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

Đỗ Phúc
Trung tâm Phát triển Công nghệ Thông tin, ĐHQG-HCM
(Bài nhận ngày 25 tháng 08 năm 2005, hoàn chỉnh sửa chữa ngày 27 tháng 02 năm 2006)

TÓM TẮT : *Bài báo trình bày một số kết quả nghiên cứu ứng dụng các thuật toán tìm tập phổ biến và luật kết hợp vào bài toán phân lớp văn bản. Mô hình vector có thành phần là các cụm danh từ phổ biến được dùng để đặc trưng văn bản. Thuật toán tách từ, gán nhãn từ loại được sử dụng để rút trích các cụm danh từ. Thuật toán tập phổ biến và luật kết hợp được sử dụng để tạo đồ thị đồng hiện các từ trong ngữ cảnh nhất định nhằm xác lập nghĩa của từ trong văn bản và kết hợp với từ điển đồng nghĩa, gần nghĩa để điều chỉnh thành phần của vector văn bản nhằm nâng cao khả năng phân lớp văn bản có xem xét ngữ nghĩa. Ngoài ra, luật kết hợp có vế phải là các thuộc tính phân lớp sẽ được sử dụng để làm luật phân lớp. Chúng tôi đã thử nghiệm giải pháp đề xuất vào bài toán phân lớp các tóm tắt bài báo khoa học trong lĩnh vực CNTT tiếng Việt*

Từ Khoá: *Cụm danh từ, Đồ thị đồng hiện, Luật kết hợp, Luật phân lớp, Tập phổ biến*

1.GIỚI THIỆU

Với sự xuất hiện của Internet, khối lượng thông tin chủ yếu và chiếm trên 80% vẫn là các thông tin văn bản. Các phương pháp phân loại văn bản trước đây đều dựa trên tiếp cận máy học, mô hình xác suất,cây quyết định, qui nạp thuộc tính, người láng giềng gần nhất, và mới đây là phương pháp support vector machine [11]. Các thuật toán này thường tập trung vào bài toán phân làm 2 lớp và gặp khó khăn với khối lượng dữ liệu lớn. Trong bài báo này, chúng tôi nghiên cứu dùng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt gồm a)Đặc trưng văn bản: bao gồm tìm dãy từ phổ biến trong tập ngữ liệu văn bản và tạo đồ thị đồng hiện nhằm xác lập nghĩa của từ đặc trưng b) Tạo luật phân lớp văn bản. Bài báo được tổ chức như sau: 1) Giới thiệu 2) Bài toán tìm tập phổ biến và luật kết hợp 3) Phân lớp văn bản bằng luật kết hợp 4) Tạo vector đặc trưng cho văn bản 5) Xây dựng bộ phân lớp văn bản 6) Thử nghiệm 7) Kết luận

2. BÀI TOÁN TÌM TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP

2.1.Các khái niệm cơ bản

Định nghĩa 1: Ngữ cảnh khai thác dữ liệu

Cho tập O là tập hữu hạn khác rỗng các giao tác và I là tập hữu hạn khác rỗng các mặt hàng, R là một quan hệ hai ngôi giữa O và I sao cho với $o \in O$ và $i \in I$, $(o,i) \in R \Leftrightarrow$ giao tác o có chứa mặt hàng i. Ngữ cảnh khai thác dữ liệu (dưới đây sẽ gọi tắt là NCKTDL) là bộ ba (O,I,R).

Định nghĩa 2: Các kết nối Galois

Cho NCKTDL (O, I, R), xét hai kết nối Galois ρ và λ được định nghĩa như sau:

$\rho: P(I) \rightarrow P(O)$ và $\lambda : P(O) \rightarrow P(I)$:

Cho $S \subset I$, $\rho(S) = \{o \in O \mid \forall i \in S, (o,i) \in R\}$

Cho $X \subset O$, $\lambda(X) = \{i \in I \mid \forall o \in X , (o,i) \in R\}$

Trong đó P(X) là tập các tập con của X.

Cặp hàm (ρ , λ) được gọi là kết nối Galois. Giá trị $\rho(S)$ biểu diễn tập các giao tác có chung tất cả các mặt hàng trong S. Giá trị $\lambda(X)$ biểu diễn tập mặt hàng có trong tất cả các giao tác của X.

Định nghĩa 3: Tập mặt hàng phổ biến

Cho NCKTDL (O,I,R) và minsupp $\in$ (0,1] là ngưỡng phổ biến tối thiểu. Cho S $\subset$ *I*, độ phổ biến của S ký hiệu là SP(S) là tỉ số giữa số các giao tác có chứa S và số lượng giao tác trong O. Nói cách khác SP(S)= $|\rho(S)|/|O|$.

Cho S $\subset$ I , S là một tập các mặt hàng phổ biến theo ngưỡng minsupp nếu và chỉ nếu SP(S) $\geq$ minsupp. Trong các phần sau tập mặt hàng phổ biến sẽ được gọi tắt là tập phổ biến. Ký hiệu FS(O,I,R,minsupp) = { S $\in$ P(I) | SP(S) $\geq$ minsupp }

Định nghĩa 4: Luật kết hợp

Cho NCKTDL (O,I,R) và ngưỡng minsupp $\in$(0,1]. Với một S$\in$ FS(O,I,R,minsupp), gọi X và Y là các tập con khác rỗng của S sao cho S = X$\cup$Y và X $\cap$Y=$\varnothing$. Luật kết hợp X với Y có dạng X$\rightarrow$Y phản ánh khả năng khách hàng mua tập mặt hàng Y khi mua tập mặt hàng X. Độ phổ biến của luật kết hợp X$\rightarrow$Y với S= X$\cup$Y là SP(S). Độ tin cậy của luật kết hợp X$\rightarrow$Y được ký hiệu là CF(X$\rightarrow$Y) và được tính bằng công thức CF(X$\rightarrow$Y)=SP(X$\cup$Y)/SP(X)

Nguyên lý Apriori:

- Cho S $\in$ FS(O,I,R,minsupp), nếu T $\subseteq$ S thì T $\in$ FS(O,I,R,minsupp)
- Cho T $\notin$ FS(O,I,R,minsupp), nếu T $\subseteq$ S thì S $\notin$ FS(O,I,R,minsupp)

2.2. Tìm tập phổ biến

Cho NCKTDL (O,I,R) và minsupp$\in$(0,1], tìm FS(O,I,R,minsupp). Thuật toán được xây dựng dựa trên nguyên lý Apriori [3],[10]. Đầu tiên thuật toán sẽ tìm các tập phổ biến có một phần tử. Sau đó các ứng viên của các tập phổ biến có hai phần tử sẽ được tạo lập bằng cách hợp các tập phổ biến có một phần tử. Một cách tổng quát, các tập ứng viên của tập phổ biến có k phần tử sẽ được tạo từ các tập phổ biến có k-1 phần tử. Gọi F_k ={S$\in$ P(I) | SP(S) $\geq$ minsupp và |S|= k }. Thuật toán sẽ duyệt từng ứng viên để tạo F_k bao gồm các ứng viên có độ phổ biến lớn hơn hoặc bằng ngưỡng minsupp.

2.3. Tìm luật kết hợp

Cho NCKTDL (O,I,R) và hai ngưỡng phổ biến minsupp$\in$[0,1] và ngưỡng tin cậy minconf$\in$(0,1], tìm tất cả các luật kết hợp r có CF(r) $\geq$ minconf và SP(r) $\geq$minsupp.
Chi tiết thuật toán tìm tập phổ biến theo nguyên lý Apriori [3],[10]:

3. PHÂN LỚP VĂN BẢN BẰNG LUẬT KẾT HỢP

3.1. Bảng quyết định

Đinh nghĩa 5. Bảng quyết định

Xét NCKTDL (O,D,R) với D =I $\cup$ C , I $\cap$ C=$\varnothing$, trong đó I là tập các mặt hàng và C là tập các nhãn xác định nhóm. Bộ ba (O, D=I $\cup$ C, R) được gọi là một bảng quyết định Lưu ý trong trường hợp |C| > 2 sẽ là bài toán phân thành nhiều lớp.

3.2 Luật phân lớp trên bảng quyết định

Định nghĩa 6. Luật phân lớp

Cho bảng quyết định (O, D=I $\cup$ C,R) và các ngưỡng minsupp, minconf, tìm các luật kết hợp có dạng r: S$\rightarrow${c}. với S $\subseteq$ I và c$\in$C . Có thể dựa vào luật kết hợp này làm các luật phân lớp dữ liệu. Theo định nghĩa về độ tin cậy của luật kết hợp r: S$\rightarrow${c} được định nghĩa là : CF(r)=$\frac{|\rho(S)\cap\rho(\{c\})|}{\rho(S)}$ và ρ(S) là tập các giao tác có chứa các mặt hàng trong S, ρ({c})

là tập các giao tác thuộc lớp c do đó $\rho(S) \cap \rho(\{c\})$ sẽ xác định các giao tác thuộc lớp c và có chứa các mặt hàng trong S. Do vậy có thể sử dụng độ tin cậy của luật kết hợp để đánh giá độ chính xác của luật phân lớp. Nếu CF(r) càng dần về 1,0 thì độ chính xác của phân lớp càng tăng. Khi CF(r) =1 thì $\rho(S) \subseteq \rho(\{c\})$, lúc này luật phân lớp có độ chính xác phân lớp là 100%. Khi áp dụng vào bài toán phân lớp văn bản, mỗi văn bản sẽ tương ứng với một giao tác, mỗi mặt hàng sẽ tương ứng với một từ đặc trưng (sẽ được giải thích trong mục đặc trưng văn bản).

3.3. Rút gọn luật phân lớp

Trong quá trình tìm luật phân lớp từ luật kết hợp, chúng ta có thể tìm được rất nhiều luật phân lớp. Để rút gọn luật phân lớp, chúng tôi chọn các luật có độ tổng quát cao hơn. Chi tiết như sau:

Định nghĩa 7.Cho hai luật phân lớp r1: p1$\rightarrow$ c , r2: p2$\rightarrow$ c. Luật r1 được gọi là tổng quát hơn r2 nếu và chỉ nếu $\rho(p2) \subseteq \rho(p1)$.

Ví dụ 1: Cho hai luật

R1:{khoá, phụ_thuộc_hàm}$\rightarrow$ { Lớp_CSDL}
R2:{khoá, phụ_thuộc_hàm, dạng-chuẩn}$\rightarrow$ { Lớp_CSDL}
Luật R1 thì tổng quát hơn luật R2 vì:

{khoá, phụ_thuộc_hàm}$\subseteq$ {khoá, phụ_thuộc_hàm, dạng-chuẩn}

Trong quá trình tạo luật phân lớp, ta có thể gặp rất nhiều luật phân lớp. Do vậy cần tiến hành rút gọn bộ luật phân lớp bằng cách loại bỏ các luật phân lớp thừa.

Định nghĩa 8. Cho hai luật R1 và R2, R1 được xếp hạng cao hơn R2 nếu:
(1) CF(R1) > CF(R2)
(2) CF(R1) = CF(R2) nhưng SP(R1) > SP(R2)
(3) CF(R1) = CF(R2) và SP(R1) > SP(R2) , nhưng vế trái của R1 có chứa ít từ khóa hơn vế trái của R2

Thuật toán 1: Rút gọn luật phân lớp **Vào:** tập luật phân lớp R **Ra:** Tập luật rút gọn
1) Sắp xếp các luật theo độ tổng quát (định nghĩa 7) 2) For each r in R 3) Tìm tất cả các luật có hạng nhỏ hơn r (định nghĩa 8) và loại bỏ khỏi R các luật có độ tin cậy nhỏ hơn r. 4) Endfor 5) For each r in R 6) Quét CSDL và tìm các giao tác thỏa luật r. 7) Nếu luật r phân lớp đúng tối thiểu cho một mẫu học thì chọn r. 8) Loại khỏi CSDL các bộ thỏa luật r. 9) Endfor 10) Return R && tập luật rút gọn

4. TẠO VECTƠ ĐẶC TRƯNG VĂN BẢN

4.1. Tìm dãy từ phổ biến Thuật toán tìm tập phổ biến được ứng dụng để tìm dãy từ phổ biến trong tập dữ liệu gồm nhiều văn bản. Mỗi văn bản được xem là một giao tác. Một tập mặt hàng $\{i_1 , i_2 , \dots , i_k\}$ với $i_1, i_2 , \dots , i_k$ là các mặt hàng sẽ trở thành dãy các từ $i_1 i_2 \dots i_k$ với $i_1, i_2 , \dots , i_k$ là các từ theo nghĩa có dấu cách hoặc dấu chấm câu đi trước và đi sau từ đó. Một văn bản sẽ hỗ trợ (mức độ phổ biến) cho dãy từ $i_1 i_2 \dots i_k$ nếu tồn tại một câu trong văn bản đó có chứa dãy từ $i_1 i_2 \dots i_k$. Thuật toán tìm tập phổ biến được cải tiến như sau:

1. Tạo F_1 tập các dãy từ chỉ chứa 1 từ và có độ phổ biến lớn hơn ngưỡng minsupp

2. Dùng thuật toán tìm tập phổ biến. Lưu ý phép hợp các tập phổ biến $S = X \cup Y$ với X, Y là các tập mặt hàng phổ biến có k-1 mặt hàng trở thành phép nối chuỗi, trong đó X lấy từ dãy phổ biến có k-1 từ và Y là dãy phổ biến có 1 từ (lấy từ F_1)

2. Trích cụm danh từ

Để tìm cụm danh từ trong văn bản, chúng ta tiến hành các bước sau: tách từ , gán nhãn từ loại, nhóm các từ đã được gán nhãn từ loại thành cụm danh từ.

4.2.1. Tách từ

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 từ "tin học". Sau khi thử nghiệm một số chương trình tách từ, chúng tôi sử dụng chương trình tách từ theo mô hình lai (mô hình WFST kết hợp mạng nơron) của nhóm nghiên cứu [5] vì kết quả tách từ đạt độ chính xác cao và được sự hỗ trợ kỹ thuật của các tác giả. Tiếp cận tách từ tiếng Việt trong [5] là một bài toán thống kê chuyển đổi trạng thái. Đầu tiên câu được xử lý loại bỏ các lỗi về cách trình bày một câu, và chuẩn hóa về cách bỏ dấu, cách viết các ký tự y, i…trong tiếng Việt. Sau đó, câu được đưa vào mô hình WFST (Weighted Finite State Transducer) để nhận diện từ láy, danh từ riêng, tên riêng người Việt, tên riêng người nước ngoài,.. Mô hình thực hiện tách câu thành các từ đi liền nhau theo các trạng thái có thể, nhận diện từ và gán trọng số thích hợp số thích hợp dựa vào tự điển (trọng số ước lượng thường rất nhỏ nên lấy log (=-log(tần suất từ/kích thước tập mẫu)).

Mô hình WFST căn cứ trên các trọng số này để chọn ra một cách tách từ thích hợp. Sau khi có được tất cả trạng thái tách từ có thể có của câu, với mỗi trạng thái, mô hình tính tổng trọng số và chọn trạng thái tách từ đúng nhất là câu có tổng trọng số nhỏ nhất.

Ví dụ 2:

Câu = "Hai công ty vừa ký kết hợp đồng sản xuất."
Sau khi qua công đoạn tách từ ta có các từ tiếng Việt trong cặp dấu ngoặc như sau:
(Ha) (công ty) (vừa) (ký kết) (hợp đồng)(sản xuất)

4.2.2. Gán nhãn từ loại bằng phần mềm VnQTag

Chúng tôi sử dụng chương trình VnQTag của nhóm tác giả [8] để gán nhãn từ loại tự động cho văn bản. Chương trình VnQTag được nhóm tác giả trên chỉnh sửa lại thành phiên bản dùng cho tiếng Việt từ phần mềm QTAG của nhóm tác giả O. Mason, Đại học Bermingham, Anh. QTAG là một bộ gán nhãn xác suất độc lập với ngôn ngữ. Phương pháp xử lý của QTAG có thể mô tả tổng quát như sau. Nó được xây dựng theo tiếp cận máy học từ khối ngữ liệu học đã được gán nhãn bằng tay. Dựa vào những dữ liệu đã học được này, bộ gán nhãn tìm những nhãn có thể được và tần số của nó cho từng từ trong kho dữ liệu mới đã được tách từ. Nếu việc tìm kiếm một từ trong danh sách từ vựng đã học thất bại thì tất cả các nhãn sẽ được gán cho từ đó. Cuối cùng, bộ gán nhãn thực hiện bước loại bỏ nhập nhằng bằng cách sử dụng thông tin về xác suất phân bố từ vựng đã được học trước đó.

Dữ liệu đầu vào của chương trình VnQTAG là văn bản đã được phân tách từ trong từng câu (kết quả của bước tách từ ở phần trên), kết quả đầu ra của chương trình là một từ loại tương ứng sẽ được gán cho từng từ trong văn bản. Hệ thống sử dụng đồng thời từ điển để liệt kê các từ loại có thể cho một từ, và một kho văn bản mẫu để loại bỏ nhập nhằng.

Cùng với chương trình VnQTAG, tác giả [8] đã cung cấp một tự điển, một tập dữ liệu huấn luyện khoảng gần 100.000 từ bộ chú thích (bộ *tag*) từ loại gồm các chú thích cho: Danh từ (N), Động từ (V), Tính từ (A), Đại từ (P), Từ chỉ định (D), Trạng từ (R), Trạng từ vị trí (S), Liên từ (C), Số (M), Thán từ (I), Còn lại (X).

4.2.3. Trích cụm danh từ

Trong tiếng Anh để gộp các từ thành cụm danh từ, chúng tôi sử dụng giải pháp được nêu trong [2],[11] trong đó cụm danh từ được định nghĩa là chuỗi gồm có danh từ hay tính từ và tận cùng bằng danh từ. Công thức tổng quát của cụm danh từ tiếng Anh là {danh từ, tính từ} * { danh từ}. Ví dụ cụm từ "computer science" là một cụm danh từ trong đó "computer" và "science" đều là danh từ, cụm từ "great man" là một cụm danh từ trong đó "great" là tình từ và "man" là danh từ. Dựa trên cấu trúc của cụm danh từ tiếng Việt được trình bày trong [4], chúng tôi xây dựng các công thức sau để rút trích cụm danh từ trong văn bản tiếng Việt đã được gán nhãn từ loại.

- Cụm danh từ gồm danh từ và danh từ đi liền sau nó: N+N (ví dụ 'cơ sở dữ liệu').
- Cụm danh từ gồm danh từ, danh từ và danh từ đi liền sau nó: N+N+N (ví dụ 'hệ thống thông tin địa lý').
- Cụm danh từ gồm danh từ và tính từ đi liền sau nó: N+A (ví dụ 'dữ liệu lớn').
- Cụm danh từ gồm danh từ, danh từ và tính từ đi liền sau nó: N+N+A (ví dụ 'cơ sở dữ liệu lớn').
- Cụm danh từ gồm danh từ và động từ đi liền sau nó: N+V (ví dụ 'phép ánh xạ').
- Cụm danh từ gồm danh từ, động từ và danh từ đi liền sau nó: N+V+N (ví dụ 'hệ thống chuyển thông điệp') .

Chúng tôi cũng sử dụng một từ điển chuyên ngành theo lĩnh vực áp dụng để nhận dạng đúng các cụm danh từ được tách.

4.3. Tạo vector đặc trưng văn bản

Khối ngữ liệu văn bản được phân tích để tìm các cụm danh từ phổ biến. Gọi M là số số văn bản trong khối ngữ liệu cần xem xét, N là số từ /cụm từ đặc trưng của khối dữ liệu, f_{ik} là tần số xuất hiện của từ/cụm từ đặc trưng thứ k trong văn bản i, n_k là số văn bản có chứa từ/cụm từ đặc trưng.. Hệ số tf-idf (term frequency, inversed document frequency) để gán trọng cho từ/cụm từ thứ k trong văn bản i như sau:

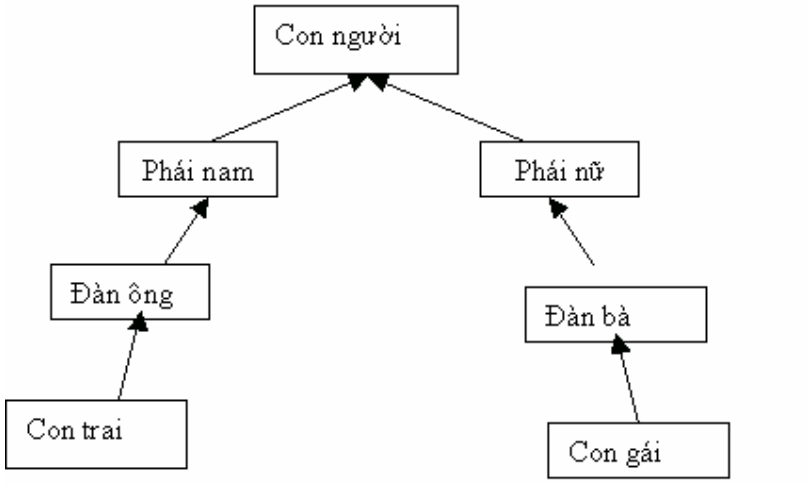
$$a_{ik} = f_{ik} \times \log\left(\frac{M}{n_k}\right)$$

Chúng tôi chọn một ngưỡng để biến đổi vector đặc trưng cho văn bản thành vector nhị phân. Thành phần thứ k của vector đặc trưng cho văn bản thứ i có trị 1 nếu $a_{ik} \geq$ Ngưỡng và có trị 0 nếu ngược lại.

4.4. Điều chỉnh thành phần của vector văn bản

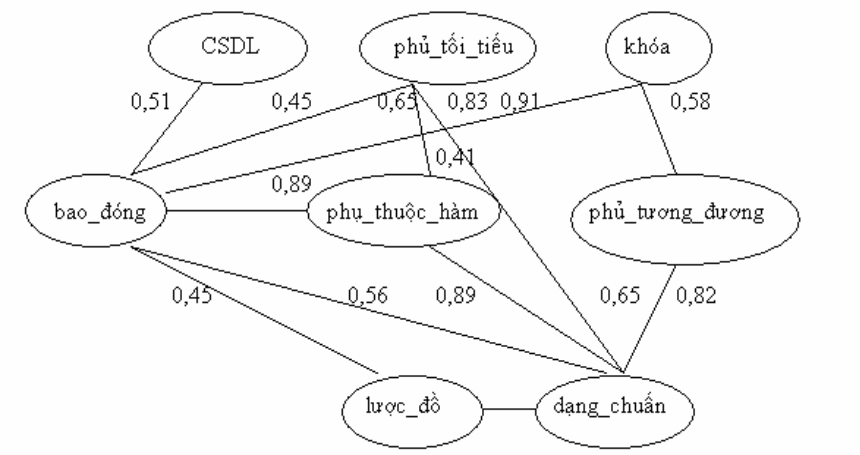
Trong tiến trình phân lớp, cần có sự so sánh giữa vector đặc trưng cho văn bản cần xếp lớp với từng vector đặc trưng lớp được tạo trong quá trình học. Các thành phần vector là các từ đặc trưng và có thể đồng nghĩa, hay gần nghĩa với nhau. Ví dụ vector thứ nhất có thành phần ứng với từ "con_người", vector thứ hai có thành phần ứng với từ "nhân_loại", rõ ràng hai từ con_người và nhân_loại gần nghĩa nhau.

Do đó cần tiến hành điều chỉnh các thành phần này trước khi đưa vào bộ phân loại. Đối với tiếng Anh, hiện có từ điển Wordnet [7] trong đó lưu trữ các tập từ đồng nghĩa và các quan hệ ngữ nghĩa (nghĩa rộng, nghĩa hẹp). Đối với tiếng Việt, chúng tôi bước đầu xây dựng một hệ thống tựa Wordnet cho tiếng Việt. Hình 1 là một đồ thị biểu diễn quan hệ "là một loại của" của các từ con người, phái nam, phái nữ, đàn ông, đàn bà, con trai,con gái..



Hình 1. Đồ thị quan hệ nghĩa rộng/nghĩa hẹp giữa các danh từ

Dựa vào khoảng cách giữa các từ trên cây có thể khẳng định hai từ đó có gần nghĩa hay không, ví dụ nếu khoảng cách là 4 thì ”con trai” và ”con gái” là gần nghĩa nhau do đó thành phần tương ứng trong vector đặc trưng văn bản sẽ được điều chỉnh.
Một trong những vấn đề cần xác định trước khi so sánh hai từ có đồng nghĩa hay gần nghĩa là vấn đề xác lập nghĩa của từ. Ví dụ từ ”khóa” có thể có nhiều nghĩa như: khóa học, khóa trong quan hệ của cơ sở dữ liệu, ổ khoá .... Hiện nay có nhiều cách tiếp cận để xác lập nghĩa của từ, chúng tôi chọn giải pháp được nêu trong [1],[12]. Tác giả đã xây dựng đồ thi các từ xuất hiện đồng thời với từ cần xét. Ví dụ : nếu “khóa” xuất hiện đồng thời với các từ như ”cơ sở dữ liệu”, ”quan hệ”, ”phụ thuộc hàm”….. thì nghĩa của khóa là khoá trong quan hệ của cơ sở dữ liệu (xem hình 2).



Hình 2: Một phần của đồ thị đồng hiện các từ đăc trưng

Chúng tôi tạo đồ thị đồng hiện như sau: Cho O là tập văn bản và FT(O) là tập các từ phổ biến đặc trưng cho các văn bản trong O. Gọi G=(V,E) là đồ thị không có hướng trong đó V là tập các cụm danh từ phổ biến V=FT(O). Đồ thị G(V,E) được tạo bằng cách sử dụng luật kết hợp các dãy từ phổ biến được khai thác từ khối ngữ liệu và sử dụng ngưỡng liên kết để tìm các miền liên thông trên đồ thị đồng hiện bằng cách loại bỏ các các cung có trọng liên kết nhỏ hơn ngưỡng. Trọng liên kết giữa cung nối hai từ hai từ a và b là $W_{a,b}$ =(1/2)(CF(a $\rightarrow$ b) + CF(b $\rightarrow$ a)). Sau đó dùng thuật giải cây bao trùm tối thiểu để tạo các cụm có mức độ gắn kết chặt (độ đồng hiện cao) và gán nhãn cho cụm. Các cụm được đặc trưng bởi các tập các từ có

trong đồ thị đặc trưng cho cụm, tập từ này được gọi là tập từ đặc trưng cho cụm. Mỗi cụm sẽ xác định nghĩa của từ. Mỗi cụm này sẽ được gán nhãn ngữ nghĩa bằng tay.

Ví dụ 2: Cụm cơ sở dữ liệu được đặc trưng bằng tập các từ trong bảng 1 :

Bảng 1: Tập từ đặc trưng cho ngữ cảnh của một số nhóm tiêu biểu :

Tập từ đặc trưng đồng hiện	**Nhóm**
CSDL, phụ_thuộc_hàm, khóa,lược_đồ_quan_hệ, dạng_chuẩn, bao_đóng, phụ_thuộc_đa_trị,chuẩn_hóa, phủ_tối_thiểu, phủ_không_dư,	CSDL nâng cao
Luật; lập_luận; logic_mờ, mạng_neuron, thuật_giải_di_truyền, lập_luận_lùi, lập_luận_tiến, cơ_sở_tri_thức, suy_diễn, logic_mờ, lập_luận_xấp_xỉ,	Công nghệ Tri thức
Luật_kết_hợp, tập_phổ biến, phân_lớp, gom_cụm, ngưỡng_minsupp, ngưỡng_minconf; dữ_liệu_lớn, nhiều_chiều, episode, mẫu_tuần_tự, cụm, nhà_kho_dữ_liệu, CSDL	Khai phá dữ liệu

Khi gặp văn bản cần phân lớp, ta tạo vetor đặc trưng cho văn bản. Qua vector này, chúng ta có thể xác định tập các từ xuất hiện đồng thời. Sau đó, chúng ta tính khỏang cách giữa tập các từ trong vector đặc trưng văn bản với tập từ đặc trưng cho cụm bằng công thức tìn khỏang cách giữa hai tập hợp bằng công thức

$$\mathbf{1 - (| X \cap Y | / | X \cup Y |)}.$$

Với X là tập hợp các từ đặc trưng cho văn bản và Y là tập hợp các từ có trong tập từ đặc trưng cho cụm. Cụm ngữ nghĩa có khỏang cách gần nhất sẽ được dùng làm nhãn ngữ nghĩa cho từ. Sau khi xác định được nghĩa, chúng tôi chọn nhánh đi lên trong đồ thị Wordnet để xác định mức độ gần nghĩa.

5. XÂY DỰNG BỘ PHÂN LỚP VĂN BẢN

Sau khi đã có tập luật phân lớp, mỗi thông điệp sẽ được rút trích và tạo vector đặc trưng. Qui trình phân lớp được thực hiện thông qua thuật toán 2 [2],[8].

1.1.1.1.1.1.1 Thuật toán 2 – Tạo bộ phân loại văn bản

1. Ứng với mỗi văn bản mới, dựa trên tập các cụm danh từ phổ biến để tạo một vector nhị phân đại diện cho thông điệp
2. Các luật phân lớp lần lượt được biến đổi thành các vector
3. Điều chỉnh các thành phần của vector đặc trưng văn bản và vector đặc trưng lớp dựa trên việc duyệt đồ thi đồng hiện để tìm nghĩa, sử dụng Wordnet để tìm từ gần nghĩa, đồng nghĩa
4. Tính độ đo tương tự dựa trên hệ số Cosine giữa vector văn bản và vector đặc trưng lớp theo công thức

$$\frac{\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i^2)^{1/2} (\sum_{i=1}^{n} y_i^2)^{1/2}}$$

5. Nếu tồn tại duy nhất một nhóm có mức độ tương tự lớn nhất ứng với luật tương ứng thì thông điệp sẽ được phân vào nhóm đó.

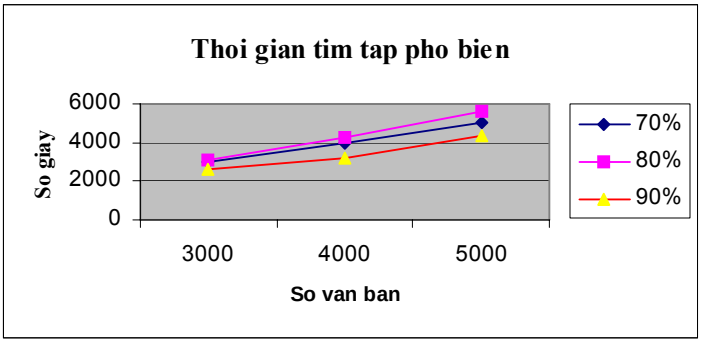
6.THỬ NGHIỆM

Chúng tôi tiến hành phân lớp các tóm tắt bài báo khoa học tiếng Việt trong lĩnh vực CNTT . Chiều dài trung bình cho mỗi tóm tắt bài báo khoa học khoảng 300 từ. Chúng tôi sử dụng khoảng 2/3 số lượng mẫu cho việc huấn luyện và phần còn lại để để kiểm tra độ chính xác của phân lớp. Ứng dụng thuật toán tìm dãy từ phổ biến, chúng tôi thu được khoảng. 1,200 cụm danh từ phổ biến với nguỡng là 2. Một số cụm danh từ tiêu biểu được liệt kê như sau: "tổng hợp, phân rã, ràng buộc, bảo toàn, toàn vẹn, dạng chuẩn, suy diễn lùi, suy diễn tiến, lập luận xấp xỉ, cơ chế giải thích, logic mờ, mạng neuron, phân nhóm , gom cụm, thuật toán học, toàn vẹn, dạng chuẩn, dạng chuẩn 1, phụ thuộc hàm, kết tự nhiên, phủ tối thiểu, hệ cơ số, cơ sở dữ liệu,tiếp cận…". Kết quả thử nghiệm tiến hành trên máy PC Pentium 4, 256MB RAM được trình bày trong bảng 2.

Bảng 2: Bảng so sánh thời gian xử lý theo các độ phổ biến khác nhau

Số văn bản huấn luyện	**3000**		**4000**		**5000**	
Độ phổ biến	Số luật kết hợp	Thời gian (giây)	Số luật kết hợp	Thời gian (giây)	Số luật kết hợp	Thời gian (giây)
70%	512	3600	846	5800	1243	7400
80%	498	3100	732	4300	1053	5600
90%	402	2600	698	3200	**987**	4356

Biểu đồ phân tích giữa thời gian xử lý, số lượng văn bản và độ phổ biến được trình bày trong hình 3.



Hình 3. Biểu đồ phân tích thời gian xử lý theo số văn bản và ngưỡng minsupp

Độ chính xác của kết quả phân lớp được trình bày trong bảng 3.

Bảng 3: Độ chính xác của kết quả phân lớp

Số văn bản huấn luyện	Số văn bản kiểm tra	Độ phổ biến	70%	80%	90%
2000	600	Độ chính xác phân lớp	43%	46%	54%
3000	1000	Độ chính xác phân lớp	49%	53%	62%
4000	1200	Độ chính xác phân lớp	54%	61%	**81%**
5000	1600	Độ chính xác phân lớp	62%	**75%**	**86%**

Một số luật phân lớp được tạo từ luật kết hợp tiêu biểu:
{phụ_thuộc_hàm, khóa, dạng_chuẩn}→ {Nhóm_cơ_sở_dữ_liệu}
{phụ_thuộc_đa_trị, lược_đồ_quan_hệ}→ {Nhóm_cơ_sở_dữ_liệu}
{ khóa, bao_đóng, phủ_tối_tiểu }→ {Nhóm_cơ_sở_dữ_liệu}
{dạng_chuẩn,phân_rã,bảo_toàn}→ {Nhóm_cơ_sở_dữ_liệu}
{mạng_neuron, thuật_toán_GA, lớp}→ {Nhóm_cơ_sở_tri_thức}
{suy_diễn_lùi, luật}→ {Nhóm_cơ_sở_tri_thức}

7.KẾT LUẬN

Bài báo trình bày các kết quả nghiên cứu về việc ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét ngữ nghĩa của từ. Thuật toán tìm tập phổ biến được cải biên cho phép tìm dãy từ phổ biến trong văn bản, Sau có thuật toán tách từ và gán nhãn từ loại được sử dụng để tìm các cụm danh từ. Từ điển Wordnet và từ đồng hiện được sử dụng để phát hiện nghĩa và điều chỉnh thành phần của vector đặc trưng. Thuật toán tìm luật kết hợp được cải biên nhằm cho phép tìm luật phân lớp văn bản. Hệ thống đề xuất được tiến hành thử nghiệm qua tập các tóm tắt bài báo khoa học.

RESEARCH ON APPLICATION OF FREQUENT SETS AND ASSOCIATION RULES TO SEMANTIC VIETNAMESE DOCUMENT CLASSIFICATION

Do Phuc
Center of Information Technology Development, VNU-HCM

ABSTRACT: *Today, the volume of electronic documents in the Internet is really huge. Therefore, the issue of developing the classification algorithms which can work effectively with large data set is a research direction of text mining. In this paper, we would like to present some results of the application of frequent sets and association rules to the document classification problem. We have applied these algorithms in i) Using the frequent sets and association rules for generating the document feature vectors, and ii) Using the association rules for classifying the documents. In the problem (i) the frequent set discovery algorithm has been improved to find the frequent terms in the corpus and document. After that, the natural language processing algorithms has been used for POS tagging and discovering the noun phrases. Besides, the association rules have been used to build the co-occurrence term graph in a particular context supporting to determine the word sense and the adjustment of the similar meaning components of document feature vector. In problem (ii), the association rules are used to generate the classification rules. The proposed system was tested with the data set of abstracts of papers in IT field.*

TÀI LIỆU THAM KHẢO

[1]. Beate Dorow (2003), *Discovering Corpus Specific Word Senses*, EACL, Hungary
[2]. Ciya Liao, Shamin Alpha, Paul Dixon(2000), *Feature Preparation in Text Categorization. Oracle Cooperation*
[3]. D. Phuc, H. Kiem (2000), *Discovering the binary and fuzzy association rules from database*, In Proc of AFSS2000 intl. Conf on Fuzzy Set and Application, Tsukuba, Japan, pp 981-986
[4]. Diệp Quang Ban, Hoàng Văn Thung (2000), *Ngữ pháp tiếng Việt,* NXB Giáo dục.

[5]. Dinh Dien, Nguyen Van Toan, Hoang Kiem (2001), *Vietnamese Word Segmentation*, In Proc of the NLPRS'01 conf,Tokyo, Japan, 2001.

[6]. Ellen M. Voorhees (1999), *Using WordNet for Text Retrieval*, WordNet, MIT Press, England, pp 285-303

[7]. G. Miller (1999), *Nouns in Wordnet*, Wordnet MIT Press, England

[8]. Nguyen Thi Minh Huyen. Laurent Romary (2003): A case study in POS Tagging of Vietnamese texts, http://www.vietlex.com\research

[9]. R. Florin, G. Ngai (2001), *Multidimensional Transformation based Learning*, Computational Language Learning

[10]. R. Agrawal & R. Srikant (1994), *Fast algorithm for mining association rules*, In proc of VLDB'94 intl conf, Santiego, Chile

[11]. Sam Scott, Stam Matwin (2000), *Feature engineering for text classification*, University of Ottawa, Canada, 2000

[12]. Yoshiki Niwa, Yoshiki Nita (1998), *Co-occurrence vectors from corpora vs. distance vector from Dictionary*, Advanced Research Laboratory, Japan,