Một giải pháp tóm tắt văn bản tiếng Việt tự động

Conference Paper · December 2012

CITATIONS
2

2 authors:

Quoc-Dinh Truong
Can Tho University
35 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Dung Nguyen Quang
Can Tho University
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Một giải pháp tóm tắt văn bản tiếng Việt tự động

Trương Quốc Định
Khoa CNTT-TT
Đại học Cần Thơ
Cần Thơ, Việt Nam
tqdinh@cit.ctu.edu.vn

Nguyễn Quang Dũng
Khoa Nông nghiệp & Sinh học ứng dụng
Đại học Cần Thơ
Cần Thơ, Việt Nam
nqdung@ctu.edu.vn

***Tóm tắt*— Trong bài báo này chúng tôi đề xuất mô hình tóm tắt văn bản tiếng Việt tự động. 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 các đỉnh biểu diễn sự tương tự về ngữ nghĩa giữa hai đỉnh (câu). Giá trị tương tự được biểu diễn dưới dạng trọng số của các cạnh. Chúng tôi sử dụng 3 thuật toán thống kê dựa trên từ vựng để tính độ tương tự câu là Jaro, Contrast Model và Jaccard. Độ quan trọng của đỉnh (câu) được tính bởi thuật toán PageRank, một giải thuật toán học dựa trên đồ thị, được tùy biến để tích hợp độ tương tự câu. Hệ thống sẽ tự động chọn các câu quan trọng nhất (mặc định là 25% tổng số câu) để đưa vào kết quả tóm tắt. Để kiểm chứng tính chính xác của mô hình đề xuất, chúng tôi so sánh kết quả tóm tắt tự động với kết quả tóm tắt của chuyên gia vì thế dữ liệu thực nghiệm sử dụng là khá khiêm tốn (gồm 5 văn bản thuộc các chủ đề khác nhau). Kết quả tóm tắt của hệ thống có độ tin cậy cao vì được đánh giá bởi tập dữ liệu đánh giá được tổng hợp từ 12 nhà khoa học uy tín. Kết quả cho thấy việc kết hợp các thuật toán thống kê với thuật toán xếp hạng dựa trên đồ thị PageRank có tích hợp độ tương tự câu cho độ chính xác khá cao, trong đó thuật toán Contrast model và Jaccard cho kết quả tóm tắt tốt nhất (51.5 và 52%). Ngoài ra, chúng tôi cũng đã thực nghiệm trên tập các bài viết thu thập từ các trang báo mạng với kết quả khả quan.**

Từ khóa : tóm tắt, đồ thị, độ đo tương đồng, PageRank

GIỚI THIỆU

Tóm tắt văn bản [1] đã trở thành một công cụ quan trọng và hữu ích để hỗ trợ và trích chọn thông tin văn bản trong thời đại thông tin phát triển nhanh chóng ngày nay. Tóm tắt văn bản thủ công (được thực hiện bởi con người) đôi khi là một nhiệm vụ khó khăn khi phải làm việc với một văn bản lớn, chứa nhiều thông tin.

Nếu phân loại tóm tắt theo hướng tiếp cận, tóm tắt văn bản có thể được phân thành các loại như: tóm tắt trích chọn (extractive) và tóm tắt trừu tượng (abstractive). Hướng tiếp cận tóm tắt trừu tượng [2] có nghĩa là hệ thống cố gắng hiểu được ý chính của tài liệu rồi sau đó diễn giải chúng dưới dạng ngôn ngữ tự nhiên. Tóm tắt trích chọn [3] được xây dựng bằng cách trích xuất các đơn vị văn bản quan trọng (câu hoặc đoạn văn) từ văn bản gốc, dựa trên phân tích từ/cụm từ, tần số, vị trí hoặc các từ gợi ý để xác định tầm quan trọng của các đơn vị và từ đó trích xuất các đơn vị quan trọng nhất như là tóm tắt.

Về phương pháp tóm tắt, hiện nay trên thế giới đã có nhiều công trình nghiên cứu áp dụng các phương pháp tóm tắt khác nhau [4]: phương pháp TF-IDF , phương pháp phân cụm (Cluster based), phương pháp tiếp cận máy học, phương pháp phân tích ngữ nghĩa tiềm ẩn (LSA), mạng nhân tạo (neural networks), phương pháp Lôgíc mờ (fuzzy logic), phương pháp hồi quy toán học (Mathematical regression) [5], phương pháp dựa trên truy vấn (Query based).

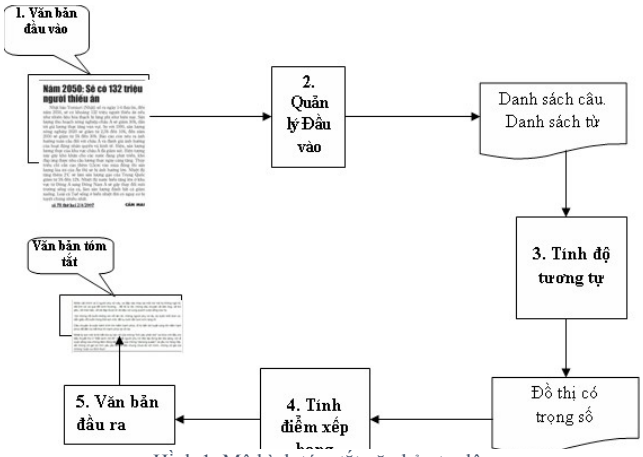
Trong 50 năm qua đã có nhiều công trình nghiên cứu tạo tóm tắt tự động văn bản tiếng Anh, Nhật, Hoa. Một số công trình tiêu biểu: Edmundson [6] đã thử nghiệm 3 tiêu chí đánh giá mới cho các câu để tạo ra tóm tắt tự động, hai trong số đó sử dụng cấu trúc văn bản; công trình của Marcu [7] thì quan tâm đến việc sử dụng phương pháp phân tích cấu trúc diễn ngôn (“discourse parsing” hoặc “rhetorical parsing”) để tạo tóm tắt tự động; công trình của Radev và cộng sự [8] sử dụng khái niệm “trọng tâm” (centroid) để tóm tắt đa văn bản bằng cách trích chọn; công trình của Mihalcea [9] thì sử dụng thuật toán dựa trên đồ thị để tạo tóm tắt tự động.

Đối với các nghiên cứu về tóm tắt tự động văn bản tiếng Việt, gần đây cũng có một số công trình nghiên cứu được công bố như: Nguyễn Lê Minh tóm tắt văn bản tiếng Việt bằng phương pháp phân cụm SVM (Support Vector Machine) [10]; Đỗ Phúc và cộng sự rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận bằng phương pháp gom cụm đồ thị [11]; Nguyễn Trọng Phúc và cộng sự thì trình bày phương pháp tóm tắt văn bản tiếng Việt dựa trên cấu trúc diễn ngôn [12]. Tuy nhiên, kết quả của các nghiên cứu này vẫn chưa được đánh giá cụ thể. Đồng thời một số công cụ có sẵn thì không thích hợp cho tiếng Việt nên kết quả tóm tắt rất thấp, không đáp ứng được yêu cầu người dùng, ví dụ như công cụ Autosummarizer của phần mềm Microsoft Word. Vì vậy trong nghiên cứu này chúng tôi đề xuất phương pháp tóm tắt văn bản tiếng Việt tự động theo hướng tiếp cận rút trích các câu quan trọng của văn bản để đưa vào tóm tắt dựa trên mô hình đồ thị.

Trong phần 2 của bài báo chúng tôi trình bày mô hình tóm tắt văn bản bao gồm các nội dung: quản lý đầu vào; tính độ tương tự; tính điểm xếp hạng. Dữ liệu thực nghiệm, phương pháp đáng giá và kết quả thực nghiệm được giới thiệu trong phần 3. Phần 4 trình bày kết luận và kiến nghị.

MÔ HÌNH TÓM TẮT

Hình 1 trình bày mô hình tóm tắt văn bản tự động được chúng tôi đề xuất.



Hình 1. Mô hình tóm tắt văn bản tự động

A.

Quản lý đầu vào

Văn bản đầu vào có định dạng *.txt hoặc *.doc. Văn bản sẽ được đưa qua bộ lọc để loại bỏ từ dừng (stopwords), những từ này mang ít nghĩa hoặc không có nghĩa, loại bỏ các ký tự không phải chữ cái hoặc chữ số. Quản lý đầu vào còn có nhiệm vụ tách văn bản thành các câu và các từ riêng lẻ để sử dụng cho mục đích tính toán sau này.

1) Tách câu và tách từ: trong nghiên cứu này chúng tôi sử dụng công cụ JVnTextPro do nhóm nghiên cứu về xử lý ngôn ngữ tự nhiên của Khoa Công nghệ - Trường Đại học Quốc gia Hà Nội nghiên cứu và xây dựng [13]. Chức năng chính của gói này như sau:

Đoạn văn bản → Gán nhã câu → Tách từ → Gán nhãn từ loại → Từ loại

Chúng tôi sử dụng JVnTextPro cho giai đoạn lập chỉ mục cho văn bản vì công cụ này có thể nhận biết được các danh từ riêng, có thể nhận biết được từ đơn và từ ghép và có độ chính xác trung bình khi tách từ khá cao (khoảng 94,5%). Chúng tôi sử dụng mô hình túi từ (bag of words) để biểu diễn văn bản, chính nhờ việc phân biệt được từ đơn và từ ghép mà ngữ nghĩa của văn bản không mất đi hoàn toàn khi sử dụng mô hình này.

2) Loại bỏ từ dừng (stopwords): Stopwords trong lĩnh vực khoa học máy tính được định nghĩa là một tập hợp các từ xuất hiện rất phổ biến trong văn bản nhưng lại không cần thiết cho phân tích ngôn ngữ học, hoặc là xuất hiện rất ít lần trong tập ngữ liệu nên cũng không đóng góp nhiều về mặt ý nghĩa. Vì là các từ không mang nhiều ý nghĩa nên có thể loại bỏ khỏi văn bản một cách an toàn. Một nguyên nhân cần loại bỏ các từ có tần suất xuất hiện cao nhưng lại không mang nhiều ý nghĩa là vì sự tồn tại của các từ này có thể làm sai lệch kết quả khi phương pháp chúng tôi đề xuất có dựa trên việc phân tích tần suất của từ. Ví dụ các từ như "như vậy", "sau đó", "một số", "chỉ", … là những từ sẽ được loại bỏ, chẳng những không ảnh hưởng đến kết quả cuối cùng mà còn có thể tăng độ chính xác. Chúng tôi sử dụng danh sách gồm 570 stopwords, được đề xuất bởi [13].

B.

Tính độ tương tự

Trong nghiên cứu của chúng tôi, văn bản được biểu diễn bằng đồ thị. Mỗi đỉnh trong đồ thị tương ứng với một câu trong văn bản, mỗi cạnh nối hai đỉnh trong đồ thị biểu diễn mối liên hệ giữa hai câu. Trọng số của mỗi cạnh chính là giá trị độ tương tự (value of similarity) giữa hai câu. Độ tương tự (trọng số của cạnh) được tính bằng một trong ba phương pháp: Jaro, Constrast Model và Jaccard.

1) Khoảng cách Jaro [14]: là một độ đo tương tự giữa hai chuỗi. Khoảng cách Jaro d_j của giữa câu s_1 và câu s_2 được tính như sau:

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \quad (1)$$

trong đó m là số từ giống nhau, t là 1/2 số bước chuyển (transpositions).

Phép chuyển vị trí sẽ được thực hiện khi hai từ giống nhau trong hai câu s_1 và s_2 có khoảng cách không lớn hơn giá trị:

$$\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1 \quad (2)$$

Mỗi từ trong câu s_1 được so sánh với tất cả các từ trong câu s_2. Transpositions được định nghĩa là số lượng từ giống nhau giữa hai câu (nhưng thứ tự trong chuỗi khác nhau) chia cho 2.

2) Mô hình tương phản (Contrast model): Chúng tôi sử dụng mô hình tương phản của Tversky [15] để tính độ tương tự.

$$s(A, B) = \alpha*g(A \cap B) - \beta*g(A-B) - \gamma*g(B-A) \quad (3)$$

Công thức ở trên có thể được sử dụng để tính độ tương tự giữa hai câu A và B. Trong đó g(A∩B) biểu diễn cho các từ chung giữa A và B, g(A-B) biểu diễn cho các từ riêng của A, g(B-A) biểu diễn cho các từ riêng của B. α, β, γ trọng số được xác định trong quá trình thử nghiệm thuật toán.

3) Hệ số Jaccard: Hệ số tương tự Jaccard [16] là một độ đo tương tự của các tập hợp dựa trên phương pháp thống kê. Chúng tôi sử dụng hệ số này để đo độ tương tự giữa hai câu A và B như sau:

$$s(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

C.

Tính điểm xếp hạng

Chúng tôi sử dụng thuật toán thuật toán PageRank [17] để tính điểm xếp hạng (độ quan trọng) các đỉnh trong đồ thị. Tuy nhiên, thuật toán PageRank gốc được áp dụng trên đồ thị có hướng được chúng tôi hiệu chỉnh để có thể áp dụng trên đồ thị vô hướng. Thuật toán PageRank sẽ được áp dụng

trên đồ thị vô hướng có trọng số biểu diễn văn bản, trong đó trọng số của cạnh nối các đỉnh là độ tương tự của hai câu được biểu diễn bởi hai đỉnh tương ứng. Thuật toán xếp hạng PageRank thực hiện các lần lặp để cập nhật giá trị xếp hạng cho các đỉnh trong đồ thị. Quá trình lặp sẽ kết thúc khi lỗi hội tụ đạt dưới ngưỡng định trước (STANDARD_ERROR_THRESHOLD) hoặc là số lần lặp đã vượt quá giá trị định sẵn (tỷ lệ với số đỉnh của đồ thị). Giá trị lỗi được tính là độ lệch chuẩn của các giá trị xếp hạng mới và cũ của các đỉnh trong đồ thị. Ngoài việc phụ thuộc vào số lượng các cạnh vào và cạnh ra của các đỉnh trong đồ thị đã được xây dựng ở thành phần tính độ tương tự, do đây là đồ thị có trọng số nên trọng số cạnh cũng sẽ được tích hợp vào mô hình tính điểm xếp hạng của PageRank như sau (trong đó W_{ATi} là trọng số cung nối đỉnh A và đỉnh Ti):

$$PR(A) = 0.25 + 0.85 * (W_{AT1} * PR(T_1)/C(T_1) + \ldots + W_{ATn} * PR(T_n)/C(T_n)) \quad (5)$$

THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Phương pháp tóm tắt mà chúng tôi đề xuất trong nghiên cứu này là rút trích các câu quan trọng nhất trong văn bản để đưa vào tóm tắt. Khi đã xác định được danh sách các câu quan trọng nhất (mặc định là 25% số câu của văn bản), chúng tôi sẽ thực hiện sắp xếp các câu này theo thứ tự xuất hiện trong văn bản để có được tóm tắt của văn bản.

Để đánh giá độ tốt của giải pháp đề xuất, chúng tôi đã thực hiện đánh giá theo hai cách: 1- Thu thập các đoạn văn bản thô thuộc nhiều chủ đề khác nhau. Chọn lựa cộng tác viên tham gia tóm tắt các văn bản đã được thu thập ở bước trước, so sánh kết quả tóm tắt bởi các cộng tác viên và của hệ thống; 2- Thu thập các bài viết trên các trang báo điện tử theo tiêu chí các bài viết này phải được tóm tắt theo cách sử dụng các câu văn trong nội dung văn bản. Thực hiện đối chiếu tóm tắt của hệ thống với tòm tắt của văn bản thu thập.

D. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm dùng cho phương pháp đánh giá thứ nhất là 5 đoạn văn bản có độ dài khác nhau và thuộc các chủ đề khác nhau. Do cần nhờ đến các cộng tác viên thực hiện tóm tắt các đoạn văn bản để so khớp nên số lượng mẫu thực nghiệm cho phương pháp này là nhỏ. Chủ đề và số lượng câu của mỗi mẫu kiểm thử được cho trong Bảng 1.

BẢNG 1. DỮ LIỆU THỰC NGHIỆM CHO PHƯƠNG PHÁP 1

Tên văn bản	Chủ đề	Tổng số câu
Nhiều chuyên gia muốn Apple thu hồi Iphone4	Công nghệ	27
5 người mất tích trong bảo đã được tìm thấy	Xã hội	30
Barca tăng cường chất thép cho cánh trái	Thể thao	18
Dự án LMF	Kỹ thuật	18
Lão ngư dân và biển cả	Văn học nghệ thuật	78

Dữ liệu thực nghiệm cho phương pháp thứ hai là 25 bài viết được thu thập từ các trang báo mạng như Vietnamnet.vn và vnexpress.net. Do các bài viết phải đáp ứng yêu cầu là có tóm tắt được rút trích từ nội dung của bài viết nên thực tế số lượng cũng không nhiều và không phong phú về chủ đề. Đa số các bài viết được thu thập thuộc chuyên mục "*Tâm sự*" và "*Bạn đọc*" của hai tờ báo điện tử trên.

E. Phương pháp đánh giá

1) Cách 1

Dữ liệu dùng để đánh giá hiệu quả chương trình trong cách 1 này là các bản tóm tắt được thực hiện thủ công do các nhà khoa học thực hiện trên 5 văn bản dùng để thực nghiệm như đã đề cập ở mục A của phần III (Bảng 1). Mặc dù kết quả tóm tắt từ mỗi nhà khoa học có độ tin cậy khá cao, tuy nhiên để đảm bảo tính khách quan của kết quả tóm tắt, chúng tôi tiến hành thu thập tóm tắt từ 12 nhà khoa học (Bảng 2) khác nhau và việc tóm tắt được thực hiện độc lập.

BẢNG 2. CÁC NHÀ KHOA HỌC THAM GIA ĐÁNH GIÁ HỆ THỐNG

Stt	Họ tên	Email
1.	GS.TS. Võ Thị Gương	vtguong@ctu.edu.vn
2.	PGS.TS. Nguyễn Minh Thủy	nmthuy@ctu.edu.vn
3.	PGS.TS. Lê Thị Mến	ltmen@ctu.edu.vn
4.	GS.TS. Nguyễn Văn Thu	nvthu@ctu.edu.vn
5.	TS. Nguyễn Thị Hồng Nhân	nthnhan@ctu.edu.vn
6.	TS. Nguyễn Thị Thu Nga	nttnga@ctu.edu.vn
7.	TS. Lê Vĩnh Thúc	lvthuc@ctu.edu.vn
8.	ThS. Trương Xuân Việt	txviet@ctu.edu.vn
9.	ThS. Nguyễn Văn Ây	nvay@ctu.edu.vn
10.	ThS. Nguyễn Thu Tâm	nttamty@ctu.edu.vn
11.	ThS. Lê Minh Lý	lmly@ctu.edu.vn
12.	ThS. Phạm Thị Phương Thảo	ptpthao@ctu.edu.vn

Độ chính xác của kết quả tóm tắt được định nghĩa như sau: (số lượng câu trùng lắp giữa kết quả thuật toán và kết quả chuyên gia) / (số lượng câu tóm tắt cần chọn). Chúng tôi đề xuất phương pháp đo như sau: sử dụng phương pháp bầu chọn (voting) để chọn ra một chuẩn vàng (gold-standard). Gold-standard là một tập hợp gồm các câu nằm trong tóm tắt được nhiều người bầu chọn nhất. Gọi result (i) là kết quả tóm tắt văn bản thứ i, công thức để tính độ chính xác (precision) của mỗi phương pháp áp dụng trên văn bản thứ i như sau:

$$Precision(i) = \frac{|result(i) \cap gold-standard(i)|}{|gold-standard(i)|} \quad (6)$$

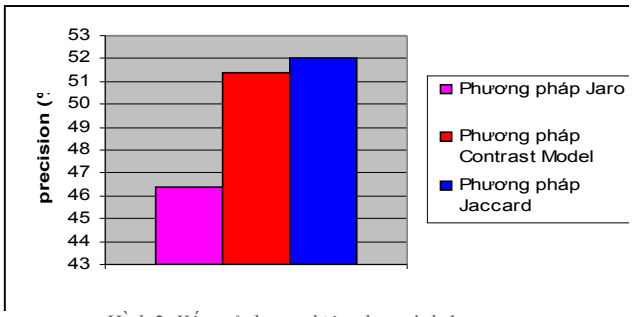
Tóm tắt của các nhà khoa học không phải lúc nào cũng trùng khớp với nhau, vì thế chúng tôi đề xuất sẽ lựa chọn các câu nào được nhiều nhà khoa học chọn nhất sẽ được đưa vào tóm tắt và được xem như là tóm tắt của các nhà khoa học. Tỷ lệ thống nhất giữa các nhà khoa học cao nhất là 67% và thấp nhất là 55%.

Chúng tôi cho hệ thống thực hiện tóm tắt trên 3 độ đo đã giới thiệu ở mục B phần II. Giá trị các tham số sử dụng cho từng độ đo được cho trong bảng 3.

BẢNG 3. GIÁ TRỊ THAM SỐ THỰC NGHIỆM

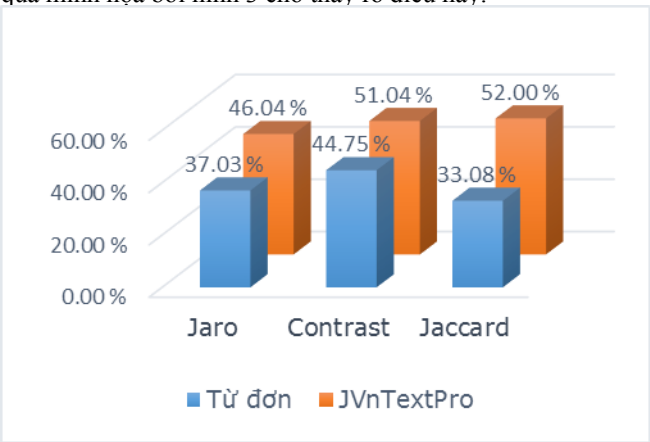
Tên phương pháp	Thuật toán tính độ tương tự	Threshold (xây dựng cạnh nối giữa các đỉnh)
Phương pháp Jaro	Jaro	0.65
Phương pháp Contrast Model	Contrast Model	5.0
Phương pháp Jaccard	Jaccard	0.25

Thực hiện so sánh kết quả đạt được khi sử dụng các độ đo khác nhau, chúng tôi có thể kết luận rằng độ đo Jaccard cho kết quả tốt nhất nhưng không khác biệt nhiều so với mô hình tương phản (contrast model), xem hình 2.



Hình 2. Kết quả thực nghiệm theo cách 1

Chúng tôi cũng thực nghiệm ảnh hưởng của quá trình tiền xử lý đối với phương pháp đề xuất. Thật vậy, để có thể tính toán chính xác độ tương đồng giữa các câu đòi hỏi quá trình tách từ phải có khả năng nhận biết đúng các từ được sử dụng trong ngữ cảnh của câu. Có nghĩa là cần phân biệt được từ đơn và từ ghép. Vì bản chất tiếng việt có nhiều từ ghép nên không thể đơn giản sử dụng khoảng trắng để tách từ, kết quả minh họa bởi hình 3 cho thấy rõ điều này.



Hình 3 Kết quả thực nghiệm khi có sử dụng và không sử dụng JVnTextPro

So sánh với các hệ thống đã có trên 5 văn bản thực nghiệm cũng cho thấy hệ thống chúng tôi xây dựng cho độ chính xác cao hơn (Bảng 4).

TextRank áp dụng cho tiếng Việt: TextRank là kết quả nghiên cứu được đề xuất bởi [18] áp dụng cho văn bản tiếng Anh. Thực nghiệm tóm tắt tương tự như cách thực nghiệm đã áp dụng cho hệ thống do chúng tôi đề xuất.

AutoSummarize (Microsoft Word 2003): Thực nghiệm tóm tắt tương tự như cách thực nghiệm đã áp dụng cho hệ thống do chúng tôi đề xuất.

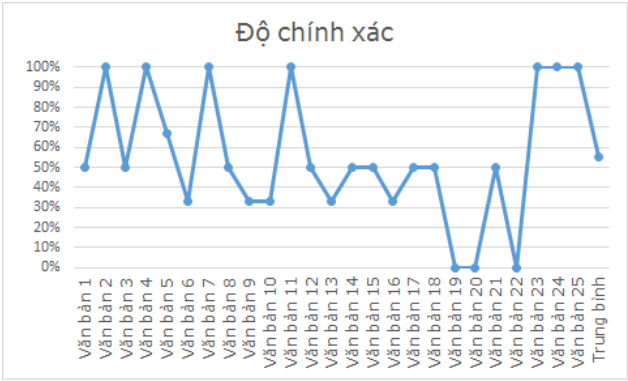
BẢNG 4. SO SÁNH KẾT QUẢ CỦA HỆ THỐNG ĐỀ XUẤT VỚI CÁC HỆ THỐNG KHÁC

Phương pháp	Tên phương pháp	Độ chính xác (%)
Đề xuất	Phương pháp Jaro	46.4
	Phương pháp Contrast Model	51.4
	Phương pháp Jaccard	52.0
Có sẵn	TextRank áp dụng cho tiếng Việt	33.2
	Microsoft Word 2003	12.4

2) Cách 2

Chúng tôi thu thập 25 bài viết trên 2 trang báo điện tử vietnamnet và vnexpress theo điều kiện các bài viết cần có tóm tắt theo kiểu rút trích nguyên văn một số câu từ nội dung của bài viết. Chúng tôi cũng đã lựa chọn các bài viết có số lượng câu tóm tắt là khá ít, dao động trong khoảng từ 1 đến 3 câu. Kết quả thực nghiệm theo cách 1 cho thấy độ đo Jaccard có kết quả tốt hơn cả nên ở cách 2 này chúng tôi chỉ thực nghiệm với độ đo Jaccard. Hình 4 minh họa độ chính xác của phương pháp tóm tắt đối với từng văn bản cũng như độ chính xác trung bình trên tập 25 văn bản.

Phân tích kết quả đạt được chúng tôi nhận thấy có 7 văn bản có kết quả tóm tắt trùng khớp 100%, phần nhiều vẫn là trùng khớp với tỷ lệ 50%, tuy nhiên vẫn còn có một số văn bản tỷ lệ trùng khớp là 0%. Tỷ lệ trùng khớp trung bình trên tập 25 văn bản là 55.3%, tỷ lệ này cũng gần với tỷ lệ thực nghiệm ở cách 1.



Hình 4. Kết quả thực nghiệm theo cách 2

KẾT LUẬN VÀ KIẾN NGHỊ

Trong bài báo này, chúng tôi giới thiệu phương pháp trích chọn tóm tắt từ nội dung văn bản theo hướng tiếp cận sử dụng cấu trúc đồ thị để biểu diễn văn bản, đây là hướng tiếp cận mới của thế giới trong những năm gần đây. Chúng tôi cũng đề xuất ứng dụng các độ khác nhau để tính độ tương tự câu trong hệ thống tóm tắt văn bản. Trong đó: 1- đây là công trình nghiên cứu lần đầu tiên tại Việt Nam sử dụng 3 thuật toán Jaro, Contrast Model và Jaccard vào công việc tóm tắt văn bản và cho kết quả khả quan; 2- đây cũng là công trình nghiên cứu đầu tiên trên thế giới tích hợp thuật toán Contrast Model vào hệ thống tóm tắt văn bản, thuật toán này thể hiện độ chính xác cao trên tập dữ liệu nghiên cứu. Kết quả thực nghiệm (ngay cả khi tập dữ liệu kiểm thử có kích thước nhỏ) đã chứng minh phần nào tính khả thi trong việc ứng dụng kết quả nghiên cứu vào thực tiễn.

Kết quả khả quan của phương pháp đề xuất có thể lý giải từ nhiều nguyên nhân: 1- Sử dụng được ưu điểm của phương pháp chỉ mục từ tiếng Việt do công cụ JVnTextPro cung cấp.

Thật vậy, trong nghiên cứu của mình, chúng tôi dựa trên hướng tiếp cận "mô hình túi từ - bag of words" để biểu diễn nội dung văn bản, phương pháp này có ưu điểm là cài đặt đơn giản nhưng có hạn chế lớn là làm mất đi ngữ nghĩa của văn bản vì không quan tâm đến vị trí của từ mà chỉ quan tâm đến tần suất xuất hiện của từ. Vì sử dụng công cụ JVnTextPro có khả năng nhận biết chính xác từ đơn và từ ghép nên ngữ nghĩa của văn bản phần nào được giữ lại so với việc xem nội dung văn bản là tập hợp các từ đơn (từ gồm 1 chữ); 2- Thuật toán PageRank dùng để xếp hạng các trang web đã chứng tỏ được tính khả thi khi được ứng dụng thành công trong các bộ máy tìm kiếm thông tin web. Khi được ứng dụng vào ngữ cảnh này, PageRank tỏ ra hiệu quả ngay cả khi đồ thị web là một đồ thị không có trọng số. Vì thế chúng tôi tin rằng sự kết hợp thuật toán xếp hạng PageRank với các độ đo tương tự (gán trọng số cho cạnh) sẽ mang lại kết quả khả quan và kết quả thực nghiệm đã phần nào chứng minh nhận xét trên khi mà độ đo Jaccard và độ đo Contrast Model đã cho kết quả tóm tắt vượt trên các hệ thống sẵn có, đặc biệt là khi so sánh với phương pháp có hướng tiếp cận tương tự là TextRank.

Một ưu điểm khác của phương pháp chúng tôi đề xuất là quá trình tóm tắt không cần tập ngữ liệu huấn luyện, cũng như không cần xem xét tính ngữ nghĩa và cấu trúc ngữ pháp của câu và việc tóm tắt được áp dụng trên từng văn bản đơn.

Tuy kết quả đạt được bước đầu là rất khả quan nhưng để có thể khẳng định chắc chắn hơn tính khả thi của giải pháp chúng tôi cần thêm thời gian thu thập dữ liệu thực nghiệm cũng như cần thêm thời gian và sự đóng góp của bạn bè đồng nghiệp trong việc trợ giúp thực hiện tóm tát các đoạn văn bản như là một kênh thông tin so khớp với kết quả của phương pháp. Chúng tôi cũng đề xuất áp dụng giải pháp tóm tắt văn bản tự động như là một công đoạn của phân nhóm tài liệu. Thay vì phân nhóm văn bản dựa trên toàn bộ nội dung của nó thì ta có thể phân nhóm dựa vào tóm tắt của nó, và nếu giải pháp này thành công thì sẽ giúp tăng đáng kể tốc độ của các ứng dụng phân nhóm văn bản theo chủ đề.

TÀI LIỆU THAM KHẢO

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, UstavInformatiky a softveroveho inzinierstva, 2008.

[2] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.

[3] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, pp. 347-352.

[4] Vishal Gupta, Gurpreet Singh Lehal (2010), "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, Vol 2, No 3 (2010), 258-268.

[5] Mohamed Abdel Fattah, Fuji Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization", Computer Speech & Language 23(1): 126-144 (2009).

[6] H. P. Edmundson, "New Methods in Automatic Extracting", J. ACM 16(2): 264-285 (1969).

[7] Daniel Marcu, "The Theory and Practice of Discourse Parsing and Summarization", A Bradford Book, MIT Press, Cambridge, Massachusetts, 2000.

[8] Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam , "Centroid-based summarization of multiple documents". Information Processing and Management, vol. 40, issue 6, pp. 919-938, 2004.

[9] 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.

[10] Nguyen, L.M., Shimazu, A., Ho, T.B., Phan, X.H., Horiguchi, S., "Sentence extraction with support vector machine ensemble", First World Congress of the International Federation for Systems Research (IFSR'05), Symposium on Data/Text Mining from Large Databases, Kobe, 15-17 November, S5-2-4, 2005.

[11] 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.

[12] Nguyen Trong Phuc, Le Thanh Huong, "Vietnamese text summarisation using discourse structures", The ICT.rda conference, Hanoi, Vietnam, 2008.

[13] Nguyen Cam Tu, "JvnTextPro: A Java-based VietNamese Text Processing Toolkit".

[14] Winkler, W. E., "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". Proceedings of the Section on Survey Research Methods (American Statistical Association): 354–359, 1990.

[15] Tversky, A., "Features of similarity", Psychological Review, 84, 327-352, 1977.

[16] Paul Jaccard, "Etude comparative de la distribution orale dans une portion des Alpes et des Jura". In Bulletin del la Socit Vaudoise des Sciences Naturelles, volume 37, pages 547-579.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", 1999.

[18] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479, 2004.