền xử lý các lớp cũng như thực hiện trên văn bản mới để loại bỏ các từ không cung cấp thông tin có giá trị. Sau đó, tần suất xuất hiện của các "thuật ngữ" trong văn bản (có thể là từ, tiếng, hay cụm từ tùy theo kiểu đồ thị được chọn để biểu diễn văn bản) được tính. Để giảm kích thước của đồ thị và thời gian tính toán đồ thị con phổ biến, chúng tôi chỉ giữ lại $f\%$ số "thuật ngữ" có tần số xuất hiện cao nhất. Chúng tôi thống kê tần số xuất hiện và tính trọng số của "thuật ngữ" theo công thức TF-IDF (Term Frequency Invert Document Frequency) [14]. Sau khi tính trọng số cho các "thuật ngữ", chúng tôi sắp xếp chúng theo thứ tự trọng số giảm dần từ cao xuống thấp và lấy $f\%$ số "thuật ngữ" để tạo lập đồ thị.

3.2. Mô hình hóa văn bản thành đồ thị

Ưu điểm chính của mô hình biểu diễn văn bản bằng đồ thị là mô hình này có thể lưu giữ các thông tin cấu trúc của văn bản ban đầu. Có nhiều cách xây dựng đồ thị từ văn bản. Tác giả [12] trình bày một số kiểu chính: kiểu chuẩn, đơn giản, khoảng cách n,khoảng cách n đơn giản và đồ thị tần số. Tất cả các kiểu này đều sử dụng sự liền kề của thuật ngữ"(term). Chúng tôi chọn mô hình đồ thị đơn giản [12] hay mô hình đồ thị có hướng để biểu diễn văn bản vì tin rằng mô hình này có thể áp dụng cho mọi loại văn bản. Trong mô hình đồ thị đơn giản, mỗi văn bản là một đồ thị. Đỉnh biểu diễn "thuật ngữ" trong văn bản. Các đỉnh được gán nhãn duy nhất là tên của "thuật ngữ". Sau bước tiền xử lý văn bản, nếu thuật ngữ "a" đứng ngay trước thuật ngữ "b" thì sẽ có cạnh nối từ đỉnh "a" đến đỉnh "b" (không kể các trường hợp phân cách bởi dấu câu). Mô hình đồ thị là mô hình độc lập ngôn ngữ: có thể áp dụng cho văn bản trên ngôn ngữ bất kỳ.

Ví dụ : Ta có văn bản sau

"*Microsoft sẽ giới thiệu hệ điều hành Vista và trưng bày các công nghệ bổ trợ được xây dựng để cải tiến hệ điều hành*".

Hình 2. Ví dụ mô hình đồ thị đơn giản

"Thuật ngữ" ở đây tương ứng với đơn vị từ của văn bản. Hình 2 là mô hình biểu diễn văn bản trên ở dạng đồ thị đơn giản khi đã qua bước loại bỏ hư từ và các từ có trọng số thấp.

3.3. Rút trích đặc trưng đồ thị

Kỹ thuật khai thác đồ thị được dùng để rút trích các đặc trưng hay các đồ thị con phổ biến. Trong từng chủ đề, chúng ta thực hiện việc rút trích đặc trưng là các đồ thị con phổ biến có tần số xuất hiện lớn hơn ngưỡng phổ biến tối thiểu minsupp. Quá trình này được

thực hiện bằng thuật toán gSpan đã được trình bày ở phần 2. Sau đó, ta tổng hợp các đồ thị con phổ biến thu được từ tất cả các chủ đề. Chúng ta xây dựng được tập các đặc trưng tập đồ thị con phổ biến. Đây là đầu vào cho bước xây dựng vectơ đại diện lớp/chủ đề tiếp theo.

3.4. Xây dựng vectơ đại diện chủ đề

Chúng tôi xây dựng vectơ nhị phân đại diện cho từng chủ đề. Mỗi chủ đề cho trước được biểu diễn thành một vectơ đặc trưng có số chiều bằng kích thước tập đồ thị con phổ biến. Đặc trưng nhận giá trị 1 nếu đồ thị con phổ biến tương ứng xuất hiện trong tập đồ thị con phổ biến của chủ đề và ngược lại. Kết quả, chúng ta thiết lập được tập vectơ đặc trưng nhị phân đại diện cho các chủ đề.

Giả sử đầu vào của quá trình huấn luyện là tập văn bản huấn luyện $D = \{d_1, d_2, ..., d_n\}$ có gán nhãn lớp và tập các lớp $C = \{C_1, C_2, ..., C_m\}$. Mỗi văn bản $d_i \in D;\ 1 \leq i \leq n$ chỉ thuộc về một lớp $C_j \in C;\ 1 \leq j \leq m$. Khi đó, sau bước mô hình hoá văn bản ta thu được tập đồ thị $G = \{G_1, G_2, ..., G_n\}$ tương ứng với các văn bản thuộc tập D và được phân chia thành m lớp phân biệt. Sau khi áp dụng thuật toán $gSpan$ lên từng lớp trong G, ta tổ hợp đồ thị con phổ biến từ tất cả các lớp và thu được tập đặc trưng $F = \{f_1, f_2, ..., f_k\}$. Vectơ đại diện cho mỗi chủ đề/lớp $C_i;\ 1 \leq i \leq m$ là vectơ $R_i = (R_{i1}, R_{i2}, ..., R_{ik})$ có k chiều với giá trị $R_{ij} = 1$ nếu đặc trưng $f_j \in F$ là một trong các đồ thị con phổ biến tìm được từ tập đồ thị biểu diễn văn bản thuộc lớp C_i và ngược lại. Kết quả, chúng ta thiết lập m vectơ nhị phân $\{R_1, R_2, ..., R_m\}$ với R_i là đại diện cho lớp C_i.

3.5. Bộ phân loại

Đầu tiên, chúng ta thực hiện các bước tiền xử lý như tách câu, loại bỏ hư từ và sử dụng tập các "thuật ngữ" đã lựa chọn trong quá trình huấn luyện để xây dựng đồ thị biểu diễn cho văn bản mới. Sau đó, ta cần xác định trong đồ thị biểu diễn văn bản mới có chứa những đặc trưng của tập đặc trưng (hay có chứa các đồ thị con phổ biến) và từ đó xây dựng vectơ nhị phân có số chiều tương ứng với kích thước của tập đặc trưng. Giá trị của từng thành phần trong vectơ biểu diễn văn bản mới thể hiện sự tồn tại hay không các đặc trưng trong đồ thị biểu diễn văn bản tương ứng. Tiếp theo, ta tính toán sự tương tự giữa vectơ biểu diễn văn bản mới với tất cả vectơ đại diện cho các chủ đề/lớp. Chúng tôi sử dụng độ đo Dice [1] vì đây là độ đo đơn giản nhưng rất hiệu quả trong việc xác định độ tương tự giữa các vectơ nhị phân. Ngoài ra, độ đo Manhattan [1] cũng được dùng để so sánh kết quả phân loại.

Giả sử vectơ v_0 biểu diễn văn bản mới X và tập $R = \{R_1, R_2, ..., R_m\}$ với R_i là vectơ đại diện cho lớp C_i. Khi đó công thức tính độ đo Dice và Manhattan giữa vectơ v_0 và vectơ R_i đại diện cho lớp C_i như sau:

$$\text{Dice}(v_0, R_i) = \frac{2|v_0 \wedge R_i|}{|v_0| + |R_i|}, \tag{1}$$

$$\text{Manhattan}(v_0, R_i) = \sum_{h=1}^{k} |(v_{oh} - R_{ih})|, \tag{2}$$

trong đó, $|v_0|$, $|R_i|$ là tổng số đặc trưng mang giá trị 1 của vectơ v_0, R_i; k - số chiều của vectơ.

Cuối cùng, dựa trên các độ tương tự Dice ta gán văn bản mới vào lớp cho giá trị Dice lớn nhất. Còn nếu sử dụng độ đo Manhattan thì lớp có giá trị Manhattan nhỏ nhất được chọn làm lớp cho văn bản mới.

4. KẾT QUẢ THỬ NGHIỆM

Để đánh giá mô hình phân loại dựa trên kỹ thuật khai thác đồ thị, chúng tôi thử nghiệm mô hình này vào bài toán phân loại văn bản tiếng Việt. Những nghiên cứu gần đây trên bài toán phân loại văn bản tiếng Việt đã sử dụng: lý thuyết tập thô [10], tiếp cận sử dụng học không giám sát [5], tiếp cận dựa trên luật kết hợp [3], Nave Bayes [8], thuật giải di truyền [7] và SVM [4]. Phần lớn các công trình này dựa trên mô hình biểu diễn văn bản là mô hình túi từ hoặc mô hình không gian vectơ và phần nào bị phụ thuộc vào bộ công cụ tách từ.

Dữ liệu thử nghiệm gồm các bài báo lấy từ các tờ báo điện tử lớn và được phân chia thành các nhóm dựa trên các chủ đề đã có trên các trang trực tuyến này. Chúng tôi thực hiện việc chọn lọc các chủ đề chung từ các trang web này. Bộ dữ liệu thử nghiệm bao gồm 3900 tập tin văn bản được chia thành 7 chủ đề như trong bảng 1.Văn bản có kích thước từ 1KB đến 15KB.

Khi áp dụng mô hình phân loại đã đề xuất lên tiếng Việt, chúng tôi chọn lựa đơn vị tiếng" tương ứng với khái niệm "thuật ngữ" để biểu diễn đỉnh trong đồ thị. Sau khi tách câu, loại bỏ hư từ, xác định trọng số cho tiếng và loại bỏ bớt các tiếng có trọng số thấp, chúng tôi thu được trung bình 40 đỉnh/đồ thị.

Bảng 1. Tập dữ liệu thử nghiệm

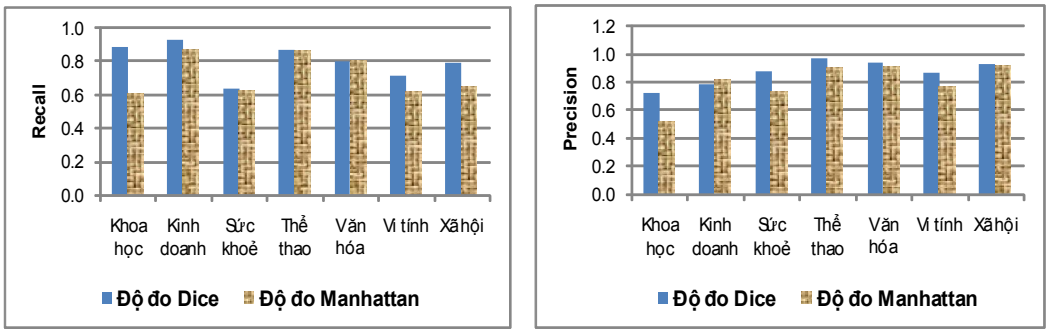
STT	Tên chủ đề	Số văn bản
1	Khoa học	358
2	Kinh doanh	654
3	Sức khoẻ	315
4	Thể thao	759
5	Văn hóa	522
6	Vi tính	457
7	Xã hội	835

Để đánh giá kết quả, chúng tôi sử dụng các chỉ số độ phủ (recall), độ chính xác (precision) và chỉ số cân bằng giữa 2 độ đo trên - F1 [11]. Chúng tôi sử dụng phương pháp đánh giá chéo (k-fold validation) để chạy thử nghiệm. Kết quả thử nghiệm tương ứng với độ đo Dice và độ đo Manhattan được trình bày trong Bảng 2. Thời gian huấn luyện trung bình là $2,5$ giây/ văn bản và thời gian thực hiện phân lớp tính từ thời điểm tiền xử lý văn bản mới cho đến khi phân lớp hoàn tất trung bình là $0,5$ giây/văn bản.

Mô hình phân lớp dùng dùng độ đo tương tự Dice cho kết quả tốt hơn mô hình dùng độ đo Manhattan. Đó là do bản chất của độ đo Manhattan, đặc biệt trong cách tính độ đo này. Độ đo Manhattan tính sự khác biệt của đặc trưng trong hai vectơ, trong khi độ đo Dice chỉ đếm số lượng đặc trưng mang giá trị 1 đồng hiện ở cả hai vectơ và chia cho tổng số đặc trưng này của cả hai vectơ. Hình 3 thể hiện biểu đồ so sánh kết quả phân loại theo các độ đo tương tự khác nhau.

Bảng 2. Kết quả thử nghiệm phân loại (5-fold validation)

Tên chủ đề	Độ đo tương tự Dice			Độ đo tương tự Manhattan		
	Độ phủ (Recall)	Độ chính xác (Precision)	Độ đo F1	Độ phủ (Recall)	Độ chính xác (Precision)	Độ đo F1
Khoa học	0.887	0.722	0.796	0.6	0.515	0.544
Kinh doanh	0.931	0.787	0.853	0.866	0.813	0.839
Sức khoẻ	0.639	0.875	0.739	0.62	0.721	0.667
Thể thao	0.873	0.968	0.918	0.86	0.896	0.878
Văn hóa	0.798	0.941	0.864	0.8	0.909	0.851
Vi tính	0.717	0.865	0.784	0.615	0.767	0.683
Xã hội	0.792	0.933	0.857	0.65	0.915	0.76
Trung bình	**0.805**	**0.87**	**0.83**	**0.716**	**0.791**	**0.746**



a) So sánh theo độ phủ (Recall)

b) So sánh theo độ chính xác (Precision)

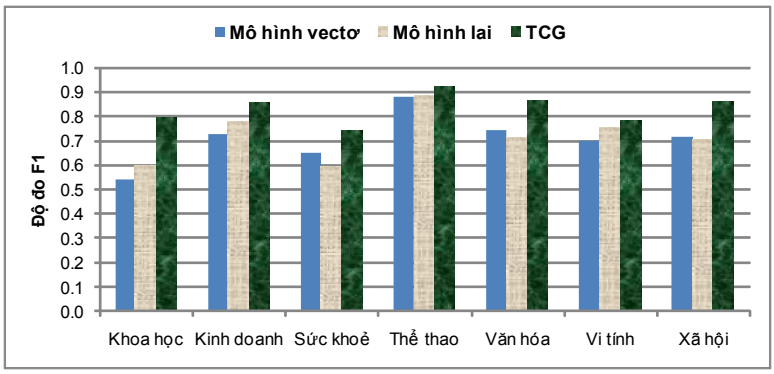
Hình 3. So sánh kết quả phân loại theo độ đo tương tự Dice và Manhattan

Chúng tôi cài đặt thuật toán k-láng giềng gần nhất (k-NN) trên mô hình không gian vectơ với độ đo Cosine [14]. Ở đây, chúng tôi sử dụng bộ công cụ tách từ tiếng Việt của nhóm tác giả [2] để xây dựng các vectơ đặc trưng văn bản. Phương pháp trích chọn đặc trưng dùng độ lợi thông tin (information gain) được áp dụng cho mô hình không gian vectơ nhằm nâng cao chất lượng phân loại. Bên cạnh đó, chúng tôi cũng thử nghiệm mô hình lai của tác giả [6] với thuật toán k-NN và độ đo tương tự Manhattan. Mô hình lai sử dụng cả dạng biểu diễn bằng đồ thị và vectơ. Bảng 3 trình bày kết quả phân loại tốt nhất của các phương pháp khác nhau theo độ đo F1.

Bảng 3. So sánh kết quả thử nghiệm phân loại theo giá trị F1 trung bình

TT	Mô hình biểu diễn văn bản	Mô tả thuật toán	Giá trị F1 trung bình
1	Mô hình vectơ	k-NN, độ đo tương tự Cosine	0.708
2	Mô hình lai	k-NN, độ đo tương tự Manhattan, "từ" tạo thành đỉnh	0.716
4	Mô hình TCG	So khớp với độ đo Dice, đỉnh tạo từ đơn vị "tiếng"	0.831

Hình 4 là đồ thị so sánh kết quả phân loại theo mô hình đồ thị TCG, mô hình lai và mô hình không gian vectơ trên các chủ đề theo độ đo F1. Kết quả của mô hình đã đề xuất TCG với phương pháp so khớp dùng độ đo Dice trên bộ dữ liệu này cho kết quả tốt nhất.



Hình 4. So sánh kết quả phân loại theo chủ đề

5. KẾT LUẬN

Qua bài báo này, chúng tôi đề xuất khung mô hình phân loại văn bản dựa trên kỹ thuật khai thác đồ thị. Hệ thống TCG sử dụng bộ phân lớp dựa trên phương pháp so khớp với vectơ đặc trưng chủ đề dùng độ đo Dice cho kết quả tốt hơn độ đo Manhattan và vượt trội so với phương pháp sử dụng mô hình không gian vectơ, cũng như mô hình lai. Kết quả thử nghiệm cho thấy tính hiệu quả của tiếp cận này. Chúng tôi tin rằng kỹ thuật này có thể áp dụng cho các kiểu dữ liệu văn bản khác như văn bản web hay email, cũng như có thể áp dụng cho các ngôn ngữ khác nhau. Chúng tôi dự định phát triển nghiên cứu theo hướng này. Bên cạnh đó, chúng tôi cũng sẽ nghiên cứu, cải tiến quá trình rút trích đặc trưng - đồ thị con phổ biến nhằm nâng cao hiệu quả và thời gian phân loại. Đồng thời thử nghiệm và so sánh với nhiều phương pháp phân loại truyền thống khác để đánh giá sâu hơn tính hiệu quả của mô hình.

Lời cảm ơn. Nhóm tác giả xin gửi lời cảm ơn đến TS. Đinh Điền và nhóm Vietnamese Computational Linguistics (VCL) Khoa CNTT, Trường ĐH Khoa học Tự nhiên, ĐHQG Tp. HCM đã cung cấp bộ tách từ tiếng Việt.

TÀI LIỆU THAM KHẢO

[1] R. Baeza-Yates, and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[2] Dinh Dien, Vu Thuy, A maximum entropy approach for Vietnamese word segmentation, *Proc. of 4*th *IEEE International Conference on Computer Science - Research, Innovation and Vision for the Future 2006 (RIVF06)*, Ho Chi Minh, Vietnam , 2 - 2006 (247–252).

[3] Đỗ Phúc, Phát triển ứng dụng thuật toán tìm luật kết hợp vào bài toán phân loại văn bản tiếng Việt, *Kỷ yếu Hội thảo Khoa học quốc gia lần thứ 2 Nghiên cứu cơ bản và ứng dụng CNTT (FAIR05)*, Tp. Hồ Chí Minh, tháng 8, 2005 (275–286).

[4] Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, Hung Quoc Ngo, A comparative study on Vietnamese text classification methods, *Proc. of 2007 IEEE InternationalConference on Computer Science - Research, Innovation and Vision for the Future (RIVF07)*, Ha Noi, Vietnam, 2007 (267–273).

[5] Huỳnh Quyết Thắng, Đinh Thị Phương Thu, 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 vector, *Kỷ yếu Hội thảo ICT.rda04*, Hà Nội, 2004 (251–261).

[6] A. Markov, M. Last, Efficient graph-based representation of web documents, *Proc. of the Third International Workshop on Mining Graphs, Trees and Sequences (MGTS 2005)*, Porto, Portugal, 2005 (52-62).

[7] Hung Nguyen, Ha Nguyen, Thuc Vu, Nghia Tran, Kiem Hoang, Internet and genetics algorithm-based text categorization for documents in Vietnamese, *Proc. of 3*$^{\text{th}}$ *International Conference Research, Innovation and Vision of the Future (RIVF05)*, Can Tho, Vietnam, 2005 (168–172).

[8] Thanh V. Nguyen, Hoang K.Tran, Thanh T.T Nguyen, Hung Nguyen, Word segmentation for Vietnamese text categorization: An online corpus approach, *Poster Proc. of 4*$^{\text{th}}$ *IEEE InternationalConference on Computer Science - Research, Innovation and Vision for the Future 2006 (RIVF06)*, HoChiMinh, Vietnam, 2006 (113–118).

[9] Nguyễn Hoàng Tú Anh, Hoàng Kiếm, Phân loại văn bản tiếng Việt dựa trên khai thác đồ thị con phổ biến, *Kỷ yếu Hội thảo Khoa học Công nghệ quốc gia lần thứ 3 "Nghiên cứu cơ bản và ứng dụng CNTT", FAIR2007*, Nha Trang, Việt Nam, 2007 (258–268).

[10] Phan Thanh Liêm, Trần Văn Quang, Nguyễn Ngọc Bình, Hồ Tú Bảo, Ứng dụng mô hình tập thô dung sai trong xử lý văn bản tiếng Việt, *Kỷ yếu Hội nghị KH kỷ niệm 30 năm ngày thành lập Viện Công nghệ thông tin*, Hà Nội, 2006, NXB KHTN & CN (498–507).

[11] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, **34** (1) (2002) 1–47.

[12] A. Schenker, M. Last, H. Bunke, A. Kandel, Classification Of web documents using graph matching, *International Journal of Pattern Recognition and Artificial Intelligence, Special*

Issue on Graph Matching in Computer Vision and Pattern Reconition **18** (3) (2004) 475–479.

[13] J. F. Sowa, Conceptual graphs for a database interface, *IBM Journal of Research and Development* **20** (4) (July, 1976) 336–357.

[14] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, *Communication of. ACM* **18** (11) (1975) 613–620.

[15] J. Tomita, H. Nakawatase, M. Ishii, Graph-based text database for knowledge discovery, *Poster Proc. of WWW04*, Manhattan, NY, USA, 2004 (454-455).

[16] X. Yan, J. Han, gSpan: Graph-based substructure pattern mining, *Proc. of ICDM02, Maebashi City*, Japan, 2002 (721–723).

[17] Y. Yang and X. Liu, A re-examination of text categorization methods, *Proc. of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR99)*, California, USA, 1999, (42–49).

[18] M. Worlein, T. Meinl, I. Fisher, M. Philippsen, A quantative comparision of the subgraph miners MoFa, gSpan, FFSM, and Gaston, *Proc. of PKDD05*, Porto, Portugal, 2005, LNAI 3721 (392–403).

Nhận bài ngày 18 - 8 - 2008
Nhận lại sau sửa ngày 24 - 8 - 2009