Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000215

ỨNG DỤNG MÔ HÌNH ĐỒ THỊ TRONG TÓM TẮT ĐA VĂN BẢN TIẾNG VIỆT

Nguyễn Thị Ngọc Tú[1], Nguyễn Thị Thu Hà [1], Lê Thanh Hương[2], Hồ Ngọc Vinh[3], Đào Thanh Tĩnh[4], Nguyễn Ngọc Cương[5]

[1] Khoa CNTT, Đại học Điện lực, 235 Hoàng Quốc Việt, Từ Liêm, Hà Nội
{ hantt, tuntn}@epu.edu.vn
[2]Viện CNTT và TT, Đại học Bách Khoa Hà Nội
huonglt@soict.hut.edu.vn
[3]Khoa CNTT, Trường Đại học Sư phạm Kỹ thuật Vinh
hnvinh.skv@moet.edu.vn
[4]Khoa CNTT, Học viện Kỹ thuật quân sự
tinhdt@mta.edu.vn
[5]Khoa Công nghệ và An ninh thông tin, Học viện An ninh nhân dân
Cuongnn.hvan@gmail.com

TÓM TẮT - *Tóm tắt đa văn bản được mở rộng từ tóm tắt đơn văn bản với mục đích tổng hợp thông tin cô đọng nhất từ nhiều nguồn văn bản khác nhau. Trong bài báo này, chúng tôi trình bày một phương pháp tóm tắt đa văn bản dựa trên cách tiếp cận mô hình đồ thị. Trọng số của mỗi câu được thể hiện tại các nút của đồ thị và độ tương tự giữa các câu là trọng số các nhánh của đồ thị. Đánh giá tóm tắt sử dụng độ đo ROUGE với 200 cụm văn bản tiếng Việt, kết quả cho thấy rằng, phương pháp chúng tôi đề xuất thực sự có hiệu quả và có thể dễ dàng triển khai thành những ứng dụng thực tế.*

Từ khóa: *tóm tắt đa văn bản, mô hình đồ thị, giảm chiều đặc trưng, mô hình chủ đề, tiếng Việt.*

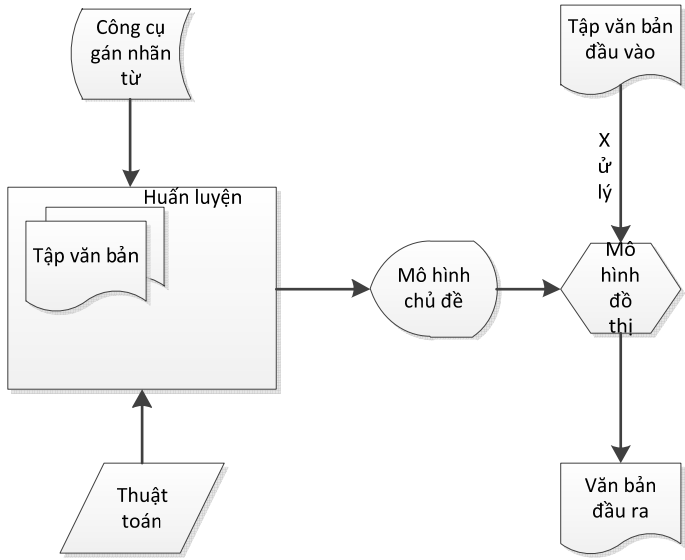
I. GIỚI THIỆU

Sự phát triển của công nghệ thông tin đã biến thế giới ngày nay thành một thế giới thông tin hoàn hảo (*perfect information world*). Cùng với sự gia tăng nhanh chóng của lượng dữ liệu trên Internet, các công cụ khai phá dữ liệu cũng được xây dựng nhằm khai phá tri thức từ những trang tin này như phân loại văn bản, tra cứu văn bản, tìm kiếm, và tóm tắt văn bản,… Trong đó, tóm tắt văn bản là một trong những hướng nghiên cứu được các nhà nghiên cứu quan tâm trong thời gian gần đây, nó làm rút gọn đi những nội dung thông tin dư thừa trong văn bản để lại văn bản tóm tắt ở trạng thái cô đọng nhất [19].

Trong suốt hơn 50 năm phát triển của tóm tắt văn bản, đã có rất nhiều các phương pháp được đề xuất nhằm mục đích xây dựng các hệ thống tóm tắt văn bản tự động thỏa mãn yêu cầu của người dùng. Các nghiên cứu về tóm tắt văn bản tập trung vào hai cách tiếp cận chính là cách tiếp cận dựa trên trích xuất (*extraction*) và tóm lược (*abstractions*). Trong đó, cách tiếp cận dựa trên trích xuất là phổ biến hơn cả, bởi độ phức tạp không quá lớn và vẫn đảm bảo được yêu cầu của một văn bản tóm tắt cần đạt được [22][30]. Mặt khác, đối với ngôn ngữ tiếng Việt, một số công cụ hỗ trợ trong việc xây dựng cách biểu diễn ngôn ngữ như VietwordNet chưa được xây dựng hoàn chỉnh, rất khó khăn trong việc xây dựng cách biểu diễn tương đương hoặc đồng nghĩa.

Phương pháp tóm tắt văn bản sử dụng mô hình đồ thị vô hướng có trọng số đã được R.Mihalcea sử dụng trong tóm tắt văn bản tiếng Anh từ năm 2004 [17]. Trong báo cáo này văn bản được biểu diễn dưới dạng đồ thị, mỗi đỉnh trong đồ thị biểu diễn một câu trong văn bản, các cạnh nối giữa hai đỉnh biểu diễn độ tương đồng ngữ nghĩa giữa hai câu tương ứng với hai đỉnh đó. Đối với tóm tắt văn bản tiếng Việt nhóm tác giả Trương Quốc Định, đã đề xuất một phương pháp tương tự trong đó có sử dụng 3 thuật toán thống kê dựa trên từ vựng để tính toán độ tương tự giữa các câu là Jaro, Contrast Model và Jaccard. Để tính độ quan trọng câu được tính theo thuật toán PageRank[1]. Trong các phương pháp sử dụng mô hình đồ thị để tạo ra các bản tóm tắt tự động thường chỉ đề cập đến độ tương đồng ngữ nghĩa của câu, nhưng đối với xử lý ngôn ngữ tự nhiên có rất nhiều các đặc trưng vì vậy việc lựa chọn các đặc trưng để tính toán có ảnh hưởng rất lớn đến chất lượng của tóm tắt. Trong bài báo này chúng tôi sử dụng mô hình đồ thị có trọng số nhưng thêm vào đó là trọng số của câu tại mỗi nút. Ngoài ra chúng tôi còn giảm chiều đặc trưng bằng mô hình chủ đề theo phương pháp tiếp cận dựa trên mô hình xác suất có điều kiện [11].

Đối với tiếng Việt, hiện nay cũng có nhiều phương pháp được đề xuất, tuy nhiên các đề xuất này thường sử dụng lại các phương pháp đã áp dụng cho tiếng Anh [25]. Một số các khác biệt về ngôn ngữ đều được xử lý thông qua các công cụ xử lý tách từ, nhận dạng từ,… [25]. Chúng tôi cũng nghiên cứu đặc điểm của ngôn ngữ tiếng Việt và nhận thấy rằng, tiếng Việt là ngôn ngữ đơn âm tiết, khó khăn khi tách từ, bởi các từ trong tiếng Việt không dựa trên khoảng trắng. Ví dụ các từ: chuẩn_bị, xử_lý,… là những từ ghép, cần phải nhận dạng và dùng các công cụ tách từ phù hợp khi xử lý. Chính vì điều này, xử lý ngôn ngữ tự nhiên tiếng Việt là một thách thức cần được giải quyết. Trong bài báo này, chúng tôi đã sử dụng một phương pháp cải tiến bài toán tóm tắt văn bản tiếng Việt so với phương pháp thông thường bằng cách sử dụng tập từ chủ đề tiếng Việt [11]. Tại pha tóm tắt, không cần sử dụng công cụ tách và gán nhãn từ để xử lý văn bản đầu vào, nhờ vậy pha tóm tắt sẽ giảm bớt độ phức tạp tính toán về mặt thời gian (Hình 1).



Hình 1. Quy trình tóm tắt văn bản tiếng Việt.

Phần còn lại của bài báo này được cấu trúc như sau: Phần II giới thiệu các nghiên cứu liên quan trong lĩnh vực tóm tắt văn bản trong và ngoài nước. Phần III giới thiệu phương pháp xây dựng mô hình chủ đề có tác dụng giảm bớt độ phức tạp tính toán về mặt thời gian và phương pháp tóm tắt văn bản tiếng Việt dựa trên mô hình đồ thị. Kết quả thực nghiệm đánh giá bằng độ đo chính xác ROUGE được trình bày trong phần IV và cuối cùng là kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Tóm tắt đa văn bản được mở rộng từ tóm tắt đơn văn bản với mục đích tổng hợp thông tin cô đọng nhất từ nhiều nguồn văn bản khác nhau. Do vậy thường các phương pháp tóm tắt đa văn bản được xây dựng từ các phương pháp tóm tắt đơn văn bản. Trong số các phương pháp hiện có thì các thuật toán học dựa trên đồ thị đã có hiệu quả tốt trong các truy vấn câu. Cụ thể một đồ thị có trọng số được xây dựng, mỗi câu được mô phỏng là một nút, mối quan hệ giữa các câu được mô hình hóa như một cạnh có hướng hoặc vô hướng [28][32][33]. Mô hình đồ thị phân lớp câu trong truy vấn tóm tắt đa văn bản cũng đã được Furu Wei và các cộng sự đề xuất trong báo cáo của mình năm 2008. Trong báo cáo này một đồ thị có trọng số được đề xuất để xác định những ảnh hưởng của các câu trong nội văn bản và liên văn bản, từ đó tạo ra một phân lớp các câu trong tóm tắt đa văn bản [29].

Các mô hình chủ đề [6] cung cấp một cách tiếp cận hiệu quả trong tóm tắt văn bản bằng cách cung cấp các chỉ dẫn xác suất rõ ràng và nghiêm ngặt hơn các phương pháp khác [15]. Đến nay, mô hình chủ đề đã được khai thác rộng rãi trong lĩnh vực tóm tắt bằng cách khai thác các chủ đề tiềm ẩn trong tập văn bản [31].

Một sự kết hợp giữa mô hình chủ đề và học bán giám sát dựa trên đồ thị cho các truy vấn trong tóm tắt đa văn bản được nhóm tác giả Yanran Li và Sujian Li đề xuất năm 2014[15]. Một mô hình đồ thị hai lớp (lớp câu và lớp chủ đề) được đưa ra với cách tiếp cận là mô hình quan hệ giữa các chủ đề và câu.

Đối với các nghiên cứu về tóm tắt tự động văn bản tiếng Việt gần đây cũng đã có một số công trình công bố: Nguyễn Lê Minh tóm tắt văn bản tiếng Việt bằng vector hỗ trợ SVM (Support Vector Machine) [20]. Đỗ Phúc và các cộng sự rút trích nội dung chính của khối thông điệp bằng phương pháp gom cụm đồ thị [2]. Nguyễn Hoàng Anh Tú với phương pháp sử dụng mô hình đồ thị trong tóm tắt văn bản tiếng Việt [26]. Ngoài ra còn có sự góp mặt của nhóm tác giả Lê Thanh Hương sử dụng cấu trúc ngôn ngữ tiếng Việt đối với hệ thống tóm tắt tự động [4]. Gần đây trong một báo cáo về " giải pháp tóm tắt văn bản tiếng Việt tự động" nhóm tác giả Trương Quốc Định và Nguyễn Quang Dũng cũng đã đề cập đến phương pháp dựa trên mô hình đồ thị có trọng số. Mỗi đỉnh của đồ thị biểu diễn một câu, cạnh nối hai câu có gán trọng số thể hiện độ tương đồng ngữ nghĩa của chúng và cuối cùng một giải thuật PageRank dựa trên đồ thị được tùy biến để tích hợp độ tương tự câu. Sau cùng các câu quan trọng nhất sẽ được trích rút trong văn bản tóm tắt [1].

III. TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN MÔ HÌNH ĐỒ THỊ

A. Xây dựng mô hình chủ đề

Các tri thức hiện nay vẫn đang được số hóa và lưu trữ trong các trang tin tức, blog, bài báo khoa học, các trang web và các mạng xã hội,... quá nhiều thông tin lưu trữ, do đó sẽ rất khó khăn để tìm kiếm và tổ chức dữ liệu, cũng như định nghĩa một dữ liệu cụ thể. Do vậy, chúng ta cần những công cụ tính toán mới giúp tổ chức, tìm kiếm và hiểu những lượng lớn thông tin.

Trong học máy và xử lý ngôn ngữ tự nhiên, mô hình chủ đề là một loại mô hình thống kê để phát hiện ra các "chủ đề" trừu tượng xảy ra trong một bộ sưu tập các tài liệu. Giả sử, cho rằng một tài liệu nói về một chủ đề cụ thể, người ta sẽ kỳ vọng từ đặc biệt để xuất hiện trong các tài liệu nhiều hơn hoặc ít hơn: "dog" và "bone" sẽ xuất hiện thường xuyên hơn trong các tài liệu về những con chó, "cat" và "meow" sẽ xuất hiện trong các tài liệu về những con

mèo và "the", "is" sẽ xuất hiện như nhau trong cả hai. Một mô hình chủ đề sử dụng mô hình toán học, cho phép kiểm tra một tập tài liệu và phát hiện, dựa trên số liệu thống kê của các từ trong mỗi tài liệu, dựa vào đó có thể dự đoán được chủ đề của văn bản là gì.

Bảng 1. Các từ chủ đề trong tập mô tả của Andrews năm 2009.

Theatre	Music	League	Prison	Rate	Pub	Market	Railway	Air
Stage	Band	Cup	Years	Cent	Guinness	Stock	Train	Aircraft
Arts	Rock	Season	Sentence	Inflation	Beer	Exchange	Station	Flying
Play	Song	Team	Jail	Recession	Drink	Demand	Steam	Flight
Dance	Record	Game	Home	Recovery	Bar	Share	Rail	Plane
Opera	Pop	Match	Prisoner	Economy	Dringking	Group	Engine	Airport
Cast	Dance	Division	serving	Cut	alcohol	news	track	Pilot

Trong nghiên cứu gần đây nhất về xây dựng mô hình chủ đề cho tiếng Việt, nhóm nghiên cứu Ha Nguyen Thi Thu đã xây dựng mô hình chủ đề dành cho tiếng Việt dựa trên tập từ lõi và xác suất điều kiện. Trong đó, từ lõi được coi là từ có tần suất xuất hiện lớn nhất trong chủ đề đó. Để xây dựng mô hình chủ đề này, các văn bản được đưa vào tập huấn luyện và được gán nhãn trước, sau đó họ tách thành tập các danh từ và sau đó dựa trên xác suất điều kiện để xây dựng tập thuật ngữ đối với mỗi chủ đề. Với cách tiếp cận này, họ đã giảm được chi phí về mặt thời gian khi xây dựng những hệ thống ứng dụng thực tế hơn so với phương pháp truyền thống [11] và cũng giảm được chi phí xử lý so với một số các công bố trước đây của họ [9].

Hình 2 dưới đây mô tả thuật toán xây dựng mô hình chủ đề.

THUẬT TOÁN XÂY DỰNG MÔ HÌNH CHỦ ĐỀ

Đầu vào:

- D: Tập văn bản huấn luyện đã được gán nhãn tương ứng với các chủ đề C;
- VnTagger: Công cụ nhận dạng, tách từ;
- C: Tập các chủ đề

Đầu ra:

- T: Tập các từ được gán nhãn tương ứng với mỗi C.

Khởi tạo:

V=ϕ ;

n=count(S); n'=count(S');

G=ϕ ;G'=ϕ ;

1. For each d_i in C_k do

1.1 $V_k \leftarrow$ Vntagger(d_i);

2. For each C_k do

2.1.1 If $w(j) \in V_k$ then

2.1.1.1 n(j) $\leftarrow$ n(j) +1; // đếm số lần xuất hiện $w(j)$ trong mỗi chủ đề C_k

2.1.1.2 N_k=argmax(n(j)); // Lấy tần suất lớn nhất của từ wj trong mỗi chủ đề C_k

3. For each C_k do

3.1 For all w in V

3.1.1 if $Pr(w(i)|N_k) <> 0$ then $V_k \leftarrow w(i)$; // cho các từ $w(i)$ vào tập V_k của C_k

Hình 2. Thuật toán xây dựng mô hình chủ đề

B. Tóm tắt văn bản tiếng Việt dựa trên mô hình đồ thị

1. Trọng số câu

Giả sử D={d_1, d_2, ..., d_n} là một tập các văn bản, D được biểu diễn thành tập các câu như sau: $D = \{S_1, S_2, \dots, S_m\}$. Với S_i là các câu được tách ra từ tập văn bản D.

Mỗi câu S_i được gán giá trị trọng số thông qua các tính trọng số của câu, có nghĩa là $<S_i,W_i>$, với mỗi câu S_i tương ứng có một trọng số W_i tương ứng. Lúc này D được biểu diễn lại như sau:

$D = \{< S_1, W_1 >, < S_2, W_2 >, \dots, < S_m, W_m >\}$.

Các nghiên cứu từ trước thường áp dụng tính trọng số của câu dựa trên phương pháp tần suất từ, tần suất nghịch đảo văn bản tf*idf. Trong bài báo sử dụng cách tiếp cận dựa trên thuật ngữ. Do đó phương pháp tính trọng số thuật ngữ được áp dụng như sau:

$$\Phi_{t_i} = \frac{N_{t_i}}{\sum_{j=1}^{m} N_{t_i}} \quad (1)$$

Trong đó:

- ϕ_{t_i} : Là trọng số của thuật ngữ t_i trong câu.

- N_{t_i} : Là số lần xuất hiện của thuật ngữ t_i trong tập văn bản.

-$\sum_{j=1}^{m} N_{t_i}$: Tổng số lần xuất hiện của tất cả các thuật ngữ trong văn bản.

Trọng số của câu được tính bằng tổng trọng số của tất cả các thuật ngữ trong câu.

$$W_i = \sum_{m=1}^{k} t_{im} \quad (2)$$

Với W_i là trọng số của câu i trong tập văn bản. Các t_{im} là các thuật ngữ trong câu thứ i.

2. Độ tương đồng câu

Trong bài báo này chúng tôi tính độ tương đồng giữa hai câu dựa trên tổng các độ đo tương tự của từng cặp từ được đối sánh.

Độ tương tự của từ:

Về mặt cấu trúc, một đoạn văn bản gồm nhiều câu, mỗi câu được tạo thành bởi một chuỗi các từ mang các thông tin cần thiết. Phương pháp này được thực hiện dựa vào thông tin về ngữ nghĩa và cú pháp của các từ trong câu.

Với một câu S_i được trích rút ra tập các thuật ngữ như sau:

$$T_i = \{t_1, t_2, \dots, t_n\}$$

Biểu thức trên được viết lại bổ sung thêm vị trí từ trong câu:

$$T_i = \{< t_1, v_1 >, < t_2, v_2 >, \dots, < t_n, v_n >\}$$

Giả sử ta có hai câu T_1, T_2:

$$T_1 = \{< t_{11}, v_{11} >, < t_{12}, v_{12} >, \dots, < t_{1n}, v_{1n} >\}$$

$$T_2 = \{< t_{21}, v_{21} >, < t_{22}, v_{22} >, \dots, < t_{2m}, v_{2m} >\}$$

Trong đó:

- $t_{1,i}$ là từ chủ đề thứ i trong câu T_1.

- $t_{2,j}$ là từ chủ đề thứ j trong câu T_2.

Do tiếng Việt chưa có hệ thống Wordnet để tính toán độ tương tự giữa hai từ, do đó, trong công thức này chúng tôi sử dụng độ đo đã được đề xuất bởi Church and Hanks (1990). Độ đo này được gọi là độ tương hỗ giữa các từ Pointwise Mutual Information (PMI):

$$PMI(W_1 = w_1, W_2 = w_2) = log_2 \frac{P(w_1,w_2)}{P(w_1).P(w_2)} \quad (3)$$

Trong đó:

- $P(W_1,W_2)$: là xác suất xuất hiện đồng thời 2 từ W_1 và W_2 trong tập văn bản huấn luyện.

- $P(W_1)$: là xác suất xuất hiện từ W_1 trong tập văn bản huấn luyện.

- $P(W_1)$: là xác suất xuất hiện từ W_2 trong tập văn bản huấn luyện.

Độ tương đồng giữa các câu

Độ tương đồng ngữ nghĩa giữa hai câu được tính theo công thức sau:

$$S_{sentences} = \sum_{k=1,m=1}^{n} PMI(w_{i,k}, w_{j,m}) \quad (4)$$

Trong đó:

- $S_{sentences}$: độ tương đồng giữa các câu s_i, s_j.

- $W_{i,k}$: từ thứ k trong câu *i*.

- $W_{j,m}$: từ thứ m trong câu *j*.

3. Xây dựng đồ thị tóm tắt văn bản

Chúng tôi sử dụng mô hình đồ thị vô hướng có trọng số với các đỉnh biểu diễn các câu cùng với trọng số của câu. Mặt khác các cạnh nối giữa hai câu có gán trọng số, trọng số này chính là độ tương đồng ngữ nghĩa giữa hai câu được kết nối bởi cạnh đó. Sau khi đã tính toán được trọng số câu và độ tương tự giữa các câu, ta dựng mô hình đồ thị ứng dụng trong tóm tắt văn bản như sau:

Bước 1: Khởi tạo đồ thị

- Mỗi đỉnh biểu diễn các câu và trọng số của nó.
- Mỗi cạnh biểu diễn trọng số là độ tương tự giữa các cặp đỉnh trong đồ thị.

Bước 2: Sắp xếp thứ tự ưu tiên

- Duyệt các đỉnh của đồ thị, các đỉnh có trọng số cao được ưu tiên.

Bước 3: Xác định chiều dài văn bản tóm tắt

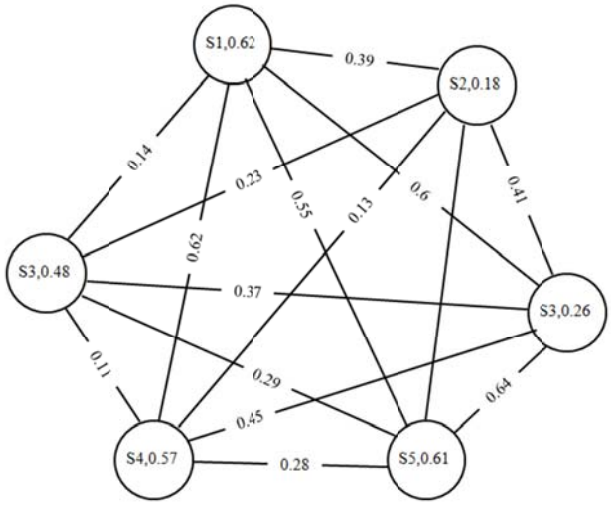
- Dựa trên tỉ lệ tóm tắt r với $r = \frac{chieu\ dai\ van\ ban\ tom\ tat}{chieu\ dai\ van\ ban\ goc}$%.

Bước 4: Lựa chọn thông tin trong văn bản tóm tắt

Dựa trên yêu cầu chiều dài văn bản tóm tắt. Xác định các câu được trích chọn ra từ văn bản gốc. Với điều kiện trọng số cao đồng thời những câu có độ tương đồng cao sẽ được loại bỏ để thay thế bằng câu khác tránh dư thừa thông tin. Thủ tục lựa chọn thông tin như sau:

- Đầu tiên lựa chọn câu có trọng số cao nhất trong đồ thị.
- Kiểm tra lại chiều dài văn bản tóm tắt, nếu chưa đủ thì lựa chọn câu tiếp theo có trọng số cao thứ 2.
- Tính toán độ tương đồng giữa hai câu đã lựa chọn, nếu độ tương đồng lớn hơn một ngưỡng σ, tiến hành loại bỏ câu này và lựa chọn câu có trọng số cao tiếp theo.
- Tiếp tục các bước lựa chọn này, cho tới khi chiều dài văn bản tóm tắt đúng với tỉ lệ yêu cầu thì dừng lại.

Ví dụ 1: Giả sử tập văn bản D gồm có 6 câu như hình 4 sau. Tóm tắt đa văn bản với σ >0.5 và tỉ lệ tóm tắt r=30%.



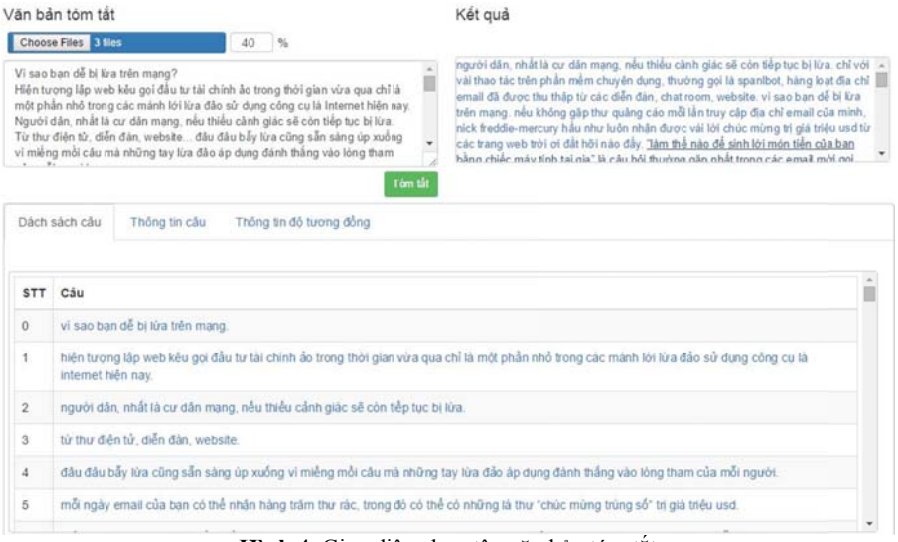
Hình 3. Mô hình đồ thị trong tóm tắt văn bản tiếng Việt.

Sau khi xây dựng được mô hình đồ thị bắt đầu xử lý tóm tắt, giả sử tóm tắt các văn bản gốc trên với tỉ lệ r = 30%. Theo tính toán trên đồ thị, câu 1 và câu 5 sẽ được lựa chọn. Tuy nhiên, do độ tương đồng giữa câu 1 và câu 5 cao >0.5, do đó, xét câu tiếp theo lựa chọn câu thứ 4, tuy nhiên độ tương đồng giữa câu 1 và câu 4 vẫn >0.5, do đó lựa chọn tiếp câu thứ 3. Kiểm tra lại chiều dài văn bản tóm tắt, cho thấy rằng, chiều dài văn bản tóm tắt đã đủ. Do vậy, văn bản tóm tắt cuối cùng bao gồm câu 1 và câu 3.

IV. KẾT QUẢ THỰC NGHIỆM

A. Một số hình ảnh hệ thống tóm tắt đa văn bản tiếng Việt dựa trên mô hình đồ thị

Để phát triển hệ thống tóm tắt đa văn bản tiếng Việt, chúng tôi đã xây dựng hệ thống ứng dụng dựa trên phương pháp đã được đề xuất. Những kết quả ban đầu khá khả quan. Thời gian xử lý tóm tắt nhanh và phản hồi trở lại giao diện người dùng. Dưới đây là một số hình ảnh của hệ thống thực nghiệm:



Hình 4. Giao diện chọn tập văn bản tóm tắt

	S0(0.763)	S1(0.737)	S2(0.763)	S3(0.763)	S4(0.00)	S5(0.684)	S6(0.368)	S7(0.404)	S8(0.474)	S9(0.684)	S10(0.00)
S0(0.763)	xxx	0.00	1.00	0.00	0.00	0.00	0.5	0.423	0.293	0.00	0.00
S1(0.737)	0.00	xxx	0.00	0.00	0.00	0.00	0.5	0.184	0.293	0.00	0.00
S2(0.763)	1.00	0.00	xxx	0.00	0.00	0.00	0.5	0.423	0.293	0.00	0.00
S3(0.763)	0.00	0.00	0.00	xxx	0.00	0.00	0.134	0.423	0.293	0.00	0.00
S4(0.00)	0.00	0.00	0.00	0.00	xxx	0.00	0.00	0.00	0.00	0.00	0.00
S5(0.684)	0.00	0.00	0.00	0.00	0.00	xxx	0.134	0.184	0.293	1.00	0.00

Hình 5. Ma trận hiển thị kết quả thực nghiệm

B. Kết quả thử nghiệm

Trong bài báo này, để đánh giá kết quả của phương pháp tóm tắt đa văn bản đã đề xuất. Chúng tôi sử dụng tập văn bản huấn luyện do nhóm tác giả Lê Thanh Hương xây dựng là kết quả của đề tài cấp Bộ Giáo dục năm 2013 [3]. Chúng tôi sử dụng 3 tập dữ liệu, tổng số gồm 200 cụm văn bản dùng cho thử nghiệm và một tập văn bản tóm tắt mẫu được tạo sẵn.

Ngoài ra, chúng tôi đã sử dụng độ đo ROUGE [34] để đánh giá kết quả đạt được theo công thức sau:

$$C_n = \frac{\sum_{C\in\{Model\ Units\}}\sum_{n-gram\in C} Count_{match}(n-gram)}{\sum_{C\in\{Model\ Unit\}}\sum_{n-gram\in C} Count(n-gram)}$$

$$Ngram(i,j) = BB \cdot exp\left(\sum_{n=i}^{j} w_n \log C_n\right)$$

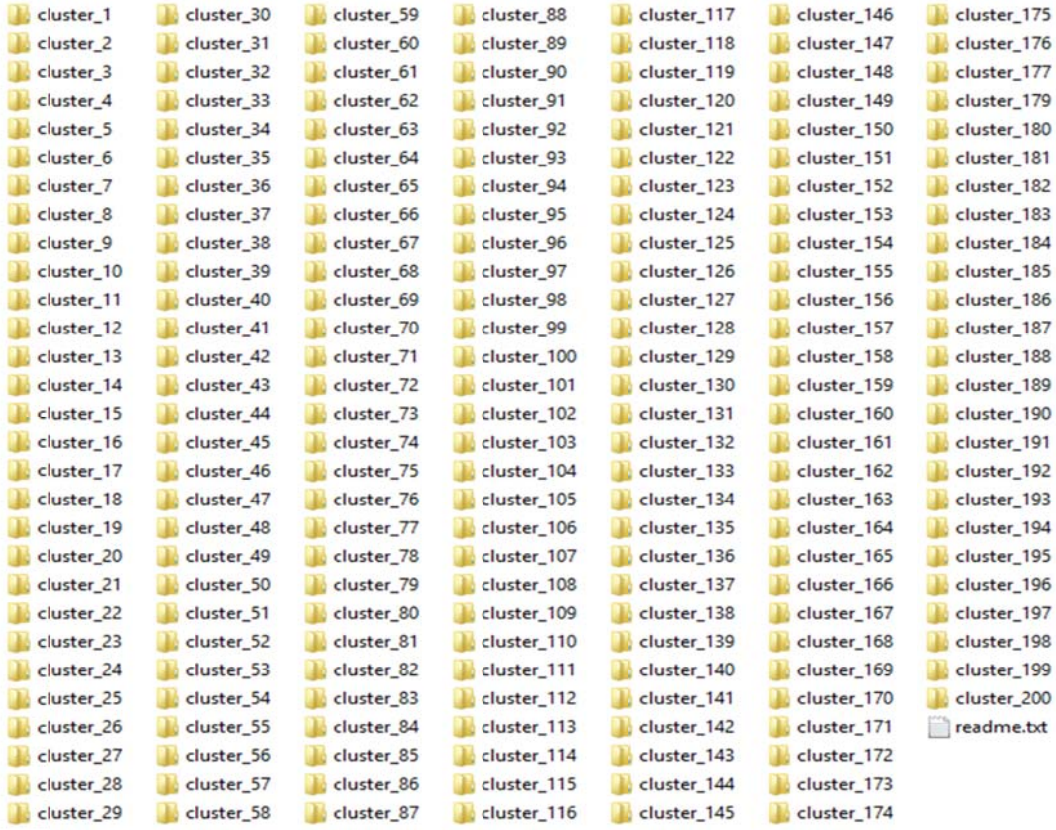
Trong đó - $Count_{match}(n\text{-}gram)$ là số lớn nhất của n-gram đồng thời trong bản tóm tắt được xem xét trong một mô hình.

- $Count(n\text{-}gram)$ là số n-gram trong một mô hình.

- i, j nhận các giá trị từ 1-4.

- w_n nhận giá trị $1/(j-i+1)$.

Trong đánh giá kết quả, chúng tôi chỉ xét Ngram với các từ chủ đề đã được xây dựng trong mô hình chủ đề, do vậy kết quả so với các công bố của tác giả Lê Thanh Hương [3] có sự khác nhau.



Hình 6. Tập dữ liệu thử nghiệm

Kết quả thử nghiệm được cho theo bảng 2. dưới đây

Bảng 2. Kết quả thử nghiệm

Tập dữ liệu	ROUGE		
	ROUGE-1	ROUGE-2	ROUGE-N
#1	0.392	0.28	0.4268
#2	0.381	0.46	0.247
#3	0.472	0.623	0.26860

Trong bảng 2, tập dữ liệu thử nghiệm gồm 3 tập (#1, #2, #3) đã được phân theo các chủ đề, các tập dữ liệu này được thử nghiệm và đánh giá với Gram-1, Gram-2 và Gram-N (với N=4). Kết quả được thể hiện tương ứng với ROUGE-1, ROUGE-2, ROUGE-N trên 3 tập dữ liệu trong khoảng 0.3 đến 0.6.

V. KẾT LUẬN

Xử lý tóm tắt đa văn bản có độ phức tạp hơn so với xử lý tóm tắt đơn văn bản bởi nguồn văn bản đầu vào không chỉ là một văn bản mà là nhiều văn bản khác nhau. Trong bài báo này, chúng tôi đã đề xuất phương pháp tóm tắt đa văn bản tiếng Việt dựa trên cách tiếp cận mô hình đồ thị có sử dụng kết hợp giảm chiều đặc trưng bằng cách sử dụng mô hình chủ đề tiếng Việt.

Với những kết quả ban đầu đạt được dựa trên thực nghiệm tập văn bản tiếng Việt cho thấy rằng, phương pháp của chúng tôi đề xuất có thời gian xử lý nhanh, văn bản tóm tắt có kết quả chấp nhận được và thực sự có ý nghĩa để phát triển thành những hệ thống tóm tắt đa văn bản tiếng Việt trong tương lai.

VI. LỜI CẢM ƠN

Chúng tôi trân trọng gửi lời cảm ơn tới nhóm tác giả Lê Thanh Hương, Hà Quang Thụy, Vũ Đức Thi đề tài cấp bộ Giáo dục năm 2013 đã hỗ trợ chúng tôi trong việc cung cấp kho ngữ liệu và công cụ đánh giá dựa trên phương pháp ROUGE. Chúng tôi cũng cảm ơn các chuyên gia về xử lý ngôn ngữ tự nhiên thuộc Đại học Công nghệ - Đại học Quốc gia Hà Nội đã đóng góp ý kiến cho việc xử lý ngôn ngữ tiếng Việt một cách hiệu quả nhất.

VII. TÀI LIỆU THAM KHẢO

[1] Trương Quốc Định, Nguyễn Quang Dũng "Một giải pháp tóm tắt văn bản tiếng Việt tự động" Hội thảo quốc gia lần thứ XV: một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông Hà Nội 03-04/12/2012.

[2] Đỗ Phúc, Mai Xuân Hùng, Nguyễn Thị Kim Phụng, "Gom cụm đồ thị và ứng dụng vào việc rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận", Tạp chí Phát triển Khoa học Công nghệ, Tập 11, Số 05 - 2008, pp. 21-32, 2008.

[3] Lê Thanh Hương, Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt, Đề tài Bộ Giáo dục, 2012-2014.

[4] Nguyễn Trọng Phúc, Lê Thanh Hương, *Tóm tắt văn bản sử dụng cấu trúc diễn ngôn*, Proc of ICTrda08, 2008.

[5] Mark Andrews, Gabriella Vigliocco, *The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation*, Topics in Cognitive Science 2 101–113, 2010.

[6] David Blei, Andrew Ng and Micheal Jordan. Latent dirichlet allocation. In The Journal of Machine Learning Research, 2003.

[7] Dipanjan Das, Andre F. T. Martins, *A Survey on Automatic Text Summarization,* November 21, 2007.

[8] Ha. N. T. T, Quynh. N. H, Tao. N. Q, *A new method for calculating weight of sentence based on amount of information and linguistic score*, International Journal of Advanced Computer Engineering, Vol.4 No.2, pp. 91-95, 2011.

[9] Ha. N. T. T, Quynh. N. H, Khanh N. T. H, Hung L. M, *Optimization for Vietnamese Text classification problem by reducing feature set*, Proc of 6th International Conference on New Trends in Information Science, Service Science and Data Mining, pp. 209-212, 2012.

[10] Ha. N. T. T, Quynh. N. H, Tu. N. N, *A Supervised Learning Method Combine with Dimensionality Reduction in Vietnamese Text Summarization*, Proc IEEE of Computer, communication and application 2013, pp. 69-73, 2013.

[11] Ha Nguyen Thi Thu, Tinh Dao Thanh, Thanh Nguyen Hai, Vinh Ho Ngoc, "'Building Vietnamese Topic Modeling Based on Core Terms and Applying in Text Classification", Proc. of Fifth IEEE International Conference on Communication Systems and Network Technologies, pp. 1284-1288, DOI 10.1109/CSNT.2015.22, 2015.

[12] Makoto Hirohata, Yousuke Shinnaka, Koji Iwano and Sadaoki Furui, *Sentence extraction – based presentation summarization techniques and evaluation metrics*, ICASSP 2005, pp. I – 1065- I – 1068, 2005.

[13] Karel Jezek, Josef Steinberger, *Automatic Text Summarization* (The state of the art 2007 and new challenges), Znalosti ,2008, ISBN 978-80-227-2827-0.

[14] Daniel Jurafsky & James, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall, 2008.

[15] Yanran Li and Sujian Li, *Query-focused Multi-Document Summarization: Combining a Topic Model with Graph-based Semi-supervised Learning*. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1197–1207, Dublin, Ireland, August 23-29 2014.

[16] C. Lopez, V. Prince, and M. Roche, *Text titling applica-tion (demonstration session, to appear),* in Proceedings of EKAW'10, 2010.

[17] Mihalcea, R., "Graph-based ranking algorithms for sentence extraction, applied to text summarization", ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics, Morristown, NJ, USA, pp. 181–184, 2004

[18] Nenkova, A. *Automatic text summarization of newswire: Lessons learned from the document understanding conference*. In Proceedings of AAAI 2005, Pittsburgh, USA.

[19] A. Nenkova and K. McKeown, *Automatic Summarization*, Foundations and Trends® in Information Retrieval Vol. 5, Nos. 2–3 (2011) 103–233.

[20] M. L. Nguyen, Shimazu, Akira, Xuan, Hieu Phan, Tu, Bao Ho, Horiguchi, Susumu, *Sentence Extraction with Support Vector Machine Ensemble*, Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society 2005-11.

[21] Seonggi Ryang, Takeshi Abekawa, Framework of Automatic Text Summarization Using Reinforcement Learning, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 256–265, Jeju Island, Korea, 12–14 July 2012.

[22] Horacio Saggion (2013) *Proceedings of NAACL-HLT 2013*, Atlanta, Georgia, 9–14 June 2013, pages 270–279;

[23] MARIA SOLEDAD PERA, YIU-KAI NG, *A Naïve Bayes Classify for web document summaries created by using word similariy and significant factor*, International Journal on Artificial Intelligence Tools, Vol. 19, No. 4 pp. 465–486, 2010.

[24] Svore, K., Vanderwende, L., and Burges, C. *Enhancing single-document summarization by combining RankNet and third-party sources*. In Proceedings of the EMNLP-CoNLL, pages 448-457, 2007.

[25] Thanh, Le Ha; Quyet, Thang Huynh; Chi, Mai Luong, *A Primary Study on Summarization of Documents in Vietnamese*, Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society 2005-11.

[26] Tu-Anh Nguyen-Hoang, Hoang Khai Nguyen, and Quang Vinh Tran (2010), "An efficient Vietnamese text summarization approach base on graph model". . RIVF, page 1-6. IEEE, (2010).

[27] Dingding Wang, Shenghuo Zhu, Tao Li, Yihong GongMulti-Document Summarization using Sentence-based Topic Models, Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 297–300, Suntec, Singapore, 4 August 2009. 2009.

[28] Xiaojun Wan, Jianwu Yang and Jianguo Xiao. Manifold-ranking based topic-focused multi-document summarization. In Proceedings of International Joint Conference on Artificial Intelligence, 2007.

[29] Furu Wei, Wenjie Li, Qin Lu, and Yanxiang He A Cluster-Sensitive Graph Model for Query-Oriented Multi-document Summarization, 2008.

[30] Kam-Fai Wong, Mingli Wu and Wenjie Li (2008), "Extractive Summarization Using Supervised and Semi-supervised Learning", *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, August 2008,pp. 985–992

[31] Jean-Yves Delort and Enrique Alfonseca. DualSum: a topic-model based approach for update summarization. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics., 2012).

[32] Dengzhong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet and Bernhard Schlkopf. *Ranking on Data Manifolds*. In Proceedings of the Conference on Advances in Neural Information Processing Systems., 2003.

[33] Dengyou Zhou, Olivier Bousquet, Thomas Navin and JasonWeston. *Learning with Local and Global Consistency*. In Proceedings of Advances in neural information processing systems, 2004.

[34] Chin-Yew Lin and Eduard Hovy (2003), "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics", *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003), May 27 - June 1, 2003, Edmonton, Canada.*

[35] http://vlsp.vietlp.org:8080/demo/.

APPLY GRAPHICAL MODEL FOR VIETNAMESE MULTI-DOCUMENT SUMMARIZATION

Nguyen Thi Ngoc Tu, Nguyen Thi Thu Ha, Le Thanh Huong, Ho Ngoc Vinh, Dao Thanh Tinh, Nguyen Ngoc Cuong

ABSTRACT - *Multi-document summarization is expanded from single-document summarization in order to compile the most important information from different sources of document. In this paper, we present a Vietnamese multi-document summarization method based on graphical model. Weighting of each sentence is represented as a node of the graph and the similarity score among sentences is on the edges of the graph. For evaluation, we used ROUGE method with 200 Vietnamese text clusters and the results shown that the method proposed is really effective and can be developed into practical applications.*