GOM CỤM ĐỒ THỊ VÀ ỨNG DỤNG VÀO VIỆC RÚT TRÍCH NỘI DUNG CHÍNH CỦA KHỐI THÔNG ĐIỆP TRÊN DIỄN ĐÀN THẢO LUẬN

Đỗ Phúc, Mai Xuân Hùng, Nguyễn Thị Kim Phụng
Trường Đại học Công nghệ Thông tin, ĐHQG.HCM
(Bài nhận ngày 26 tháng 08 năm 2007, hoàn chỉnh sửa chữa ngày 03 tháng 03 năm 2008)

TÓM TẮT: *Bài báo trình bày kết quả nghiên cứu xây dựng một hệ thống gom cụm các thông điệp trên diễn đàn thảo luận nhằm hỗ trợ trích lược nội dung chính trong khối thông điệp. Các thông điệp trên diễn đàn là một dạng văn bản. Để gom cụm thông điệp, cần tìm kiếm mô hình đặc trưng văn bản. Các tiếp cận trước đây đã sử dụng mô hình tập hợp từ hay vector từ để đặc trưng văn bản. Các mô hình này đã bỏ sót các thông tin quan trọng trong văn bản như vị trí của từ trong văn bản, quan hệ ngữ nghĩa giữa các từ, các liên kết trên các văn bản web... Gần đây đã có các công trình nghiên cứu sử dụng đồ thị để đặc trưng văn bản. Sau khi biểu diễn các thông điệp bằng đồ thị, chúng tôi đã chọn giải pháp gom cụm đồ thị bằng mạng Kohonen vì mạng Kohonen có thể gom cụm dữ liệu mà không cần chỉ định trước số cụm. Ngoài ra mạng Kohonen có khả năng biểu diễn trực quan khối văn bản trên màn hình máy tính thông qua lớp ra Kohonen 2D. Chúng tôi đã tiến hành nghiên cứu cách tính khoảng cách giữa hai đồ thị dựa trên đồ thị con chung lớn nhất và cách cập nhật trọng số của mạng Kohonen dựa trên đồ thị có trọng bằng thuật giải di truyền sau đó tiến hành thử nghiệm và phân tích kết quả.*

Từ khóa: *Đồ thị trung bình có trọng, Gom cụm đồ thị, Khoảng cách đồ thị, Mạng Kohonen, Thuật giải di truyền*

1. GIỚI THIỆU

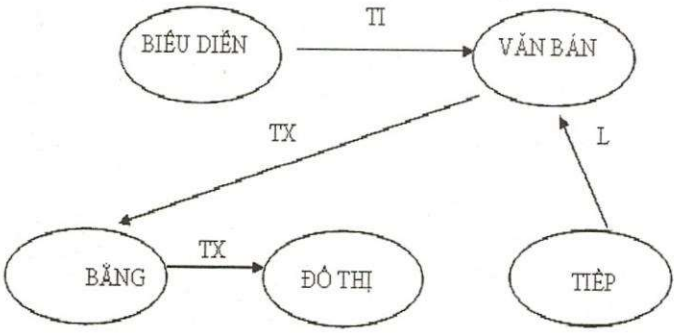
Trong các hệ thống trực tuyến, diễn đàn thảo luận là phương tiện hữu hiệu để trao đổi thảo luận. Khối lượng thông tin trao đổi trên diễn đàn thảo luận là rất lớn, hàng tháng có thể lên đến hàng ngàn thông điệp. Với số lượng này, người quản lý diễn đàn sẽ rất khó khăn khi cần nắm bắt các nội dung chính của thông tin trao đổi trên diễn đàn trong một giai đoạn [4]. Bài báo trình bày kết quả nghiên cứu xây dựng một hệ thống gom cụm các thông điệp trên diễn đàn thảo luận, hỗ trợ rút trích nội dung chính trong khối thông điệp. Các thông điệp trên diễn đàn là một dạng văn bản. Để gom cụm thông điệp, cần tìm kiếm mô hình đặc trưng cho văn bản. Các tiếp cận trước đây đã sử dụng mô hình tập hợp từ hay vector từ để đặc trưng cho văn bản. Các mô hình này đã bỏ sót các thông tin quan trọng trong văn bản như vị trí của từ trong văn bản, quan hệ ngữ nghĩa giữa các từ, các liên kết của các văn bản web... Gần đây đã có các công trình nghiên cứu sử dụng đồ thị để đặc trưng văn bản và đã chứng minh được tính vượt trội khi biểu diễn văn bản theo mô hình đồ thị [1],[3],[6]. Sau khi đặc trưng văn bản bằng đồ thị cần phát triển hệ thống gom cụm đồ thị.

Bài báo trình bày cách sử dụng mạng Kohonen để gom cụm các đồ thị đặc trưng văn bản và rút trích các ý chính từ khối văn bản hỗ trợ tạo trích lược thông tin chính trong khối văn bản. Mạng Kohonen do T. Kohonen phát triển vào những năm 1980 và đã được ứng dụng vào bài toán gom cụm phẳng. Mạng Kohonen có thể gom cụm dữ liệu mà không cần chỉ định trước số cụm, ngoài ra mạng Kohonen có khả năng biểu diễn trực quan khối văn bản trên màn hình máy tính thông qua lớp ra Kohonen 2D. Chúng tôi đã sử dụng mạng Kohonen để gom cụm đồ thị và tiến hành các nghiên cứu đề xuất cách tính khoảng cách giữa hai đồ thị dựa trên đồ thị con chung lớn nhất của chúng và cách cập nhật trọng số của đồ thị trọng trên các nút của

lớp ra Kohonen theo tiếp cận thuật giải di truyền. Bài báo được tổ chức như sau: 1) Giới thiệu 2) Biểu diễn văn bản bằng đồ thị 3) Mạng Kohonen 4) Gom cụm đồ thị bằng mạng Kohonen và rút trích ý chính 5) Thử nghiệm và bàn luận 6) Kết luận

2. BIỂU DIỄN VĂN BẢN BẰNG ĐỒ THỊ

Trong phần này, chúng tôi giới thiệu hai tiếp cận dùng đồ thị để đặc trưng cho văn bản [1],[3],[6]. Tiếp cận thứ nhất do Adam Schenker đề xuất [1]. Trong tiếp cận này, mỗi từ xuất hiện trong văn bản, trừ các phụ từ như "thì", "mà", "là", "bị"... là các từ chứa ít thông tin đều được biểu diễn bằng một đỉnh trong đồ thị biểu diễn văn bản. Nhãn của đỉnh là từ mà nó biểu diễn. Cho dù từ có xuất hiện nhiều lần trong văn bản, từ đó cũng được biểu diễn bằng một đỉnh duy nhất. Các cung của đồ thị được tạo như sau: nếu từ t2 đi liền sau từ t1 trong một đơn vị s của văn bản thì sẽ có một cung có hướng nối từ đỉnh biểu diễn cho từ t1 hướng đến đỉnh biểu diễn từ t2 và nhãn của cung này là s. Đơn vị s của văn bản có thể là tiêu đề, kết luận, đoạn văn, liên kết... Mỗi loại đơn vị sẽ được gán các tên nhãn khác nhau. Một ví dụ tiêu biểu cho đồ thị biểu diễn văn bản theo cách này được trình bày trong hình 1. Hình bầu dục chỉ các đỉnh và nhãn tương ứng, các cung được gán nhãn tiêu đề (TI), liên kết (L), văn bản (TX). Ví dụ văn bản có tiêu đề "BIỂU DIỄN", có liên kết đến văn bản với nhãn liên kết là "TIẾP" và nội dung văn bản là "VĂN BẢN BẰNG ĐỒ THỊ".



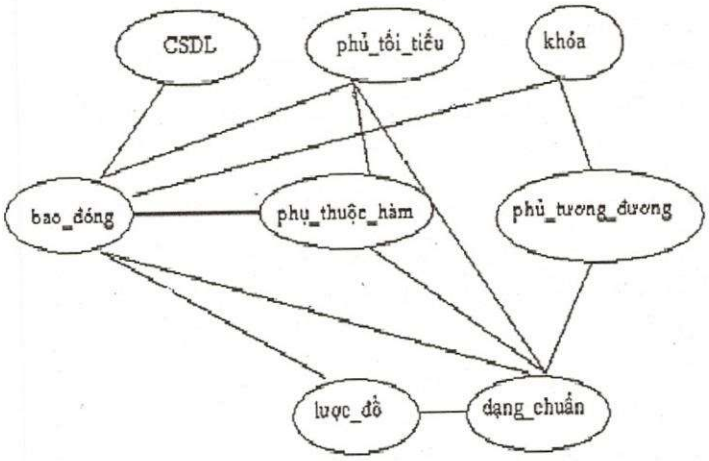
Hình 1. Đồ thị biểu diễn văn bản

Để nối hai từ có nghĩa tương tự nhau, chúng tôi dùng cung có nhãn là TS (text similarity). Ví dụ từ "túc cầu" và "bóng đá" là hai từ có nghĩa giống nhau. Trong tiếng Anh, từ điển Wordnet được sử dụng để đo sự tương đồng về nghĩa của hai từ. Đối với tiếng Việt, chúng tôi đã xây dựng từ điển từ đồng nghĩa và gần nghĩa cho các từ thông dụng và từ chuyên ngành CNTT. Một tiếp cận khác dùng đồ thị để biểu diễn văn bản được trình bày trong [6]. J.Tomita và cộng sự đã dùng đồ thị đồng hiện để biểu diễn văn bản. Đồ thị đồng hiện được tạo theo các bước sau:

Rút trích các từ phổ biến trong văn bản.

Tính các thành phần có ý nghĩa dựa trên tần suất xuất hiện đồng thời của hai từ trong một câu, đoạn văn bản ....Nếu tần suất xuất hiện đồng thời của hai từ lớn hơn một ngưỡng cho trước thì sẽ xuất hiện một cung nối hai từ này.

Một đồ thị đồng hiện tiêu biểu theo tiếp cận này được trình bày trong hình 2.



Hình 2. Đồ thị đồng hiện của văn bản

Chúng tôi sử dụng đồ thị đồng hiện để đặc trưng các thông điệp trên diễn đàn thảo luận. Bên cạnh đó, phần mềm tách từ tiếng Việt [4] đã được sử dụng để tách đúng các từ đơn, từ ghép của văn bản tiếng Việt nhằm tạo chính xác các đỉnh trong đồ thị đồng hiện biểu diễn văn bản tiếng Việt.

3. MẠNG KOHONEN

3.1. Gom cụm từ lớp ra Kohonen

Mỗi liên kết giữa đầu vào và đầu ra của mạng Kohonen tương ứng với một trọng số. Tổng đầu vào của mỗi nơron trong lớp Kohonen bằng tổng các trọng của các đầu vào nơron đó. Tiến trình huấn luyện mạng Kohonen sẽ điều chỉnh các trọng số dần dần theo mẫu học. Kết quả huấn luyện sẽ tạo trên lớp ra Kohonen các cụm dữ liệu ứng với nhóm các nút gần nhau trên lớp ra Kohonen. Các mẫu học sẽ thuộc về cụm có khoảng cách gần nhất từ nó đến nơron trong cụm. Theo tính chất của thuật giải huấn luyện trên mạng Kohonen, các cụm có vị trí gần nhau trên mạng Kohonen sẽ chứa các đối tượng có mức độ tương tự cao (tập văn bản có nội dung tương tự nhau) [7].

3.2. Thuật giải huấn luyện mạng Kohonen

Chức năng cơ bản của thuật giải huấn luyện mạng Kohonen truyền thống là gom các vector trọng của các nơron trên lớp ra Kohonen thành các cụm rời nhau [7]. Thuật giải huấn luyện mạng Kohonen truyền thống như sau:

Bước 1: Khởi tạo ngẫu nhiên các trọng số trên lớp ra Kohonen và gán Nc(t) là bán kính của vùng láng giềng. Khởi gán biến chu kỳ t=1

Bước 2: Đưa vào mạng một mẫu học hay vector nhập v(t) và chuẩn hóa v(t)

Tính khoảng cách Euclide từ vector nhập v(t) đến tất cả các vector trọng của tất cả các nơron trên lớp ra Kohonen và chọn nơron có khoảng cách Euclide dE nhỏ nhất từ vector học v(t) đến trọng ứng với nút đó.

dE (v,wic jc) = min (dE(vi,wij))

Trong đó i,j là các chỉ số hợp lệ được xác lập theo kích thuớc của lớp ra Kohonen.

Bước 3: Cập nhật các trọng số của các nút nằm trong vùng lân cận của nút chứa nơron chiến thắng (ic,jc) theo công thức:

$$\underline{w}_{ij}(t+1) = \underline{w}_{ij}(t) + \gamma\,(\underline{v} - \underline{w}_{ij}(t)) \quad (1)$$

Trong đó ic-Nc(t) ≤ i ≤ ic + Nc(t) và jc-Nc(t) ≤ j ≤ jc + Nc(t)

Hệ số γ có trị nằm trong đoạn [0,1], đây là hệ số học và sẽ giảm theo thời gian.

Bước 4. Cập nhật t = t + 1, đưa mẫu học kế tiếp vào mạng Kohonen và quay về bước 2 cho đến khi đạt được điều kiện hội tụ hay vượt quá số lần lặp qui định.

4.GOM CỤM ĐỒ THỊ BẰNG MẠNG KOHONEN VÀ RÚT TRÍCH Ý CHÍNH

Dữ liệu nhập vào mạng Kohonen là tập các đồ thị đặc trưng văn bản. Sau khi huấn luyện, các đồ thị nhập sẽ được gom vào các nút trên lớp ra của mạng Kohonen [7].

4.1.Khởi tạo đồ thị trọng

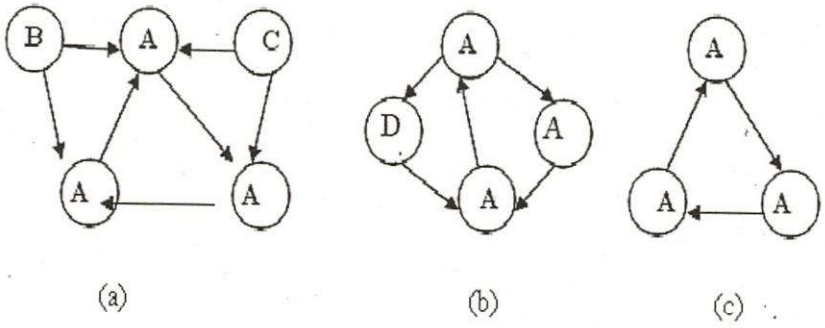
Một nơron trên lớp ra Kohonen là một đồ thị trọng được tạo ngẫu nhiên dựa trên tập dữ liệu nhập vào mạng Kohonen.

4.2. Khoảng cách giữa hai đồ thị

H Bunke [5] đã đề xuất công thức tính khoảng cách giữa hai đồ thị. Cho hai đồ thị G1 và G2, khoảng cách giữa hai đồ thị G1, G2, ký hiệu là d(G1,G2) được tính như sau:

$$d(G_1, G_2) = 1 - \frac{|\,mcs(G_1, G_2)\,|}{\max(|\,G_1, |\,G_2\,|)} \quad (2)$$

Trong đó mcs là đồ thị con chung lớn nhất và |.| là kích thước của đồ thị, trong nghiên cứu này kích thước là số đỉnh của đồ thị. Xét hai đồ thị ở hình 3.a và 3.b



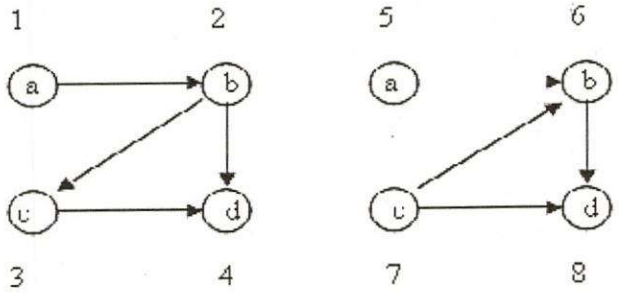
Hình 3. Tính khoảng cách giữa hai đồ thị

Hình (3.a) là đồ thị G1, hình (3.b) là đồ thị G2 , hình (3.c) là đồ thị con chung lớn nhất của hai đồ thị G1 và G2. Khoảng cách giữa hai đồ thị G1 và G2 là:

$$d(G_1, G_2) = 1 - \frac{|\,mcs(G_1, G_2)\,|}{\max(|\,G_1, |\,G_2\,|)} = 1 - \frac{3}{5} = 0{,}4$$

Theo W Henry S. [9], bài toán tính đồ thị con chung lớn nhất là bài toán thuộc lớp bài tóan NP đầy đủ, nhìn chung có ba cách giải bài toán. Một là phương pháp sử dụng clique tối đại, hai là chiến lược sử dụng kỹ thuật backtracking và không sử dụng clique tối đại, phương pháp thứ ba là các kỹ thuật khác. Chúng tôi đã sử dụng tiếp cận của W. Henry S. [9] để tạo đồ thị kết hợp (association graph) và dùng kỹ thuật tìm clique tối đại trên đồ thị kết hợp của hai

đồ thị G1 và G2 để tìm đồ thị con chung lớn nhất của chúng. Cho hai đồ thị G1=(V1,E1) và G2=(V2,E2) như trong hình 4.

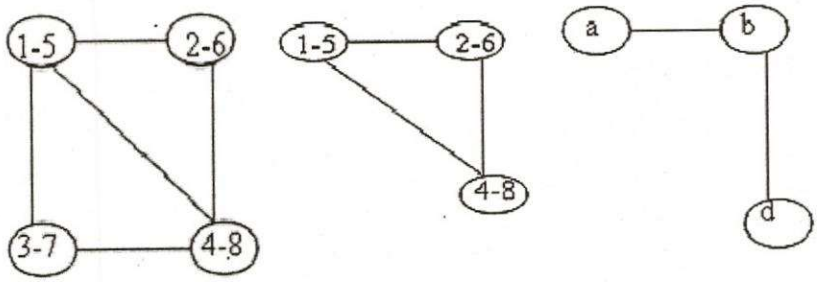


Hình 4. Hai đồ thị

Đồ thị kết hợp G=(V,E) của hai đồ thị G1=(V1,E1) và G2=(V2,E2) là đồ thị có $V \subseteq V1xV2$, cạnh (u1,v1) và (u2,v2) là kề nhau nếu:

(u1,u2) $\in$ E1 và (v1,v2) $\in$ E2

(u1,u2) $\notin$ E1 và (v1,v2) $\notin$ E2



(a) **(b)** **(c)**

Hình 5. Đồ thi kết hợp và đồ thì con chung lớn nhất

Đồ thị kết hợp của hai đồ thị G1, G2 là đồ thị nằm trong hình (5.a); clique tối đại nằm ở hình (5.b) và đồ thị con chung lớn nhất nằm ở hình (5.c).

4.3. Cập nhật đồ thị trọng trên các đỉnh của lớp ra Kohonen

Trong quá trình huấn luyện mạng Kohonen cần cập nhật trọng số của các nơron trong vùng lân cận nơron chiến thắng. Lưu ý khu dùng mạng Kohonen để gom cụm đồ thị, mỗi nơron trên lớp ra Kohonen là một đồ thị chứ không phải là vector trọng như trong mô hình truyền thống. Để cập nhật đồ thị trọng, H. Bunke [5],[8] sử dụng khái niệm đồ thị trung bình có trọng (weighted means graph) của cặp đồ thị để cập nhật đồ thị trọng. Cho hai đồ thị G1 và G2, đồ thị G là đồ thị trung bình có trọng của hai đồ thị G1 và G2 nếu với số α sao cho $0 \le \alpha \le d(G1,G2)$ ta có:

$$d(G1,G) = \alpha \quad (3)$$

và

$$d(G1,G2) = \alpha + d(G,G2) \quad (4)$$

Từ công thức (3) và (4), suy ra d(G1,G2)=d(G1,G)+d(G,G2)

Để cập nhật trọng số của mạng Kohonen ta sử dụng công thức (1). Có thể viết lại công thức (1) dưới dạng.

$$y_{new}-y_{old} = \gamma(x-y_{old}) \quad (5)$$

hay

$$x - y_{new} = (1-\gamma)(x-_{yold}) \quad (6)$$

Nếu thay thế G1=x, G=ynew, G2=yold, toán tử "-" bằng hàm tính khoảng cách "d(-,-)", thì công thức (5) và (6) sẽ trở thành công thức (7), (8) sau đây:

$$D(G,G2) = \gamma d(G1,G2) \quad (7)$$

Và

$$d(G1,G) = (1-\gamma)\, d(G1,G2) \quad (8)$$

Nếu đặt $\alpha = (1-\gamma)\, d(G1,G2)$, thì công thức (7),(8) sẽ trở thành công thức (3),(4). Nói cách khác G là đồ thị có trọng của hai đồ thị G1 và G2.

Từ công thức (7),(8) ta có công thức (9) như sau:

$$\frac{d(G_1,G)}{d(G,G_2)} = \frac{1-\gamma}{\gamma} \quad (9)$$

Thuật giải di truyền được sử dụng để tìm đồ thị trung bình có trọng của hai đồ thị G1 và G2 với các đặc điểm sau:

4.3.1.Khởi tạo quần thể

Đồ thị được biểu diễn bằng ma trận kề có đỉnh V=V1∪V2. Nhiễm sắc thể là các tập các đồ thị ứng viên của đồ thị trung bình có trọng (tập hợp ma trận). Tập các đồ thị ban đầu được khởi tạo ngẫu nhiên dựa trên hai đồ thị G1 và G2.

4.3.2.Phép toán lai ghép: lai ghép hai ma trận

Lai theo các đường chéo của ma trận.

Trước khi lai ghép:

1	1	0	0
1	1	1	1
0	1	1	0
0	1	0	1

Nhiễm sắc thể

Nhiễm sắc thể

1	0	1	1
0	1	0	1
1	0	0	1
1	1	1	1

Sau khi lai ghép:

1	0	0	0
0	1	1	1
0	1	1	0
0	1	0	1

Nhiễm sắc thể

Nhiễm sắc thể

1	1	1	1
1	1	0	1
1	0	0	1
1	1	1	1

4.3.3.Phép toán đột biến

Chọn một vị trị i,j ngẫu nhiên trong ma trận a (ma trận nhiễm sắc thể), ấn định giá trị ngẫu nhiên (thêm bớt cạnh trong đồ thị) và gán trị cho aij và aji

1	1	0	0
1	1	1	1
0	1	1	0
0	1	0	1

Nhiễm sắc thể

Nhiễm sắc thể sau đột biến

1	1	0	1
1	1	1	1
0	1	1	0
1	1	0	1

4.3.4.Hàm phù hợp

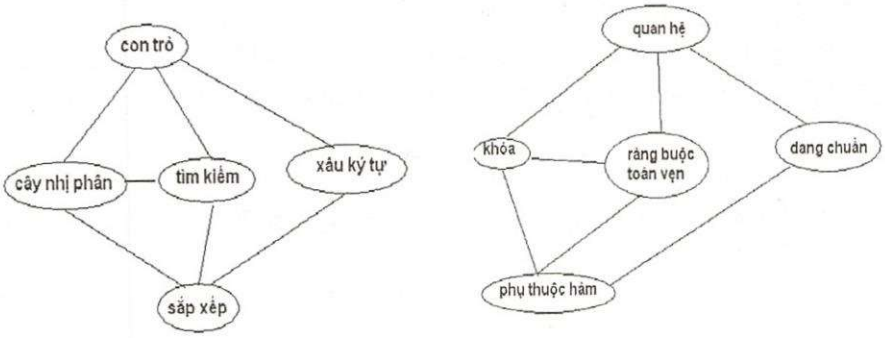
Cực tiểu hàm sau:

$$f(NST) = \frac{d(G_1, G)}{d(G, G_2)} - \frac{1-\gamma}{\gamma}$$

Trong đó γ là hằng số xác định tốc độ học và $0 \leq \gamma \leq 1$. Hàm này có trị càng nhỏ thì độ phù hợp của nhiễm sắc thể càng cao.

4.4. Trích rút nội dung chính trên đồ thị trọng của lớp ra Kohonen

Một số đồ thị có trọng tiêu biểu trên lớp ra Kohonen được trình bày ở hình 6. Có thể rút trích từ các đồ thị có ra các ý chính trong khối ngữ liệu văn bản, ví dụ các đoạn văn hoặc các câu có chứa các từ hoặc cụm từ đồng hiện như trong các đồ thị trọng sẽ mang ý chính trong tập văn bản của cụm.



Hình 6. Một vài đồ thị trọng trên các đỉnh của lớp ra Kohonen

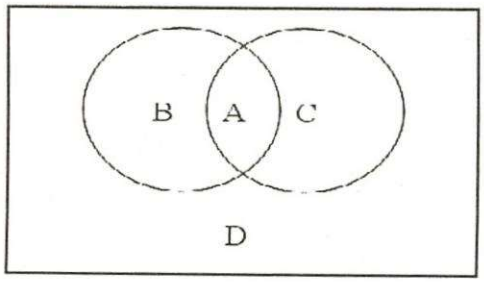
Một vài đồ thị trọng có chứa các nút không có nút kề. Các nút này sẽ bị loại khỏi đồ thị trọng.

5. THỬ NGHIỆM VÀ BÀN LUẬN

5.1. Thử nghiệm và bàn luận kết quả gom cụm

Các hệ số Precision, Recall và F-measure được sử dụng để đánh giá kết quả gom cụm. Chúng tôi so sánh kết quả gom cụm thông điệp theo giải pháp đề xuất và so sánh với kết quả gom cụm thông điệp bằng tay (do người làm). Kết qủa gom cụm thông điệp bằng tay dựa trên các chủ đề con trên diễn đàn thảo luận qua mạng, trong đó mỗi chủ đề là một cụm.

Xét tập có n thông điệp, sau khi gom cụm bằng tay ta có m cụm, và sau khi gom cụm bằng hệ thống phần mềm gom cụm văn bản có k cụm. Trong quá trình thử nghiệm ta có $m \leq k$. Để đánh giá kết quả của hệ thống, ta tiến hành xác định ba hệ số Precision, Recall và F-measure giữa hai cụm trong hai hệ thống.



Hình 7. Quan hệ giữa hai cụm

Gọi a=|A|, b=|B| và c=|C|.Trong hình 7, cụm mi do con người tạo ra là A∪B gồm có a+b văn bản, cụm ki do hệ thống gom gồm là A∪C có a + c văn bản. Hai cụm trên có phần chung là A và gồm a văn bản.

Hệ số Precision giữa hai cụm trên được ký hiệu là P (Precision) phản ánh độ chính xác của truy vấn và được tính bằng công thức:

$$P = \frac{a}{a+c} \tag{10}$$

Hệ số Precision cho biết tỉ lệ giữa số văn bản được gom cụm đúng. Nếu P=1 thì các văn bản trong cụm ki nằm trong các văn bản của cụm mi

Hệ số Recall giữa hai cụm mi và ki được ký hiệu là R (recall) và được tính bằng công thức (11). Nếu R =1 thì các văn bản trong cụm mi thuộc các văn bản nằm trong cụm ki

$$R = \frac{a}{a+b} \tag{11}$$

Có thể kết hợp hai hệ số Precision và Recall lại thành hệ số F-Measure. Hệ số F-Measure được tính bằng công thức:

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{12}$$

Giá trị α càng cao sẽ tác động mạnh đến hệ số Recall, ngược lại giá trị α thấp sẽ tác động mạnh lên hệ số Presicion Thông thường hệ số α trong công thức (12) được chọn là 0.5. Khi đó công thức (12) được viết lại:

$$F_{0.5} = \frac{2PR}{P+R} \tag{13}$$

Brew C. [2] đề nghị cách đánh giá như sau: Tương ứng với một cụm trong kết quả gom cụm của hệ thống ta đi tính giá trị của độ đo F-measure với tất cả các cụm được gom bằng tay. Chọn ra giá trị của F-measure cao nhất và loại cụm này ra. Tiếp tục công việc trên, cho các cụm còn lại. Tổng các giá trị F-measure càng cao thì hệ thống gom cụm càng chính xác.

Tập kết quả thử nghiệm gom cụm có 500 thông điệp thuộc về 5 chủ đề khác nhau, mỗi chủ đề có 100 thông điệp. Kích thước lớp ra Kohonen là 8x8; Chu kỳ lặp max là 5000; Chu kỳ cập nhật bán kính vùng lân cận là 50. Đồ thị đồng hiện được sử dụng để đặc trưng thông điệp.

5.1.1.Phương pháp gom cụm vector

Kết quả: Phương pháp gom cụm bằng tay: cho 5 cụm mỗi cụm có 100 thông điệp . Phương pháp gom cụm văn bản trong đó vector được sử dụng để biểu diễn văn bản. Số cụm thu được

sau khi gom cụm là 8 cụm. Sử dụng các công thức (10),(11),(13) để tính các hệ số Precision, Recall, F-measure. Ta có kết quả tính F-measure như trong bảng 1:

Bảng 1. Kết quả tính F-measure giữa gom cụm bằng tay và gom cụm vector

Người / Máy	Cụm 1	Cụm 2	Cụm 3	Cụm 4	Cụm 5
Cụm 1	0.32	0.12	0.12	0.35	0.13
Cụm 2	0.12	0.00	0.35	0.12	0.43
Cụm 3	0.00	0.14	0.23	0.23	0.23
Cụm 4	0.12	0.23	0.35	0.35	0.31
Cụm 5	0.00	0.24	0.12	0.12	0.23
Cụm 6	0.00	0.34	0.54	0.43	0.12
Cụm 7	0.13	0.23	0.12	0.12	0.35
Cụm 8	0.00	0.12	0.00	0.12	0.23
Max	0.32	0.34	0.54	0.43	0.43

Tổng Max cho gom cụm vector = 0.32+0.34+0.54+0.43+0.43= 2.06

5.1.2.Phương pháp gom cụm đồ thị được sử dụng để biểu diễn văn bản

Số cụm văn bản thu được là 6 cụm Tính giá trị của các hệ số giữa kết quả của phương pháp gom cụm đồ thị và phương pháp gom cụm bằng tay. Sử dụng các công thức (10),(11),(13) để tính các hệ số Precision, Recall, F-measure. Ta có kết quả tính F-measure như trong bảng 2:

Bảng 2. Kết quả tính F-measure giữa gom cụm bằng tay và gom cụm đồ thị

Người / Máy	Cụm 1	Cụm 2	Cụm 3	Cụm 4	Cụm 5
Cụm 1	0.35	0.12	0.22	0.35	0.34
Cụm 2	0.21	0.23	0.23	0.12	0.54
Cụm 3	0.54	0.12	0.12	0.42	0.23
Cụm 4	0.12	0.23	0.68	0.56	0.23
Cụm 5	0.35	0.32	0.23	0.34	0.35
Cụm 6	0.12	0.12	0.23	0.12	0.12
Max	0.54	0.32	0.68	0.56	0.54

Tổng Max cho gom cụm đồ thị = 0.54+0.32+0.68+0.56+0.54=2.64

Chúng tôi cũng đã thử nghiệm với 5 tập mẫu ngẫu nhiên khác mỗi tập có 500 văn bản. Kết quả được nêu trong bảng 3.

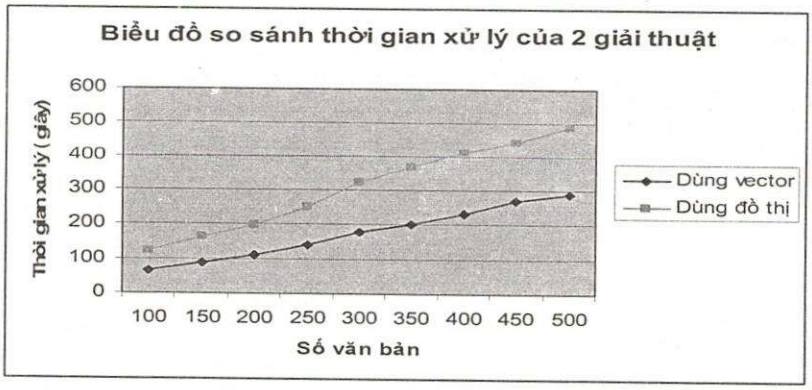
Bảng 3. Kết quả thử nghiệm với các tập mẫu ngẫu nhiên

Tập mẫu ngẫu nhiên	Tổng Max cho gom cụm vector	Tổng Max cho gom cụm đồ thị
Tập mẫu 1	2,06	2,64
Tập mẫu 2	2,21	2,98
Tập mẫu 3	3,01	3,34
Tập mẫu 4	2,34	3,12
Tập mẫu 5	2,23	2,56

Nhận xét: Qua thử nghiệm và tính tổng giá trị lớn nhất của hệ số F-Measure cho nhiều tập mẫu khác nhau, chúng tôi nhận thấy tổng giá trị lớn nhất của hệ số F-Measure của hệ thống gom cụm văn bản được biểu diễn bằng đồ thị lớn hơn nhiều so với hệ thống gom cụm văn bản được biểu diễn bằng vector. Điều này khuyến khích chúng tôi tiếp tục phát triển phương pháp biểu diễn văn bản đồ thị nhằm thay thế các phương pháp biểu diễn bằng gói từ và vector từ với mục đích nâng cao chất lượng gom cụm.

5.2. Bàn luận về độ phức tạp tính toán của giải pháp đề xuất

So với giải thuật sử dụng vector với mạng Kohonen truyền thống, giải thuật đề xuất sử dụng đồ thị và vẫn giữ nguyên các phần chính yếu của giải thuật học trên mạng Kohonen. Giải thuật học mới chỉ thay đổi ở cách thức tính khoảng cách giữa mẫu nhập và đồ thị trọng và cách thức cập nhật đồ thị có trọng. Đối với việc sử dụng vector trên mạng Kohonen theo tiếp cận cũ, các công đoạn này bao gồm việc tính khoảng cách Euclide và việc cập nhật trọng qua việc điều chỉnh thành phần của vector trọng trong vùng lân cận nơron chiến thắng. Theo tiếp cận sử dụng đồ thị trên mạng Kohonen, các công việc này được thay thế bằng việc tính khoảng cách của hai đồ thị và công việc tính đồ thị trung bình có trọng (cập nhật đồ thị trọng). Các đồ thị được biểu diễn bằng ma trận kề. Việc tính khoảng cách giữa hai đồ thị và tính đồ thị trung bình có trọng theo tiếp cận thuật giải di truyền là các công việc đòi hỏi thời gian xử lý. Hình 8 là biểu đồ so sánh thời gian xử lý trung bình của hai phương pháp với tập thử nghiệm chứa 100, 150, 200, 250, 300, 350,400, 450, 500 thông điệp. Thời gian được tính bằng phút là thời gian xử lý của giải thuật học. Kết quả thu được từ việc chạy chương trình trên máy PC Pentium 4, 3GB với bộ nhớ 500 Mbyte RAM.



Hình 8. So sánh thời gian xử lý của giải thuật dùng vector và giải thuật dùng đồ thị

Chúng tôi cũng đã tiến hành lựa chọn 5 tập mẫu khác nhau mỗi tập mẫu có chừng 500 văn bản tiến hành thử nghiệm như trên. Kết quả thử nghiệm cả hai cách bằng tay và bằng thuật toán đề xuất với thời gian xử lý trung bình như sau:

Bảng 4. So sánh thời gian xử lý với các tập mẫu ngẫu nhiên

Tập mẫu ngẫu nhiên	Chênh lệch trung bình về thời gian xử lý (lần)
Tập mẫu 1	1,62
Tập mẫu 2	1,78
Tập mẫu 3	1,65
Tập mẫu 4	1,92
Tập mẫu 5	1,56

Các số liệu trên biểu đồ cho thấy giải thuật đề xuất có thời gian xử lý trung bình gấp gần 1,7 lần so với giải thuật Kohonen sử dụng vector. Tuy vậy, như đã nhận xét ở phần trên, giải thuật đề xuất cho kết qủa gom cụm có độ chính xác vượt trội giải thuật Kohonen sử dụng vector và mở ra hướng cải tiến chất lượng kết quả gom cụm trên các lớp ra của mạng Kohenen bằng tiếp cận đồ thị.

6. KẾT LUẬN

Bài báo trình bày kết quả nghiên cứu xây dựng một hệ thống gom cụm thông điệp trên diễn đàn thảo luận nhằm hỗ trợ trích lược nội dung chính trong khối thông điệp. Các thông điệp là một dạng văn bản. Mô hình đồ thị được sử dụng để biểu diễn văn bản nhằm thể hiện các quan hệ cấu trúc, ngữ nghĩa giữa các từ, khái niệm,... Phần mềm tách từ tiếng Việt đã được sử dụng để tách đúng các từ đơn, từ ghép trong văn bản tiếng Việt. Mạng Kohonen đã được sử dụng để gom cụm các đồ thị biểu diễn văn bản. Khoảng cách giữa hai đồ thị được tính dựa vào đồ thị con chung lớn nhất giữa hai đồ thị. Tiếp cận đồ thị kết hợp và clique tối đại được sử dụng để tìm đồ thị con chung lớn nhất giữa hai đồ thị qua đồ thị kết hợp. Để cập nhật đồ thị trọng trên các nút của lớp ra Kohonen, chúng tôi đã sử dụng tiếp cận đồ thị trung bình có trọng được tính bằng thuật giải di truyền.Giải thuật đề xuất đã được thử nghiệm để gom cụm thông điệp trên diễn đàn thảo luận và phân tích kết quả về mặt chất lượng và thời gian xử lý.Chúng tôi đang tiếp tục nghiên cứu cải tiến việc tính toán khoảng cách đồ thị và đồ thị trọng để giảm độ phức tạp tính toán.

GRAPH CLUSTERING AND APPLICATION TO THE EXTRACTION OF MAIN IDEAS IN COLLECTION OF ONLINE FORUM MESSAGES

Do Phuc, Mai Xuan Hung, Nguyen Thi Kim Phung
Univesity of Information Technology, VNU-HCM

ABSTRACT: *This paper presents the results of building a graph clustering system for grouping the similar messages of forum of e-learning system and extracting the main ideas in the collection of messages. Message is a kind of text. To cluster the messages, we need a model for representing the documents. The traditional approaches used the models of bag of words or vector model for representing the documents. These models discard the important structural information of document such as word position, the semantic relation of words in document, the links of web pages... Recently, there are several works using the graph for representing the documents. After representing the documents by graph, Kohonen neural network was used for grouping the graphs. One of the advantages of Kohonen neural network is to cluster the data without specifying the number of clusters. Besides, Kohonen neural output layer is a document map which can put on the computer display for easily accessing the similar documents. The graph distance based on the maximum common sub-graph and the updated operation of Kohonen neural network based on the weighted means of two graphs was chosen. Our proposed solution with the messages in our online forum was tested and discuss the results were analysed.*

Keywords: *Weighted means of graphs, Graph clustering, Graph Distance, Kohonen neural network, Genetic algorithm*

TÀI LIỆU THAM KHẢO

[1]. Adam Schenker et al. *Classification of Web documents using a graph model*, In Proc of the 7 th int'l conf of document analysis and Recognition (ICDAR'2003), (2003)

[2]. Brew C, Schulte im Walde. *Spectral Clustering for German Verbs*, In Proc of the Conf in Natural Language Proocessing, Philadenphia, PA, pp 117-124, (2002).

[3]. Do Phuc. *Using graph mining, frequent sub-graph for document classification*, In Proc of the int'l IEEE RIVF'06 conf, pp 173-176, (2006).

[4]. Do Phuc, Nguyen Thi Kim Phung. *Using the Naïve Bayes model and natural language rocessing for classifying messages on online forum*, In Proc of the int'l IEEE RIVF'07 conf, pp247-252, (2007).

[5]. H Bunke, Kim Shearer. *Graph distance metric based on the maximal common subgraph,* Pattern Recognition letter 19, pp 225-229, (1998).

[6]. J. Tomita et al. *Graph based text database for Knowledge discovery*-In Proc of int'l conference, WWW 2004, (2004).

[7]. Kaski, S., Honkela, T., Lagus, K., and Kohonen. T.*WEBSOM--self-organizing maps of document collections*. Neuro *computing*, volume 21, (1998).

[8]. Simon Günter, Horst Bunke. *Self-organizing map for clustering in the graph domain,* Pattern Recognition Letters, v.23 n.4, pp.405-417, (2002).

[9]. W Henry Suters.: *A new approach and faster exact methods for the maximum common subgraph*, web: http://www.cs.utk.edu/~library/TechReports/2005/ut-cs-05-568.pdf.,(2002)