MÔ HÌNH BIỂU DIỄN VĂN BẢN THÀNH ĐỒ THỊ

Nguyễn Hoàng Tú Anh, Nguyễn Trần Kim Chi, Nguyễn Hồng Phi
Trường Đại học Khoa học Tự nhiên, ĐHQG –HCM
(Bài nhận ngày 09 tháng 04 năm 2008, hoàn chỉnh sửa chữa ngày 26 tháng 09 năm 2008)

TÓM TẮT: *Biểu diễn văn bản là một bước tiền xử lý rất quan trọng trong nhiều lĩnh vực như khai thác dữ liệu văn bản, truy vấn thông tin, xử lý ngôn ngữ tự nhiên. Bài báo này trình bày tổng quan mô hình biểu diễn văn bản thành đồ thị. Mô hình đồ thị có thể giữ lại các thông tin cấu trúc như vị trí, thứ tự xuất hiện và sự gần nhau của từ, trong khi chúng bị loại bỏ trong mô hình không gian vectơ truyền thống. Chúng tôi xây dựng thử nghiệm hệ thống phân lớp văn bản tiếng Việt dựa trên mô hình biểu diễn văn bản thành đồ thị.*

Từ khoá*: Mô hình đồ thị, biểu diễn văn bản, phân lớp văn bản.*

1. GIỚI THIỆU

Hiện nay, chúng ta dùng các mô hình biểu diễn để giải quyết hầu hết những vấn đề liên quan đến văn bản. Chúng đóng vai trò trung gian giữa ngôn ngữ tự nhiên dạng văn bản và chương trình xử lý trong các lĩnh vực khai thác dữ liệu văn bản, truy vấn thông tin, xử lý ngôn ngữ tự nhiên. Sau khi được tái thể hiện, văn bản trở thành những cấu trúc dữ liệu trực quan, đơn giản và có thể xử lý được. Vì vậy, các mô hình biểu diễn không ngừng phát triển, hàm chứa được nhiều hơn những suy nghĩ mà con người muốn diễn đạt, đồng thời nâng cao hiệu quả sử dụng. Mô hình biểu diễn văn bản truyền thống như: mô hình túi từ và không gian vectơ là các mô hình được sử dụng phổ biến nhất. Mô hình không gian vectơ [7] biểu diễn văn bản như một vectơ đặc trưng của các thuật ngữ (từ) xuất hiện trong toàn bộ tập văn bản. Trọng số các đặc trưng thường được tính qua độ đo *TF*IDF*. Tuy nhiên, mô hình này không nắm bắt được các thông tin cấu trúc quan trọng như trật tự xuất hiện của các từ, vùng lân cận của từ, vị trí xuất hiện của từ trong văn bản. Để giải quyết các hạn chế trên, mô hình đồ thị được đề xuất và được đánh giá có nhiều tiềm năng vì tận dụng được các thông tin quan trọng về cấu trúc mà mô hình túi từ và không gian vectơ đã bỏ qua.

Mô hình đồ thị biểu diễn văn bản, cụ thể là mô hình đồ thị khái niệm (Conceptual Graphs_ CGs), được John F. Sowa trình bày lần đầu tiên vào năm 1976 [9]. Hiện nay, mô hình đồ thị không ngừng phát triển dựa trên ý tưởng của mô hình CGs, được ứng dụng vào dãy rộng các bài toán liên quan đến xử lý văn bản và trở nên khá phong phú. Khi ứng dụng vào từng loại bài toán khác nhau, các thành phần thích hợp nhất trong văn bản trở thành đỉnh của đồ thị và mối quan hệ hiệu quả nhất giữa các đỉnh được chọn để xây dựng cạnh của đồ thị. Đỉnh của đồ thị có thể biểu diễn câu, từ, hay câu kết hợp từ. Cạnh có thể dùng để thể hiện những mối quan hệ khác nhau giữa các đỉnh như: trật tự xuất hiện, tần số đồng hiện, vị trí xuất hiện, độ tương đồng.

Mục đích của bài báo này là nghiên cứu, hệ thống các biến thể của mô hình biểu diễn văn bản bằng đồ thị nhằm cung cấp cho người đọc cái nhìn tổng quan về mô hình này. Bên cạnh đó, chúng tôi cũng áp dụng thử nghiệm mô hình biểu diễn văn bản bằng đồ thị vào bài toán phân lớp văn bản tiếng Việt .

Các phần tiếp theo của bài báo được tổ chức như sau. Phần 2 giới thiệu tổng quan mô hình biểu diễn văn bản bằng đồ thị. Phần 3 giới thiệu hệ thống phân lớp văn bản sử dụng mô hình đồ thị kết hợp thuật toán khai thác đồ thị con phổ biến. Phần 4 trình bày kết quả thực nghiệm của hệ thống và cuối cùng là phần kết luận.

2. MÔ HÌNH HÓA VĂN BẢN THÀNH ĐỒ THỊ

Hiện nay, trên thế giới có một số công trình xử lý văn bản sử dụng mô hình đồ thị. Các mô hình đồ thị tương đối đa dạng và mỗi mô hình mang nét đặc trưng riêng. Sau quá trình nghiên cứu và tổng hợp, chúng tôi xin giới thiệu một số mô hình đồ thị biểu diễn văn bản chính có những đặc tính khái quát sau.

Mỗi đồ thị là một văn bản hoặc biểu diễn cho tập văn bản. Đỉnh của đồ thị có thể là câu, hoặc từ, hoặc kết hợp câu và từ. Cạnh nối giữa các đỉnh là vô hướng hoặc có hướng, thể hiện mối quan hệ trong đồ thị. Nhãn đỉnh thường là tần số xuất hiện của đỉnh. Còn nhãn cạnh là tên mối liên kết khái niệm giữa 2 đỉnh, hay tần số xuất hiện chung của 2 đỉnh trong một phạm vi nào đó, hay tên vùng mà đỉnh xuất hiện.

Ví dụ trong bài toán rút trích thông tin, đỉnh là từ [11] hay từ kết hợp câu [14], cạnh thể hiện tần số đồng hiện. Trong bài toán phân lớp văn bản, đỉnh là từ, cạnh thể hiện trật tự xuất hiện của từ hay vị trí xuất hiện của từ trong văn bản [1] [5] [8]. Còn trong bài toán tóm tắt văn bản thì đỉnh là câu, cạnh thể hiện sự tương đồng giữa các câu [6].

Do từ lưu giữ được nhiều thông tin cấu trúc nhất nên mô hình đồ thị sử dụng đỉnh là từ được nghiên cứu sâu hơn và có nhiều biến thể nhất. Chúng tôi tổng hợp các mô hình đồ thị chính và phân thành các nhóm như sau:

Ø Mô hình đồ thị sử dụng đỉnh là từ trong văn bản (ký hiệu từ số 1 → 10).

- Mô hình đồ thị sử dụng mạng ngữ nghĩa (mô hình số 1, 2, 3). Ưu điểm của nhóm mô hình này là mô hình hoá văn bản một cách trực quan, logic, thể hiện được quan hệ ngữ nghĩa giữa các khái niệm và cho kết quả truy vấn thông tin chính xác hơn.
- Mô hình đồ thị không sử dụng mạng ngữ nghĩa (mô hình số 4 → 10). Nhóm mô hình này khai thác được các thông tin cấu trúc của văn bản (thứ tự xuất hiện, vị trí, vùng lận cận của từ trong văn bản) nhanh chóng, đơn giản và không phụ thuộc vào mạng ngữ nghĩa nên dễ dàng cài đặt các ứng dụng phân lớp, gom cụm.

Ø Mô hình đồ thị sử dụng đỉnh là câu (mô hình số 11). Thế mạnh của mô hình này là khả năng lưu trữ mối liên kết giữa các câu, thứ tự xuất hiện câu và hỗ trợ tốt cho quá trình trích chọn câu quan trọng của văn bản để đưa vào bản tóm tắt bằng tiếp cận không giám sát.

Ø Mô hình đồ thị sử dụng đỉnh là câu và từ (mô hình số 12). Mô hình này tận dụng được mối liên quan giữa từ với câu, cũng như sự đồng hiện của từ trong câu để tăng hiệu quả của bài toán rút trích thông tin văn bản.

Chúng tôi tóm tắt những đặc trưng chính và lĩnh vực ứng dụng cơ bản của các mô hình biểu diễn văn bản bằng đồ thị trong bảng 1.

Trong các mô hình được giới thiệu ở trên, có những mô hình được mở rộng từ mô hình khác. Ví dụ như đồ thị dạng chuẩn là mô hình mở rộng của đồ thị đơn giản, đồ thị khoảng cách n là mô hình mở rộng của đồ thị khoảng cách n đơn giản với nhãn cạnh là vị trí của từ trong cấu trúc văn bản. Sau đây, chúng tôi sẽ trình bày chi tiết một số mô hình đại diện với đỉnh biểu diễn từ. Đó là mô hình đồ thị khái niệm, đồ thị hình sao, đồ thị tần số xuất hiện vô hướng, đồ thị đơn giản, đồ thị khoảng cách n đơn giản.

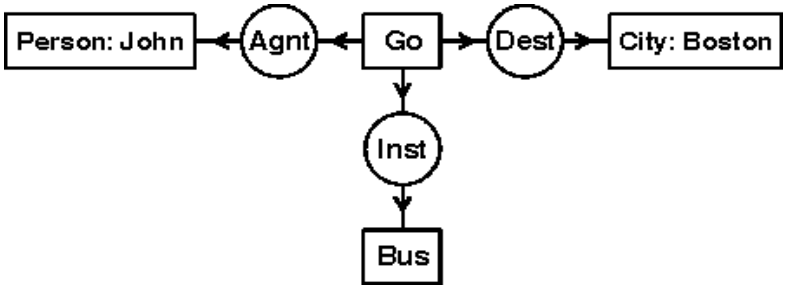
Bảng 1. Mô tả các mô hình biểu diễn văn bản bằng đồ thị

Mô hình	Tên riêng của mô hình	Đỉnh			Cạnh			Lĩnh vực ứng dụng
		Ý nghĩa	Số loại đỉnh	Nhãn	Ý nghĩa	Hướng	Nhãn	
1	Đồ thị khái niệm _ CGs	Từ	2	Không	Liên kết khái niệm	Có	Không	Truy vấn thông tin, thiết kế CSDL
2	CGs cải tiến vô hướng	Từ	1	Không	Liên kết khái niệm	Không	Không	Tìm kiếm thông tin trên Web
3	Đồ thị khái niệm cải tiến	Từ	1	Không	Liên kết khái niệm	Có	Có (cấu trúc ngữ pháp)	Gom cụm văn bản
4	Đồ thị hình sao	Từ / cấu trúc	1	Có (tần số xuất hiện)	Liên kết từ và đỉnh cấu trúc trung tâm	Không	Có (vị trí từ trong cấu trúc văn bản)	Phân loại email
5	Đồ thị tần số vô hướng	Từ	1	Có (tần số xuất hiện)	Liên kết từ xuất hiện chung trong cấu trúc	Không	Có (tần số xuất hiện chung)	Tìm kiếm thông tin trên Web
6	Đồ thị đơn giản	Từ	1	Có (tên từ)	Từ a xuất hiện ngay trước từ b	Có	Không	Phân lớp, gom cụm văn bản
7	Đồ thị khoảng cách n đơn giản	Từ	1	Không	Giữa từ a trước từ b có ít hơn n từ	Có	Không	Phân lớp văn bản
8	Đồ thị khoảng cách n	Từ	1	Không	Giữa từ a trước từ b có ít hơn n từ	Có	Có (số từ giữa a và b + 1)	Phân lớp văn bản
9	Đồ thị dạng chuẩn	Từ	1	Có (tên từ)	Từ a xuất hiện ngay trước từ b	Có	Có (vị trí từ trong cấu trúc vb)	Phân lớp, gom cụm văn bản
10	Đồ thị tần số	Từ	1	Có (tần số xuất hiện)	Từ a xuất hiện ngay trước từ b	Có	Có (tần số 2 từ xuất hiện liên tiếp)	Phân lớp văn bản
11	Đồ thị đỉnh là câu	Câu	1	Có (trọng số đỉnh)	Liên kết hai câu có từ chung	Có/ Không	Có (Độ tương tự giữa 2 câu)	Tóm tắt văn bản
12	Đồ thị song phương	Câu, từ	2	Không	Từ xuất hiện trong câu	Không	Có (tần số xuất hiện của từ trong câu)	Rút trích thông tin

2.1. Mô hình đồ thị khái niệm (Conceptual Graphs - CGs)

Mô hình đồ thị khái niệm sử dụng mạng ngữ nghĩa để biểu diễn văn bản thành đồ thị. Mỗi từ trong văn bản là một khái niệm và được biểu diễn bằng đỉnh hình vuông. Đỉnh hình oval thể hiện mối quan hệ giữa các khái niệm. Các đỉnh hình vuông được nối với nhau dựa trên mối quan hệ trong mạng ngữ nghĩa và qua trung gian là đỉnh hình oval. Ưu điểm của CGs là mô hình hoá văn bản một cách trực quan, chính xác và logic. Điểm hạn chế của CGs là khá phức tạp, đòi hỏi phân tích ngữ nghĩa sâu, chuyên biệt và phải phụ thuộc vào lĩnh vực.

Ví dụ 1: Ta có câu: "*Jonh is going to Boston by bus*".

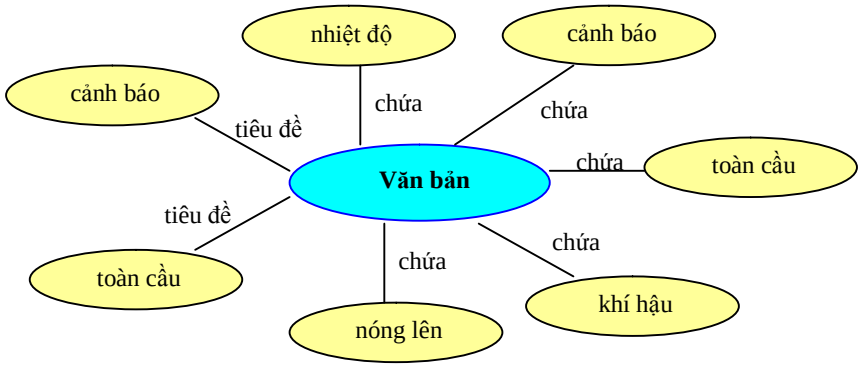


Hình 1. Ví dụ mô hình đồ thị khái niệm [15]

Mô hình đồ thị khái niệm biểu diễn câu trên như trong hình 1. Trong đó: các khái niệm là [Go], [Person: John], [City: Boston] và [Bus], các mối quan hệ là (Agnt) – tác nhân, (Dest) – nơi đến và (Inst) – phương tiện.

2.2. Mô hình đồ thị hình sao

Trong đồ thị hình sao, đỉnh trung tâm là nét khái quát cấu trúc của văn bản. Sau khi đỉnh trung tâm được xác lập, các đỉnh còn lại sẽ được triển khai. Ngoài đỉnh trung tâm, các đỉnh còn lại biểu diễn từ trong văn bản. Đỉnh thuộc khu vực nào trong văn bản sẽ có cạnh nối từ đỉnh đó đến đỉnh trung tâm. Cạnh nối giữa các đỉnh được gán nhãn, thể hiện mối quan hệ giữa các đỉnh. Ví dụ khi chúng ta mô hình hoá một văn bản thì nhãn của cạnh có thể là: "tiêu đề", "chứa" như trong hình 2. Thế mạnh của mô hình đồ thị hình sao khi áp dụng vào bài toán phân lớp nói chung và đặc biệt trong phân loại email là nắm bắt được các thông tin cấu trúc của email (phần tiêu đề, phần nội dung), mối quan hệ giữa từ với các phần cấu trúc (đồng hiện của từ trong các phần tiêu đề, nội dung, ...).

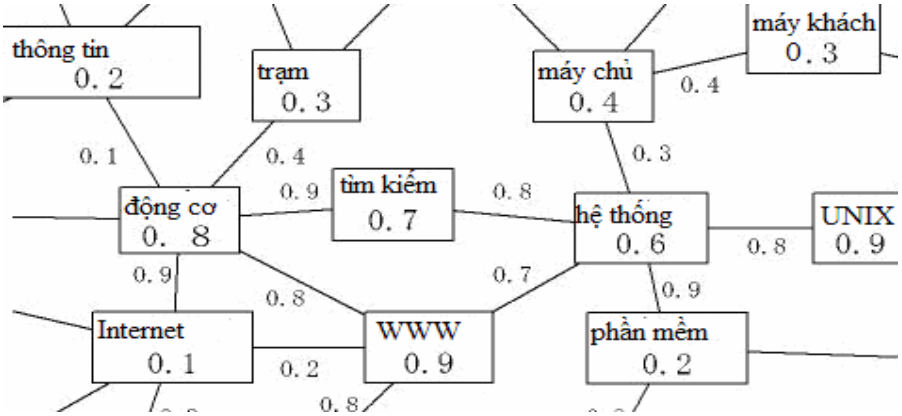


Hình 2. Ví dụ mô hình đồ thị hình sao

2.3. Mô hình đồ thị vô hướng sử dụng tần số xuất hiện

Trong mô hình đồ thị vô hướng sử dụng tần số xuất hiện, đỉnh và cạnh đều được gán nhãn, nhãn của đỉnh và cạnh là tần số xuất hiện của đỉnh và cạnh tương ứng. Nhãn đỉnh là tần số xuất hiện của từ trong văn bản. Cạnh được nối giữa hai đỉnh nếu hai từ xuất hiện chung trong tập hợp (câu hoặc nhóm từ hoặc trang) và có tần số xuất hiện chung lớn hơn ngưỡng cho phép. Nhãn cạnh là tần số xuất hiện chung của 2 từ trong tập hợp. Hình 3 là ví dụ mô hình đồ thị vô hướng sử dụng tần số xuất hiện. Ưu điểm của mô hình là khai thác được mối quan hệ giữa từ

với từ trong cấu trúc văn bản, cũng như tần số xuất hiện của từ và hỗ trợ cho quá trình tìm kiếm thông tin nhanh chóng.



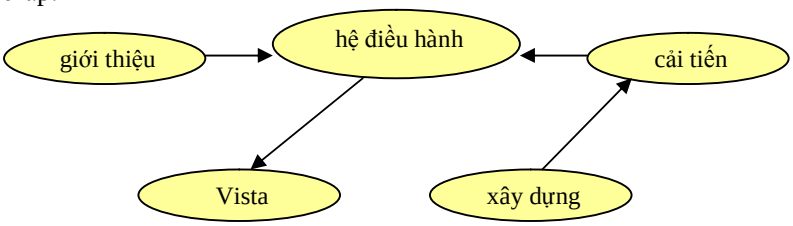
Hình 3. Ví dụ mô hình đồ thị vô hướng sử dụng tần số xuất hiện [11]

2.4. Mô hình đồ thị có hướng, cạnh không gán nhãn

Mô hình này còn được gọi là mô hình đồ thị đơn giản [8]. Mỗi đỉnh biểu diễn một từ riêng biệt và chỉ xuất hiện một lần trên đồ thị (ngay cả khi từ đó xuất hiện nhiều lần trong văn bản). Nhãn đỉnh là duy nhất và là tên của từ. Sau bước tiền xử lý văn bản, nếu từ "a" đứng ngay trước từ "b" sẽ có cạnh nối từ đỉnh "a" đến đỉnh "b" (không kể các trường hợp phân cách bởi dấu câu). Điểm mạnh của mô hình là lưu trữ được các thông tin cấu trúc như thứ tự xuất hiện, vị trí của từ trong văn bản và làm tăng hiệu quả của bài toán phân lớp cũng như gom cụm văn bản.

Ví dụ 2: Ta có câu 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 4 là mô hình biểu diễn văn bản trên sau khi đã qua bước loại bỏ bớt hư từ và các từ có trọng số thấp.



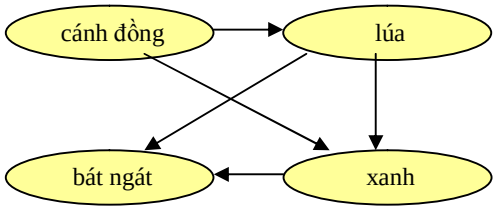
Hình 4. Ví dụ mô hình đồ thị đơn giản

2.5. Mô hình đồ thị có hướng, cạnh không gán nhãn, cạnh là khoảng cách n giữa hai từ trong văn bản

Mô hình này còn có tên gọi khác là mô hình khoảng cách n đơn giản. Trong cách biểu diễn này, người dùng cung cấp tham số n. Thay vì chỉ quan tâm từ "A" trực tiếp ngay trước từ "B", ta còn chú ý đến *n từ* đứng trước từ "B". Cạnh được xây dựng giữa hai từ khi giữa chúng có số từ xuất hiện nhiều nhất là (n-1) từ (ngoại trừ trường hợp các từ được phân cách bởi các dấu câu). Ưu điểm của mô hình là tận dụng được mối quan hệ giữa các từ, vùng lân cận của từ trong câu và có thể áp dụng vào bài toán phân lớp văn bản.

Ví dụ 3: Ta có câu sau: "*Cánh đồng lúa xanh bát ngát*".

Với n=2, hình 5 là mô hình biểu diễn câu trên.



Hình 5. Ví dụ mô hình đồ thị khoảng cách n đơn giản

Các mô hình còn lại là biến thể của các mô hình trên với các khác biệt đã được mô tả trong bảng 1.

3. HỆ THỐNG PHÂN LỚP VĂN BẢN TIẾNG VIỆT

Phân lớp văn bản là quá trình gán văn bản vào một hoặc nhiều chủ đề đã xác định trước. Phân lớp văn bản tiếng Việt là một lĩnh vực nghiên cứu quan trọng, được quan tâm trong thời gian gần đây. Tiếng Việt khác với tiếng Anh ở chỗ ranh giới giữa các từ không phải chỉ là những khoảng trắng và nó đòi hỏi phải xử lý tách từ trước. Bản thân bài toán tách từ trong tiếng Việt là bài toán khó. Khó khăn thứ hai là chưa có kho dữ liệu chuẩn cho tiếng Việt như Reuter, NewGroups,… để có thể so sánh kết quả phân lớp. Gần đây, đã có một số tiến triển đáng kể trong bài toán phân lớp văn bản tiếng Việt [3] [10]. Tuy nhiên, các công trình nghiên cứu này đều dựa trên mô hình không gian vectơ.

Nhằm tận dụng các ưu điểm của mô hình đồ thị, chúng tôi xây dựng thử nghiệm hệ thống phân lớp văn bản tiếng Việt dựa vào mô hình đồ thị biểu diễn văn bản và sử dụng thuật toán khai thác đồ thị con phổ biến để xác định đặc trưng cho từng chủ đề. Để tránh phụ thuộc vào bài toán tách từ và vì đơn vị từ được tạo thành bởi một hay nhiều tiếng [2], chúng tôi sử dụng tiếng để làm đỉnh của đồ thị.

Trong quá trình huấn luyện, đầu vào của hệ thống là tập văn bản huấn luyện D = {d_1, d_2, …, d_n} phân chia theo chủ đề và tập chủ đề C = { c_1, c_2, …, c_r}. Trong quá trình phân lớp, văn bản mới sẽ được xác định chủ đề dựa trên sự tương tự với các đặc trưng. Hình 6 là mô hình chính của hệ thống phân lớp.

Trong đó:

- (b): Mô hình hoá văn bản trong D thành tập đồ thị G = {g_1, g_2, …, g_n}. Chúng tôi dùng mô hình đồ thị đơn giản với mỗi tiếng là một đỉnh trong đồ thị. Với ưu điểm của mô hình đồ thị, nếu chúng ta tách tiếng mà không cần tách từ thì vẫn lưu giữ được cấu trúc của từ trong văn bản.

- (c): Trong từng chủ đề, chúng ta tìm tập đồ thị con phổ biến có tần số xuất hiện lớn hơn ngưỡng phổ biến tối thiểu minsupp. Chúng tôi sử dụng thuật toán gSpan [12] để tìm các đồ thị con phổ biến do đây là thuật toán được đánh giá là nhanh và có thể biến đổi phù hợp với mô hình đồ thị có hướng. Nhiệm vụ phức tạp nhất trong bài toán khai thác đồ thị con phổ biến là vấn đề đẳng cấu đồ thị, có độ phức tạp NP khi nhãn đỉnh không duy nhất. Tuy nhiên, với mô hình biểu diễn văn bản bằng đồ thị đơn giản và nhãn đỉnh là duy nhất thì độ phức tạp của thuật toán giảm xuống còn $O(n^2)$.

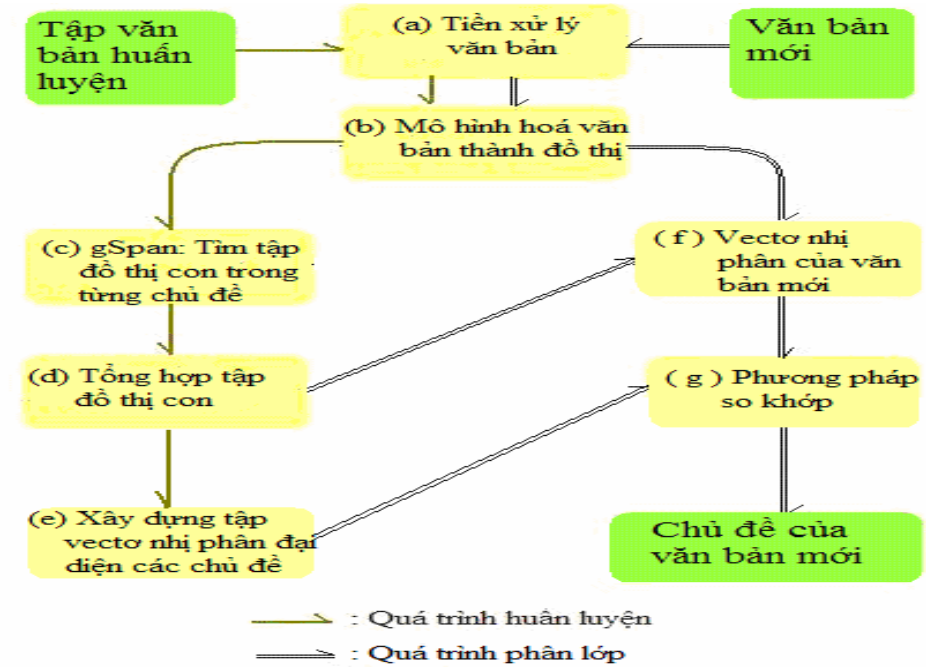
- (d): Tổng hợp đồ thị con trong tất cả các chủ đề, ta có tập đồ thị con phổ biến S = {s_1, s_2, …, s_m}

- (e): Xây dựng vectơ đặc trưng cho từng chủ đề và là vectơ nhị phân m chiều thông qua tập S. Nếu đồ thị con phổ biến thuộc S xuất hiện trong tập đồ thị con phổ biến của chủ đề thì đặc trưng tương ứng của vectơ nhận giá trị 1 và ngược lại. Chúng ta xây dựng được tập vectơ đặc trưng nhị phân $F = \{f_1, f_2, \ldots, f_r\}$.

- (g): Văn bản mới được biểu diễn thành đồ thị, sau đó chuyển thành vectơ nhị phân v_0 có m chiều tương ứng với m đồ thị con phổ biến của tập S. Chúng tôi sử dụng phương pháp so khớp với độ đo Dice [4] để tính khoảng cách giữa vectơ v_0 và vectơ đặc trưng chủ đề. Văn bản mới thuộc chủ đề cho độ đo có giá trị lớn nhất. Công thức tính độ đo Dice giữa vectơ đặc trưng chủ đề và vectơ v_0 :

$$Dice(v_0, f_j) = \frac{2|v_0 \wedge f_j|}{|v_0| + |f_j|} \quad (1)$$

Trong đó: $f_j \in F$, $|v_0|$, $|f_j|$: tổng số đặc trưng mang giá trị 1 của v_0, f_j.



Hình 6. Sơ đồ hệ thống phân lớp văn bản

4. KẾT QUẢ THỬ NGHIỆM

Để đánh giá mô hình biểu diễn văn bản bằng đồ thị, chúng tôi thu thập bộ dữ liệu bao gồm 2500 tập tin văn bản (là tóm tắt bài báo lấy từ một số báo điện tử như VnExpress[1], TuoiTre Online[2], ThanhNien Online[3]). Bộ dữ liệu bao gồm 6 chủ đề như trong bảng 2. Sau khi tiền xử lý văn bản (gồm các bước như tách câu, tách tiếng, loại bỏ hư từ) chúng tôi thu được trung bình 40 đỉnh/đồ thị.

[1] http://www.vnexpress.net
[2] http://www.tuoitre.com.vn
[3] http://www.thanhnien.com.vn

Để đánh giá kết quả phân lớp, 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 [13]. Chúng tôi sử dụng phương pháp đánh giá chéo (k-fold validation) để chạy thử nghiệm trên máy tính Pentium 1.5G và bộ nhớ 256MB.

Bảng 2. Tập dữ liệu huấn luyện

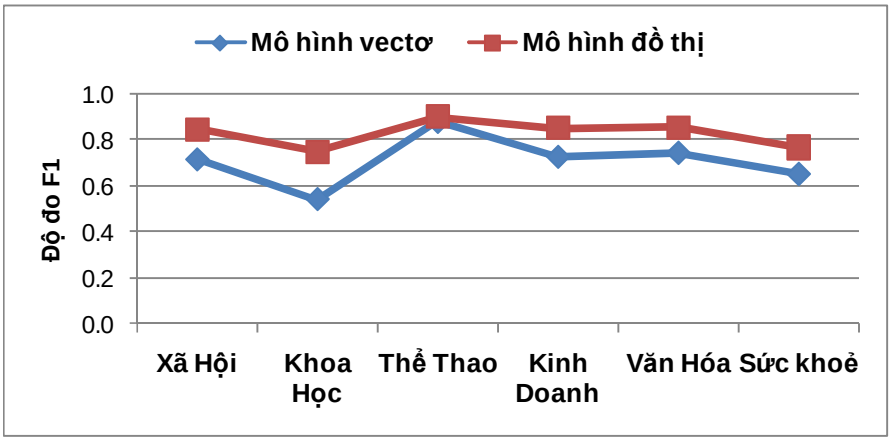
STT	Tên chủ đề	Số văn bản
1	Xã Hội	400
2	Khoa Học	350
3	Thể Thao	450
4	Kinh Doanh	450
5	Văn Hóa	400
6	Sức khoẻ	450

Kết quả thử nghiệm được trình bày trong bảng 3 với thời gian huấn luyện trung bình là 2.8 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 0.9 giây / văn bản.

Bảng 3. Kết quả thử nghiệm (5-fold validation)

Tên chủ đề	Độ phủ (Recall)	Độ chính xác (Precision)	Độ đo F1
Xã Hội	0.79	0.915	0.848
Khoa Học	0.705	0.8	0.75
Thể Thao	0.86	0.946	0.901
Kinh Doanh	0.866	0.843	0.854
Văn Hóa	0.8	0.941	0.856
Sức khoẻ	0.702	0.85	0.769
Trung bình	**0.787**	**0.888**	**0.833**

Chúng tôi cài đặt thuật toán k-láng giềng gần nhất (k-NN) trên mô hình không gian vectơ với độ đo Cosine [7] để so sánh với mô hình biểu diễn văn bản bằng đồ thị của chúng tôi. Hình 7 là đồ thị so sánh kết quả phân lớp theo từng mô hình trên các chủ đề. Mô hình biểu diễn văn bản bằng đồ thị cho kết quả phân lớp tốt hơn.



Hình 7. Kết quả phân lớp theo chủ đề

5. KẾT LUẬN

Bài báo nghiên cứu và tổng hợp các mô hình biểu diễn văn bản thành đồ thị. Chúng tôi đã xây dựng thử nghiệm hệ thống phân lớp văn bản tiếng Việt dựa trên mô hình biểu diễn văn bản bằng đồ thị. Mô hình đồ thị cho phép lưu trữ các thông tin cấu trúc quan trọng của văn bản như vị trí, sự đồng hiện hay thứ tự của từ. Kết quả thử nghiệm cho thấy mô hình đồ thị cho kết quả phân lớp tốt hơn mô hình không gian vectơ truyền thống. Để đánh giá chính xác hơn nữa, chúng tôi dự kiến sẽ thu thập và xây dựng bộ dữ liệu thử nghiệm lớn. Đồng thời, chúng tôi dự kiến sẽ thử nghiệm áp dụng các loại mô hình đồ thị khác nhau vào bài toán phân lớp để xác định loại mô hình phù hợp nhất.

GRAPH – BASED MODEL FOR TEXT REPRESENTATION

Nguyen Hoang Tu Anh, Nguyen Tran Kim Chi, Nguyen Hong Phi
University of Science, VNU-HCM

ABSTRACT: *Text representation models are very important pre-processing step in various domains such as text mining, information retrieval, natural language processing. In this paper we summarize graph-based text representation models. Graph-based model can capture structural information such as the location, order and proximity of term occurrence, which is discarded under the standard text vector representation models. We have tested this graph model in Vietnamese text classification system.*

Keyword*: Graph model, text representation, text classification.*

TÀI LIỆU THAM KHẢO

[1]. Aery M., *INFOSIFT: adapting graph mining techniques for document classification,* University of Texas at Arlington, 12/2004.

[2]. Đinh Điền, *Xử lý Ngôn ngữ tự nhiên,* NXB Đại học Quốc gia Tp. HCM, (2004).

[3]. Đỗ Phúc, *Nghiên cứu ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét ngữ nghĩa,* Tạp chí Phát triển Khoa học & Công nghệ, Tập 9, số 2, pp.23-32, (2006).

[4]. Khreisat L., *Arabic Text Classification Using N-Gram Frequency Statistics _ a Comparative Study,* WORLDCOMP'06 – DMIN'06, (2006).

[5]. Markov A., Last M., *A Simple, Structure-Sensitive Approach for Web Document Classification*, Proc. of AWIC 2005, LNAI 3528, pp. 293-298, (2005).

[6]. Mihalcea R., Tarau P., *TextRank: Bringing Order into Texts*, Proc. of EMNLP'04, pp.404-411, (2004).

[7]. Salton G., *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, (1989).

[8]. Schenker A., Last M., Bunke H., Kandel A, *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 Recognition, Vol.18, No.3, pp. 475-479, (2004).

[9]. Sowa J.F., *Conceptual Graphs for a DataBase Interface*, IBM Journal of Research and Development 20(4), 336–357, July, (1976).

[10]. Thanh V. Nguyen, Hoang K. Tran, Thanh T.T. Nguyen, Hung Nguyen, *Word Segmentation for Vietnamese Text Categorization*, Poster Proc. of RIVF'06, pp.113-118, (2006).

[11]. Tomita J., NakawataseH., Ishii M., *Graph-based Text Database for Knowledge Discovery*, Poster Proc. of WWW'04, pp. 454–455, (2004).

[12]. Yan X., Han J., *gSpan: Graph-Based Substructure Pattern Mining*, Proc. of IEEE ICDM'02, pp.721-723, (2002).

[13]. Yang Y., Liu X., *A re-examination of text categorization methods*, Proc. of ACM SIGIR'99, pp. 42-49, (1999).

[14]. Zha H., *Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering*, Proc. of ACM SIGIR'02, pp113-200, (2002).

[15]. http://www.jfsowa.com/cg/cgexamp.htm