TINH CHỈNH ĐẶC TRƯNG TỪ TRONG GOM NHÓM TẬP CÂU HỎI TIẾNG VIỆT

Nguyễn Xuân Hậu - Ngô Thị Khánh Tường*

Tóm tắt

Nghiên cứu này trình bày về các kỹ thuật tinh chỉnh đặc trưng từ trong tập dữ liệu hỏi đáp Tiếng Việt phục vụ cho việc gom nhóm dữ liệu. Dựa vào kết quả đánh giá gom nhóm thử nghiệm các kỹ thuật tinh chỉnh đặc trưng trên tập dữ liệu thử nghiệm, từ đó đưa ra các đặc tính phù hợp của đặc trưng dùng cho việc gom nhóm tập dữ liệu hỏi đáp Tiếng Việt.

Từ khóa: *tinh chỉnh đặc trưng, rút trích đặc trưng, hệ thống hỏi đáp, gom nhóm*

1. Giới thiệu

Việc tinh chỉnh đặc trưng nhằm làm giảm đáng kể số chiều dữ liệu không những làm nhanh hơn khi thực hiện gom nhóm mà còn làm tăng độ chính xác khi gom nhóm dữ liệu. Chúng tôi sẽ tiến hành đánh giá các kỹ thuật tinh chỉnh đặc trưng trong các tập câu hỏi Tiếng Việt như lọc các hư từ, các từ xuất hiện ít, các từ xuất hiện nhiều, lọc giữ lại danh từ, cụm danh từ và động từ, phép biến đổi dữ liệu SVD [9] và đánh giá ảnh hưởng của chúng tới quá trình gom nhóm tập dữ liệu câu hỏi Tiếng Việt. Để đánh giá sự ảnh hưởng việc tinh chỉnh đặc trưng trong gom nhóm dữ liệu. Chúng tôi sử dụng các độ đo phản ánh chất lượng nhóm dữ liệu. Ngoài ra, thuật toán gom nhóm K-means và sử dụng độ đo Euclidean để tính khoảng cách các phần tử xuyên suốt trong quá trình đánh giá.

2. Tập dữ liệu

Để đánh giá các kỹ thuật tinh chỉnh đặc trưng trên tập dữ liệu hỏi đáp Tiếng Việt, chúng tôi tiến hành xây dựng bộ dữ liệu thử nghiệm như sau:

Tập dữ liệu thử nghiệm được thu thập từ website Đối thoại doanh nghiệp thành phố Hồ Chí Minh (hiện tại website này có hơn 12000 câu hỏi) [3]. Trong đó, có 2 tập con:

Tập thứ nhất: TH – tập hợp những cặp hỏi đáp gồm 4 chủ đề khác nhau bao gồm: *"các sắc thuế trong nội địa, kế hoạch & đầu tư, bảo hiểm xã hội và hải quan"*.

Tập thứ hai: CST- tập hợp những cặp hỏi đáp trên cùng một chủ đề "*các sắc thuế trong nội địa*".

Sau khi thu thập, chúng tôi thực hiện các bước tiền xử lí nhằm chỉnh sửa lại dữ liệu theo đúng ý nghĩa vốn có của nó như sau:

- Bổ sung dấu và sửa các lỗi chính tả, từ viết tắt, bổ sung dấu hỏi...

* ThS, Trường Cao đẳng Công nghiệp Tuy Hòa

- Loại bỏ phần tiêu đề của câu hỏi.

- Sau đó thực hiện xử lí để đưa tập câu hỏi (tập dữ liệu) về dạng có thể xử lí được: bằng cách, chúng tôi thực hiện lần lượt các bước sau để trích từ khóa: tách câu, tách từ, phân loại từ loại, cập nhật các hư từ và các từ xuất hiện nhiều nhưng không có ý nghĩa trong hệ thống; Xây dựng vector đặc trưng từ cho từng câu hỏi trong tập câu hỏi .

3. Đặc trưng tập dữ liệu thử nghiệm

Tập dữ liệu thứ nhất (TH) thu thập các câu hỏi trên bốn chủ đề khác nhau như: "*các sắc thuế, kế hoạch và đầu tư, bảo hiểm xã hội, hải quan*". Việc đánh giá gom nhóm trên tập dữ liệu này ta có thể dựa trên hai loại độ đo: độ đo dựa vào thông tin nội tại bên trong của dữ liệu hoặc các độ đo dựa vào thông tin bên ngoài (như là dựa trên nhãn các phần tử). Đối với tập dữ liệu này thì thông tin bên ngoài là chủ đề mà phần tử đó thuộc. Chúng tôi xem đó như là một nhãn được gán từ trước cho các đối tượng dữ liệu (chủ đề là nhãn của tập dữ liệu).

Tập dữ liệu thứ hai (CST) các câu hỏi được thu thập trên cùng một chủ đề là *"các sắc thuế "*. Do các câu hỏi trên cùng một chủ đề nên việc đánh giá gom nhóm trên tập dữ liệu này chỉ sử dụng độ đo dựa vào thông tin nội tại của phần tử (không dựa vào sự gán nhãn cho trước). Các độ đo này thể hiện sự gắn kết của các phần tử trong nhóm và sự khác biệt với các phần tử thuộc nhóm khác.

Mục đích của việc tạo ra hai tập dữ liệu trên là để chúng ta có một cái nhìn tổng quát và toàn diện hơn khi tiến hành đánh giá so sánh sự phù hợp của kỹ thuật tinh chỉnh đặc trưng dựa trên các độ đo khác nhau. Đặc biệt, các độ đo dựa trên nhãn đã có, cho ta một cái nhìn khách quan hơn khi chỉ tiến hành đánh giá, vì đánh giá sự phù hợp của phương pháp gom nhóm dựa trên thông tin nhãn cho trước cũng có nghĩa là nếu các phần tử trong cùng một nhóm thuộc về một nhãn thì phương pháp gom nhóm được đánh giá tốt. Còn nếu các phần tử trong cùng một nhóm thuộc về nhiều nhãn khác nhau thì phương pháp gom nhóm không tốt.

Thống kê của 2 tập dữ liệu :

Chủ đề	Số lượng
Các sắc thuế	486
Kế hoạch và đầu tư	219
Bảo hiểm xã hội	154
Hải quan	146

Bảng 1 : Tập dữ liệu TH

Chủ đề	Số lượng
Các sắc thuế	1005

Bảng 2 : Tập dữ liệu CST

	TH	CST
N	1005	1005
(min n_d, max n_d)	(10, 501)	(11, 501)
Dim	3912	3502
K	4	1

Bảng 3: Thống kế 2 tập dữ liệu

***Trong đó* :**

N: Số lượng phần tử (câu hỏi) trong kho dữ liệu.

(min n_d :max n_d): Số lượng từ (term) nhỏ nhất và lớn nhất trong một câu hỏi (phần tử) trong tập dữ liệu.

Dim: Số chiều của tập dữ liệu.

K : Số nhãn của tập dữ liệu.

4. Đánh giá các kỹ thuật tinh chỉnh đặc trưng

a. Lọc hư từ (stopword), các từ xuất hiện nhiều nhưng không có nghĩa và các từ loại danh từ, cụm danh từ, động từ (NV)

Trong hầu hết các công trình nghiên cứu về xử lí ngôn ngữ luôn đề nghị xử lí loại bỏ hư từ và loại bỏ các từ không có nghĩa, vì nó không những không có nghĩa mà còn làm nhiễu hơn trong các quá trình xử lí ngôn ngữ. Tập hư từ chúng tôi sử dụng từ [1] và có thêm những từ xuất hiện nhiều nhưng không có ý nghĩa trong Hệ thống hỏi đáp đối thoại doanh nghiệp của TP HCM. Gần đây, cũng có vài nghiên cứu đề nghị lọc lấy những từ loại là danh từ, nhóm danh từ và động từ [4] trong các xử lí gom nhóm, phân loại dữ liệu. Trong bước xử lí lọc lấy danh từ và động từ chúng tôi sử dụng công cụ phân loại từ loại JVnTagger-SP8.3 [1] là một phần của đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP)

Bộ lọc	Dim(%)	NMI	Purity	DB	Avg	Max : Min
Không lọc	100	0.221	0.606	-2.312	-151.900	279 :1
Hư từ	94.4	0.337	0.719	-2.273	-100.783	288 :1
NV	89.5	0.247	0.656	-2.180	-120.675	268 :1
Hư từ + NV	**86.1**	**0.367**	**0.753**	**-2.305**	**-88.401**	**292 :1**

Bảng 4: Lọc hư từ và NV trên kho dữ liệu TH

Bộ lọc	Dim(%)	Accuracy(%)	DB	Avg	Max : Min
Không lọc	100	22.89	-2.675	-157.789	118 :1
Hư từ	93.6	29.23	-2.538	-101.601	207 :1
NV	89.3	23.80	-2.670	-122.968	281 :1
Hư từ + NV	**85.4**	**29.77**	**-2.475**	**-91.027**	**315 :1**

Bảng 5 : Lọc hư từ và NV trên kho dữ liệu CST

Từ kết quả thử nghiệm trên, chúng tôi thấy lọc hư từ, các từ không có ý nghĩa trong kho dữ liệu và kết hợp với lọc NV cho kết quả tốt nhất trong xử lí gom nhóm tập dữ liệu hỏi đáp Tiếng Việt. Ngoài ra, khi lọc hư từ và NV cũng làm giảm đi một phần đáng kể số chiều của tập dữ liệu, làm giảm đáng kể thời gian xử lí của thuật toán gom nhóm.

b. Lọc những từ xuất hiện nhiều

Có những nghiên cứu [6], [12] đề xuất lọc bỏ những từ xuất hiện thường xuyên trong toàn tập dữ liệu, vì nó không có nghĩa trong gom nhóm và cũng giảm đi số chiều của tập dữ liệu.

U(%)	Dim(%)	NMI	Purity	Avg	DB	(Max :Min)
100	100	0.221	0.606	-151.900	-2.312	279 :1
50	99.6	0.196	0.585	-127.237	-2.768	435 :1
30	**99.1**	**0.288**	**0.709**	**-108.418**	**-2.454**	**475 :1**
20	**98.3**	**0.333**	**0.703**	**-87.610**	**-2.306**	**467 :1**
10	96.4	0.200	0.603	-62.480	-2.192	714 :1
7	95.1	0.157	0.573	-51.826	-2.069	775 :1
5	93.7	0.137	0.554	-42.730	-1.710	833 :1
3	90.7	0.094	0.520	-32.199	-2.000	872 :1
2	88.1	0.093	0.519	-26.030	-1.968	877 :1
1	82.3	0.070	0.504	-17.684	-1.495	921 :1
0.5	74.0	0.082	0.504	-10.931	-0.701	969 :0

Bảng 6 : Loại bỏ các từ xuất hiện nhiều hơn U% trong kho dữ liệu TH

U(%)	Dim(%)	Accuracy(%)	Avg	DB	Max : Min
100	100	22.87	-157.789	-2.675	218 :1
50	99.6	34.83	-126.314	-2.642	317 :1
30	98.8	36.76	-101.782	-2.676	401 :1
20	97.9	38.65	-78.130	-2.340	448 :1
10	95.9	61.45	-54.998	-2.520	698 :1
7	94.6	75.30	-47.013	-2.096	845 :1
5	93.1	76.74	-39.404	-1.904	877 :1
3	90.5	80.54	-30.074	-1.834	944 :1
2	87.2	87.86	-21.512	-1.456	1028 :1
1	80.6	83.80	-13.514	-0.997	1043 :1
0.5	74.2	78.73	-9.895	-0.800	1061 :1

Bảng 7: Loại bỏ các từ xuất hiện nhiều hơn U% trong kho dữ liệu CST

Kết quả thử nghiệm lọc bỏ các từ xuất hiện nhiều hơn một số U% (từ 20% đến 50%) cho kết quả cải thiện nhưng không rõ ràng lắm. Có điều đáng lưu ý là khi lọc các từ xuất hiện nhiều số chiều không giảm số chiều dữ liệu mà còn làm cho việc gom nhóm có khuynh hướng phân dữ liệu về một nhóm.

c. Lọc các từ ít xuất hiện trong tập dữ liệu

Các từ xuất hiện ít có thể xem như nhiễu hay ít có ý nghĩa trong hoạt động gom nhóm dữ liệu. Trong các công trình nghiên cứu đối với Tiếng Anh đã chỉ ra việc lọc các từ xuất hiện bé hơn L lần [3,30] mang lại nhiều kết quả tốt và được sử dụng nhiều trong gom nhóm dữ liệu. Vì thế, chúng tôi tiến hành thử nghiệm lọc các từ xuất hiện kém hơn L lần trong 2 tập dữ liệu trên.

L	Dim(%)	NMI	Purity	Avg	DB	(Max :Min)
1	100	0.221	0.606	-152.900	-2.312	279 :1
2	54.6	0.234	0.632	-150.184	-2.461	289 :1
3	**41.1**	**0.231**	**0.631**	**-146.944**	**-2.445**	**257 :1**
5	**28.7**	**0.277**	**0.703**	**-143.296**	**-2.800**	**271 :1**
9	**20.6**	**0.267**	**0.687**	**-138.193**	**-2.526**	**246 :1**
12	17.6	0.236	0.645	-133.736	-2.453	285 :1
20	12.2	0.236	0.646	-126.451	-2.471	249 :1
28	9.8	0.254	0.665	-120.106	-2.552	264 :1
30	9.0	0.246	0.663	-118.824	-2.529	292 :1
38	7.9	0.265	0.680	-112.393	-2.458	93 :1
50	6.4	0.272	0.698	-105.649	-2.752	257 :1

Bảng 8: Lọc các từ xuất hiện bé hơn L trong kho dữ liệu TH

L	Dim(%)	Accuracy	Avg	DB	Max : Min
1	100	22.89	-157.789	-2.675	108 :1
2	**57.3**	**23.09**	**-154.439**	**-2.625**	**298 :1**
3	**43.5**	**21.91**	**-153.410**	**-2.598**	**300 :1**
5	**31.9**	**22.18**	**-151.337**	**-2.625**	**285 :1**
9	22.4	19.91	-146.044	-2.598	274 :1
12	19.4	20.28	-143.417	-2.597	268 :1
20	13.0	16.83	-135.935	-2.686	256 :1
28	11.0	16.38	-129.294	-2.709	242 :1

30	10.4	15.03	-126.455	-2.646	257 :1
38	8.9	16.83	-121.777	-2.618	249 :1
50	7.0	16.11	-112.919	-2.577	248 :1

Bảng 9 : Lọc các từ xuất hiện bé hơn L trong kho dữ liệu CST

Từ kết quả trên, chúng tôi có nhận xét, việc lọc các từ xuất hiên bé hơn L lần (từ 2 đến 9) trong 2 tập dữ liệu không những cải thiện được kết quả gom nhóm mà còn số chiều của tập dữ liệu giảm đi đáng kể.

d. Cách đánh trọng số các từ

Trong phần này chúng tôi sẽ đánh giá sự ảnh hưởng của cách đánh trọng số trong mô hình vector. Dựa trên 4 cách đánh trọng số cơ bản như sau : Tf – tần suất từ, Tf*idf – tần suất từ* nghịch đảo tần xuất tài liệu, To – số lần xuất hiện của từ, Bo – cách đánh trọng số nhị phân (xuất hiện là 1, không xuất hiện là 0)

Trọng số	**NMI**	**Purity**	**(Max :Min)**
To	0.221	0.606	279 :1
Tf	0.335	0.775	92 :1
Bo	0.388	0.759	218 :1
Tf*idf	**0.381**	**0.800**	**91 :1**

Bảng 10: Cách đánh trọng số khác nhau trên kho dữ liệu TH

Trọng số	**Accuracy(%)**	**Max : Min**
To	22.89	108 :1
Tf	30.32	91 :1
Bo	30.86	164 :1
Tf*idf	**42.81**	**109 :18**

Bảng 11: Cách đánh trọng số khác nhau trên kho dữ liệu CST

Từ kết quả thử nghiệm chỉ ra cách đánh trọng số **Tf* idf** nâng cao được chất lượng gom nhóm dữ liệu hơn ba độ đo To, Tf, Bo. Đặc biệt, việc đánh trọng số theo **Tf*idf** cho kết quả gom nhóm các nhóm có số phần tử đồng đều.

e. Các độ đo tương đồng

Trong xử lí gom nhóm có những công trình đánh giá sự thích hợp các độ đo tương đồng trên tập dữ liệu [1]. Từ đó chọn ra được một độ đo thích hợp nhất cho dữ liệu đó [4]. Sau đây chúng tôi sẽ tiến hành kiểm nghiệm các độ đo thông dụng khác nhau trên hai tập dữ liệu TH và CST. Để từ đó, chỉ ra độ đo phù hợp trên các tập dữ liệu thử nghiệm của chúng tôi.

Độ đo	NMI	Purity	DB	Avg	Max : Min
Euclidean	0.221	0.606	-2.312	-152.900	179 :1
Cosine	**0.340**	**0.770**	**-3.456**	**-171.782**	**95 :1**
Jaccard	0.347	0.769	-3.268	-174.843	154 :0
Mahatan	0.109	0.523	-1.427	-171.333	928 :0
Dice	0.311	0.736	-3.223	-175.755	195 :1
Correlation	**0.344**	**0.779**	**-3.456**	**-171.818**	**98 :1**

Bảng 12: Các độ đo tương đồng khác nhau trên kho dữ liệu TH

Độ đo	Accuracy	DB	Avg	Max : Min
Euclidean	22.89	-2.675	-157.789	108 :1
Cosine	**28.88**	**-3.396**	**-181.423**	**120 :1**
Jaccard	69.78	-2.447	-192.370	776 :0
Mahatan	93.55	-15.04	-173.837	1029 :0
Dice	35.29	-2.972	-185.972	258 :1
Correlation	**29.88**	**-3.362**	**-181.385**	**123 :1**

Bảng 13: Các độ đo tương đồng khác nhau trên kho dữ liệu CST

Từ kết thử nghiệm trên chỉ ra rằng các độ đo Euclidean, jaccard, Mahatan, Dice không thích hợp cho xử lí gom nhóm dữ liệu (dữ liệu chưa được trích chọn đặc trưng), các độ đo thích hợp cho xử lí gom nhóm dữ liệu trên là độ đo **Cosine, Correlation.**

f. Sử dụng phép biến đổi SVD (Singular value decomposition)

Khi xử lí tập câu hỏi trong hệ thống hỏi đáp, mặc dù độ dài câu hỏi không dài nhưng khi mô hình vector hóa thì có số chiều rất cao, lên đến hàng ngàn chiều. Chính vì vậy, thời gian cần xử lí gom nhóm rất lớn nên trong rất nhiều công trình nghiên cứu xử lí gom nhóm dữ liệu văn bản người ta thường sử dụng các phép biến đổi để đưa dữ liệu về dạng ít chiều hơn [5][10]. Một phương pháp biến đổi phổ biến dùng trong xử lí ngôn ngữ hiện nay là SVD, chúng không những làm giảm đáng kể số chiều của dữ liệu mà còn làm tăng độ chính xác hơn trong các bước xử lí gom nhóm dữ liệu văn bản.

Dim	Accuracy	DB	Avg	Max: Min
2	95.11	-0.663	0.000	63:12
4	90.14	-1.019	-0.001	112:1
5	**85.89**	**-1.163**	**-0.001**	**134:5**
7	83.44	-1.236	-0.002	110:1
10	80.81	-1.288	-0.003	190:1
15	77.29	-1.415	-0.006	202:1
25	76.84	-1.578	-0.013	291:1
45	66.52	-1.963	-0.029	346:1

Bảng 4.15: Sử dụng biến đổi SVD đưa kho dữ liệu CST giảm số chiều còn Dim

Từ kết quả trên chỉ ra việc sử dụng phép biến đổi SVD để giảm số chiều còn lại trong đoạn [4:30] kết quả gom nhóm cho kết quả tốt hơn nhiều so với khi chưa sử dụng phép biến đổi SVD. Ngoài ra, việc sử dụng phép biến đổi SVD làm tăng đáng kể tốc độ xử lí của thuật toán gom nhóm.

5. Kết luận

Ngày nay, cùng với sự bùng nổ thông tin đã tạo ra vô số kho dữ liệu số khổng lồ và việc tìm kiếm, khai thác thông tin trong các kho dữ liệu khổng lồ đó đòi hỏi tốn nhiều thời gian và công sức. Chính vì vậy nên việc lựa phương pháp gom nhóm và tinh chỉnh các đặc trưng thích hợp cho việc gom nhóm các kho dữ liệu khổng lồ thành các nhóm nhỏ hơn để nhanh và chính xác hơn cho công việc tìm kiếm, khai thác là cần thiết.

Trong nghiên cứu này, chúng tôi đã thử nghiệm đánh giá các kỹ thuật tinh chỉnh đặc trưng từ trên tập dữ liệu thu thập từ hệ thống hỏi đáp đối thoại doanh nghiệp TP HCM. Qua quá trình thử nghiệm đánh giá, chúng tôi rút ra kết luận như sau: Lọc loại bỏ đặc trưng từ xuất hiện nhiều hơn từ [20% - 30%] trong kho dữ liệu, Loại bỏ những từ xuất hiện nhỏ hơn [3-9] lần trong kho dữ liệu, loại bỏ hư từ và các từ xuất hiện nhiều nhưng không có nghĩa, lọc những cụm danh từ, danh từ và động từ, đánh trọng số bằng phương pháp Tf*idf, sử dụng phương pháp biến đổi giảm chiều SVD còn từ [4-15] chiều. Tất cả những tinh chỉnh trên không những làm giảm số chiều đáng kể của tập dữ liệu mà còn nâng cao đáng kể chất lượng của thuật toán gom nhóm❑

TÀI LIỆU THAM KHẢO

[1] A Huang (2008), *Similarity Measures for Text Document clustering*, Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC 2008), Christchurch New Zealand, pp 49-56.

[2] A Rangrej, et al (2011), *Comparative study of clustering techniques for short text documents*, Proceedings of the 20th international conference companion on World wide web (WWW '11), ACM, pp 111-112.

[3] C ISSAL, M EBBESSON (2010), *Document Clustering*, Master of Science thesis, Chalmers University of Technology, Sweden.

[4] Eduard Hovy, et al (2000), *Question Answering in Webclopedia*, In Proceedings of the Ninth Text REtrieval Conference (TREC-9 (2000), pp. 655-664.

[5] G Cong, et al (2008), *Finding Question-Answer Pairs from Online Forums*, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore.

[6] Hasan, et al (1999), *Document Clustering: Before and After Singular Value Decomposition*, Sapporo, Japan, Information Processing Society of Japan, pp. 47-55.

[7] K Lerman (1999), Document clustering in reduced dimention vector model, USC information sciences institute, (unpublished, last visited 09/02/2011).

[8] Ji-Rong Wen, et al (2001), *Clustering User Queries of a Search Engine*. In WWW '01: Proceedings of the 10th international conference on World Wide Web, pp. 162-168.

[9] NA Samat,et al (2009), *Malay documents clustering algorithm based on singular value decomposition*, Journal of theoretical and applied information technology (JATIT), pp 180-186.

[10] Hồ Tú Bảo (2010), *Các công cụ xử lí tiếng Việt như: tách từ, gán nhãn từ loại thuộc Đề tài cấp nhà nước*, nhánh đề tài xử lí văn bản, download từ Website http://vlsp.vietlp.org:8080/demo/.

[11] Trần Mai Vũ, et al (2009), *Hệ thống hỏi đáp tiếng Việt sử dụng mối quan hệ rút trích ngữ nghĩa trong kho văn bản tiếng Việt*, Hội thảo CNTT quốc gia, Hà nội.

[12] UBND TPHCM, Website đối thoại doanh nghiệp Tp. HCM, Website http://www.doithoaidn.hochiminhcity.gov.vn/.

[13] Phan Thị Tươi, Nguyễn Chánh Thành, Huynh T.N.T (2010), *Question semantic analysis in Vietnamese QA system*, Adv. In intelligent inform and Database System, SCI 283, pp 29-40.

Abstract

Extracting methods of word features in Vietnamese question set clustering

The research focuses on extracting methods of word features in Vietnamese question set serving for clustering. Depending on clustering experimental results of word features extracting methods on Vietnamese question sets, we have showed validity values of word features in Vietnamese question clustering.

Key words: *word features, extracting methods, question set, clustering*