RÚT TRÍCH VÀ TÓM TẮT NỘI DUNG TRANG WEB TIẾNG VIỆT

Đỗ Phúc, Hồ Anh Thư
Trung tâm Phát triển Công nghệ Thông tin – Đại học Quốc gia Tp. HCM
(Bài nhận ngày 25 tháng 08 năm 2005, hoàn chỉnh sửa chữa ngày 31 tháng 10 năm 2005)

TÓM TẮT: *Tóm tắt văn bản, hay tóm tắt tự động văn bản, là kỹ thuật tự động tạo ra một bản tóm lược hay tóm tắt của một hay nhiều văn bản. Kỹ thuật này đã được phát triển trong nhiều năm khởi đầu là công trình của Luhn năm 1958. Những năm gần đây, sự gia tăng sử dụng Internet đã thu hút sự quan tâm về các kỹ thuật tóm tắt văn bản. Bài báo trình bày các bước xây dựng hệ thống tạo tóm tắt các trang Web tiếng Việt dựa trên việc trích các câu từ tài liệu gốc để tạo tóm tắt. Hệ thống được phát triển theo hướng thống kê kết hợp với việc áp dụng các kết quả xử lý ngôn ngữ tự nhiên đã có (như tách từ, gán nhãn từ loại) nhằm tăng hiệu quả của các bản tóm tắt và mở ra triển vọng giải quyết bài toán ngữ nghĩa trong tạo tóm tắt văn bản tiếng Việt.*

1. GIỚI THIỆU

Tóm tắt là qui trình làm cô đọng một văn bản nguồn thành một phiên bản rút ngắn mà vẫn giữ được các thông tin chính của văn bản đó [7]. Đây là một công việc khó, đặc biệt đối với xử lý bằng máy. "Hầu hết các nghiên cứu về tóm tắt văn bản đều chuyển dịch bài toán này thành bài toán "tóm tắt bằng cách rút trích các đoạn văn bản": xếp hạng các câu trong văn bản dựa trên độ trội của câu" [10]. Hướng nghiên cứu để tạo ra bản tóm tắt thật sự, như của người thực hiện, được gọi là *abstract*, ta gọi là *tóm lược*, và hướng nghiên cứu sau được gọi là *extract*, ta gọi là *trích lược*. Đối với việc tạo tóm tắt theo hướng *trích lược*, bản tóm tắt được tạo ra là tập hợp của các phần văn bản được trích nguyên văn từ văn bản gốc. Một lĩnh vực con của tóm tắt văn bản là tóm tắt trang Web. Trong lĩnh vực này đã có phát triển một hướng tóm tắt mới phù hợp với đặc trưng của các tài liệu Web là tóm tắt dựa vào *ngữ cảnh* của trang Web, bao gồm thông tin trên chính trang Web và các thông tin mô tả từ bên ngoài của các trang Web khác dành cho trang Web đang xét, xem [8]. Bên cạnh việc tóm tắt trang Web còn có nhu cầu tóm tắt cả *Website* gồm nhiều trang Web. Hiện nay, các nghiên cứu về tóm tắt đều theo hướng sử dụng xử lý ngôn ngữ tự nhiên để làm tăng hiệu quả tóm tắt [6],[7]. Trong nước, cũng có đề tài bắt đầu nghiên cứu về tóm tắt văn bản tiếng Việt [4], tuy vậy công trình này chưa áp dụng các phương pháp xử lý ngôn ngữ tự nhiên, để rút trích từ đặc trưng văn bản mà chỉ dựa trên việc thống kê các dãy từ phổ biến. Đối với lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt, cho đến nay đã có một số nghiên cứu cho kết quả tương đối và có thể áp dụng trong việc tóm tắt văn bản như tách từ, gán nhãn từ loại, phát hiện cụm danh từ ... tiếng Việt [1],[3],[9],[15]. Việc sử dụng các kết quả xử lý ngôn ngữ tự nhiên mở ra triển vọng giải quyết bài toán ngữ nghĩa trong tạo tóm tắt văn bản tiếng Việt.

2. TÓM TẮT VĂN BẢN TỰ ĐỘNG

2.1. Các yêu cầu đối với tóm tắt văn bản

Nhìn chung, các bản tóm tắt văn bản cần đảm bảo các yêu cầu:

- Cung cấp đủ thông tin chính yếu của văn bản.
- Có tính bao quát toàn bộ nội dung chính của văn bản
- Độ nén của bản tóm tắt. Việc đảm bảo độ nén được xét đến qua hai cấp: rút ngắn số câu, và rút ngắn câu.
- Tính mạch lạc, tính chặt chẽ.

2.2. Quy trình tóm tắt bao gồm các bước sau

- **Bước 1.** Tiền xử lý tài liệu, loại bỏ các phần không cần thiết trong tài liệu.
- **Bước 2.** Tách đoạn, tách câu.

- **Bước 3.** Xử lý ngôn ngữ tự nhiên bao gồm những việc được liệt kê bên dưới.

+ **Xử lý 1.** Tách từ. Một ứng dụng về xử lý trên văn bản nếu muốn xử lý sâu hơn về mặt ngôn ngữ như ngữ nghĩa của từ, cấu trúc văn phạm, phân biệt các thành phần trong câu, ... thì công việc bắt buộc phải làm đầu tiên là tách từ. Để phục vụ cho bước tách từ tự động tiếng Việt trong hệ thống, chúng tôi sử dụng chương trình theo [9].

+ **Xử lý 2.** Gán nhãn từ loại. Kết quả của công đoạn này là mỗi từ trong câu sẽ được gán cho một nhãn từ loại như danh từ, động từ, tính từ, trạng từ, ... Chúng tôi sử dụng phần mềm VnQTag [3] để gán nhãn từ loại tự động cho văn bản.

+ **Xử lý 3.** Phát hiện các cụm từ (cụm danh từ, cụm động từ, cụm tính từ, ...). Do tính chất của cụm từ, cụm danh từ sẽ miêu tả rõ hơn và xác định rõ đối tượng hơn danh từ, nên nếu ta dùng cụm danh từ để đặc trưng câu, hay biểu diễn câu, thì các đối tượng được thông báo trong câu sẽ được trích ra chính xác hơn.

+ **Xử lý 4.** Xét quan hệ ngữ nghĩa giữa các từ. Việc xét quan hệ ngữ nghĩa giữa các từ giúp tăng hiệu quả so sánh độ tương tự giữa các câu. Thông thường, để đánh giá quan hệ ngữ nghĩa giữa các từ người ta dùng một từ điển ngữ nghĩa có sẵn, chẳng hạn, trong tiếng Anh có từ điển Wordnet trong đó gom các từ đồng nghĩa theo từng nhóm và có phân cấp các nhóm theo quan hệ ngữ nghĩa cha-con. Hiện nay đã có các nỗ lực xây dựng một tự điển Wordnet như vậy cho tiếng Việt.

– **Bước 4.** Biểu diễn tài liệu. Trong giới hạn của việc tóm tắt dựa trên việc trích câu, một tài liệu có thể được xem là một tập các câu, biểu diễn tài liệu là biểu diễn các câu trong tập đó. Các câu sau bước xử lý ngôn ngữ tự nhiên sẽ được chia nhỏ và bao gồm các đơn vị nhỏ hơn, là từ hoặc ngữ, mỗi đơn vị như vậy được gọi là mục từ (term). Sau đó sẽ sử dụng chúng để tạo mô hình biểu diễn văn bản.

– **Bước 5.** Rút các câu trội trong văn bản. Việc rút câu trội được dựa trên nguyên lý: các câu nổi trội là các câu chứa các thông tin được nhắc lại nhiều lần trong văn bản, hay là thông tin quan trọng trong văn bản. Do vậy, các câu trội sẽ được chọn ra từ các nhóm câu tương tự có số lượng câu trong nhóm lớn nhất. Ngoài ra, có những nghiên cứu sử dụng các heuristic, hoặc phương pháp học để rút ra các qui luật chọn câu quan trong dựa trên cấu trúc văn bản [19].

– **Bước 6.** Tạo tóm tắt. Bước này sử dụng các câu trội đã chọn được ở bước trước để tạo nên bản tóm tắt. Trong đề tài này, tác giả chỉ đơn thuần ghép các câu trội trích được để tạo thành bản tóm tắt, nhưng có giữ lại đúng thứ tự các câu trong văn bản.

3. TÓM TẮT TÀI LIỆU WEB

Tóm tắt tài liệu Web thực chất là bài toán con phát triển từ bài toán tóm tắt văn bản. Các tài liệu Web sau bước tiền xử lý để trích lấy thông tin cũng sẽ được tạo tóm tắt theo trình tự tương tự như với tóm tắt văn bản thông thường. Tuy nhiên, do đặc điểm của các tài liệu Web, ta cần có quá trình tiền xử lý phù hợp để trích được đúng thông tin làm dữ liệu cho quá trình tóm tắt. Hơn nữa, trong một số trường hợp, nhằm tăng hiệu quả của hệ thống, việc tiền xử lý này không bắt buộc phải được thực hiện trước khi xử lý tạo tóm tắt, mà có thể được thực hiện trực tiếp ngay trong lúc xử lý tạo tóm tắt, như trong đề tài này.Việc tách lấy nội dung của trang Web bao gồm hai việc: tách lấy thông tin dạng văn bản của trang Web và chọn vùng văn bản chính của trang Web.

3.1.Tách lấy thông tin dạng văn bản

Việc tách lấy thông tin dạng văn bản đồng nghĩa với việc loại bỏ các thông tin định dạng, thông tin xử lý động của trang Web. Các trang Web hiện nay được thiết kế theo chuẩn HTML bao gồm các thẻ (*tag*) định dạng cho các thành phần nội dung trong trang Web. Các thông tin định dạng, xử lý, các hình ảnh, ... cần được loại bỏ, chỉ giữ lại những phần thông tin bằng lời mà trang Web muốn thông báo cho người xem.

Yêu cầu trên có thể được thực hiện nhờ trên trang Web, tất cả những thông tin không phải văn bản và không cần hiển thị trên bề mặt trang Web đều được bao trong hai dấu "<" và

">", một chuỗi "<.....>" được gọi là *HTML tag* (hay *thẻ đánh dấu trang HTML*), trong phần này ta gọi tắt là *tag*. Các *tag* luôn bắt đầu bằng "tên tag" tương ứng với chức năng xử lý của tag đó trên trang Web, theo sau đó là các thuộc tính thêm cho *tag* đó, nếu có.

3.2.Chọn vùng văn bản chính

Trong trang web, ngoài thông tin chính của trang web, thường chứa nhiều thông tin phụ khác như: thông tin quảng cáo, thông báo phụ, các đề mục, menu, ... Mục tiêu của bước này là xác định một vùng văn bản chính của trang Web nhằm giảm nhẹ phần dữ liệu không cần thiết phải xử lý trong phần gom cụm và rút câu trội tạo tóm tắt. Việc xác định vùng văn bản chính bao gồm hai bước:

- Phân chia các vùng văn bản dựa vào cấu trúc trang Web. Trong trang Web, các *tag* thường xuyên được dùng để phân tách vùng văn bản khác nhau là *<TABLE>*(kẽ bảng) và *<HR>* (lằn phân cách). Do đó, ở đây ta sử dụng các dấu hiệu này để xác định các ranh giới giữa các vùng văn bản khác nhau.
- Phân chia lại các vùng văn bản dựa vào nội dung. Việc phân vùng dựa vào các *tag HTML* như trên vẫn có thể không đúng do các trường hợp thiết kế khác lạ. Sự phân vùng không chính xác ở bước trên có thể được khắc phục bằng việc xét đến nội dung của các vùng văn bản. Những vùng có nội dung tương tự nhau (ngưỡng để được xem là tương tự được cho trước) sẽ được ghép lại thành một vùng văn bản lớn hơn. Việc so sánh độ tương tự của hai vùng văn bản thực chất là so sánh độ tương tự của hai nhóm câu, nên cách xử lý sẽ giống như xử lý tính độ tương tự của hai nhóm câu trong bước rút câu trội. Do có so sánh độ tương tự giữa các câu nên bước này chỉ được thực hiện sau bước biểu diễn câu. Sau khi phân chia tài liệu thành các vùng văn bản có nội dung không liên quan nhau, ta chọn vùng văn bản có kích thước lớn nhất là vùng văn bản chính để xử lý tạo tóm tắt. Từ đây, việc xử lý tóm tắt được thực hiện giống như xử lý cho tài liệu thuần văn bản thông thường.

3.3. Qui trình tóm tắt trang Web

Qui trình xử lý của chương trình tạo tóm tắt cho trang Web tiếng Việt như sau:

Bước 1. Tiền xử lý trang Web. Bao gồm:

- ***Bước 1.1. Trích thông tin dạng văn bản.*** Tách và loại bỏ các thông tin định dạng, thông tin xử lý của trang Web, các hình ảnh, ..., chỉ giữ lại phần thông tin dạng văn bản của trang Web.
- ***Bước 1.2. Phân vùng văn bản tạm thời.*** Phân chia tạm thời các vùng nội dung trên trang Web dựa vào độ gần về không gian của các vùng văn bản xuất hiện trong trang Web.

Bước 2. Tách câu. Phân tách các câu trong các phần văn bản đã có được qua bước 1.

Bước 3. Tách từ. Phân tách các từ trên mỗi câu đã phân tách ở bước 2.

Bước 4. Gán nhãn từ loại. Gán nhãn từ loại cho từng từ trên các câu.

Bước 5. Đặc trưng câu. Bao gồm:

- ***Bước 5.1.*** Loại bỏ những câu không hợp lệ (không phải là câu thật sự)
- ***Bước 5.2.*** Biểu diễn các câu trên không gian vec-tơ.

Bước 6. Tạo tóm tắt. Bao gồm:

- ***Bước 6.1. Xác định vùng văn bản chính.*** Dựa vào thông tin phân vùng văn bản tạm thời có được ở Bước 1, ghép các vùng có độ tương tự cao và chọn vùng văn bản có kích thước lớn nhất để xử lý trong các bước tiếp theo.
- ***Bước 6.2. Gom cụm câu.*** Gom các câu tương tự nhau thành từng nhóm
- ***Bước 6.3 Rút câu trội.*** Với mỗi nhóm câu, chọn ra câu trội nhất để tạo bản tóm tắt, số lượng câu được chọn để tạo tóm tắt tỉ lệ với tổng số câu của các nhóm.

4. RÚT CÂU TRỘI

4.1. Biểu diễn câu vào không gian vec-tơ

Trong bước này, chương trình sử dụng mô hình vec-tơ và độ đo TF để biểu diễn các câu. Từng câu trong tài liệu sẽ được biểu diễn vào trong vec-tơ không gian, mỗi thành phần trong vec-tơ chỉ đến 1 từ tương ứng trong danh sách mục từ chính. Vec-tơ không gian là vec-tơ có kích thước bằng số mục từ trong danh sách chính. Mỗi phần tử là độ quan trọng của mục từ tương ứng ở trong câu. Độ quan trọng của 1 mục từ *i* trong câu *j* được tính bằng độ đo TF, như sau:

$$w_{i,j} = \frac{tf_{i,j}}{\sqrt{\sum_j tf_{i,j}^2}}$$

Trong đó tf_{ij} là tần số xuất hiện của mục từ *i* trong câu *j*.

4.2. Gom cụm các câu

Chúng tôi chọn phương pháp gom cụm phân cấp để gom cụm các câu tương tự nhau lại. Trước tiên, cần xác định hàm đo độ tương tự để đo độ tương tự giữa từng cặp câu trong tài liệu.

Xét độ tương tự giữa hai câu:

Với không gian biểu diễn tài liệu được chọn là không gian vectơ và trọng số TF, độ đo tương tự được chọn là cosin của góc giữa hai vectơ tương ứng của hai câu cần xét: cho S_i và S_k là hai câu được biểu diễn bằng các vec-tơ $<w_1^i, .., w_t^i>$ và $<w_1^k, .., w_t^k>$ tương ứng, độ tương tự giữa chúng là:

$$Sim(S_i, S_k) = \frac{\sum_{j=1}^{t} w_j^i . w_j^k}{\sqrt{\sum_{j=1}^{t} \left(w_j^i\right)^2 \bullet \sum_{j=1}^{t} \left(w_j^k\right)^2}}$$

Trên các vectơ biểu diễn cho các câu lúc này ta chưa xét đến các quan hệ ngữ nghĩa giữa các mục từ, do đó các từ đồng nghĩa sẽ không được phát hiện, dẫn đến kết quả xét độ tương tự giữa các câu chưa tốt. Ta xét ví dụ, cho hai câu như sau:

S_1: *Nhân loại* đang đứng trước nguy cơ thiếu nhiên liệu.

S_2: Trong quá khứ *loài người* đã có những biến đổi lớn.

Nếu ta không xét đến quan hệ ngữ nghĩa giữa các từ thì hai câu trên không có liên hệ gì cả và độ tương tự bằng 0. Nhưng thực chất, ta thấy rằng, từ *"nhân loại"* và từ *"loài người"* là đồng nghĩa, hai câu trên đều nói về loài người, do đó giữa hai câu có một sự liên quan nhất định và với công thức tính độ tương tự như trên thì độ tương tự giữa hai câu này phải khác 0.

Việc xét quan hệ ngữ nghĩa giữa các từ giúp tăng hiệu quả so sánh độ tương tự giữa các câu. Trong bài báo này, để đánh giá quan hệ ngữ nghĩa giữa các từ chúng tôi dùng từ điển đồng nghĩa Longman tiếng Anh (LDOCE – The Longman Dictionary of Comtemporary English) được dịch sang tiếng Việt, trong đó bao gồm các lớp từ đồng nghĩa. Các từ đồng nghĩa được gom vào trong cùng lớp. Với việc dùng từ điển đồng nghĩa như trên, mỗi mục từ trong danh sách mục từ chung của tài liệu sẽ được xác định một lớp ngữ nghĩa tương ứng, và ta cũng xác định được những mục từ nào đồng nghĩa với nhau. Khi xét độ tương tự giữa hai câu, ta điều chỉnh trọng số cho các mục từ đồng nghĩa trước khi tính độ tương tự theo công thức trên.

Thuật toán gom cụm phân cấp:

Thuật toán thực hiện như sau, cho tập các câu trong tài liệu $S=\{S_i\}_{i=1..N}$:

Bước 1:Gán từng câu vào một cụm của chính nó và xác định độ tương tự giữa từng cặp cụm $\{S_i\}$ và $\{S_k\}$ bằng giá trị độ tương tự là $Sim(S_i, S_k)$.

Bước 2: Tìm cặp cụm gần nhất (có giá trị độ tương tự cao nhất) và trộn chúng lại thành một, do đó giảm bớt một cụm.

Bước 3:Tính độ tương tự giữa cụm mới với mỗi cụm cũ.

Bước 4: Lặp lại bước 2 và 3 cho đến khi tất cả cụm được trộn lại thành một cụm, hoặc giá trị độ tương tự của cặp tương tự nhất trong tập nhỏ hơn một ngưỡng $0 \le \alpha \le 1$ cho trước.

Có thể thực hiện bước 3 theo ba cách: có thể xem độ tương tự của hai tập là độ tương tự cao nhất của một cặp 2 thành phần được lấy từ mỗi cụm (đây gọi là khoảng cách kết nối đơn giản, SLD), hoặc ngược lại là giá trị độ tương tự nhỏ nhất (gọi là khoảng cách kết nối hoàn toàn, CLD), hoặc có thể lấy trung bình độ tương tự của tất cả thành phần của cụm này với tất cả thành phần của cụm kia (khoảng cách kết nối trung bình, ALD).

Sau khi gom cụm xong, ta loại bỏ những cụm chỉ có 1 phần tử. Xếp thứ tự ưu tiên các cụm dựa vào số lượng câu có trong mỗi cụm, cho $\{C_1, ., C_p\}$ là tập các cụm còn lại đã được xếp ưu tiên.

4.3. Rút câu trội

Trong phần này, ta rút các câu trội trong từng cụm để tạo tóm tắt.

Cách 1:Tính hàm f cho từng cụm C_i. Cách tính hàm f như bên dưới được tham khảo từ công trình của nhóm tác giả J.-Y. Delort[8].Hàm f sử dụng hai thành phần chiều dài của các câu và độ gần của các câu so với trọng tâm[1] của cụm để xếp độ ưu tiên của các câu trong cụm. Trước khi tính $f(S_i)$, các câu của các cụm được xếp thứ tự theo kích thước và độ gần của chúng so với trọng tâm của cụm của chúng trong hai bảng băm $R1$ và $R2$, trong đó $Rx[S_i]$ là thứ tự ưu tiên của câu S_i theo $x = 1, 2$. Sau đó, giá trị $f(.)$ của mỗi câu S_i được cho bởi công thức:

$$f(S_i) = R1[S_i] \times \sqrt{R2[S_i]}$$

Sau khi tính f cho từng cụm. Câu trội trong mỗi cụm là câu có f cao nhất.

Cách 2: Cách này sử dụng *qui tắc tăng cường lẫn nhau* (Mutual Reinforcement Principle) của tác giả Zha (2002). Lý luận của qui tắc này là: một mục từ sẽ có chỉ số trội cao nếu nó xuất hiện trong nhiều câu có độ trội cao, và ngược lại, một câu có độ trội cao nếu nó có chứa nhiều mục từ có độ trội cao. Ta sẽ áp dụng qui tắc này trong từng cụm để tìm câu trội nhất trong các cụm đó.

Trong phần biểu diễn các câu vào trong vec-tơ ta đã có mỗi câu là một vec-tơ trong đó các thành phần của vec-tơ là các trọng số của mỗi mục từ trong không gian vec-tơ trên câu đó. Như vậy tập hợp các vec-tơ này trong một cụm sẽ tạo thành một đồ thị liên kết hai thành phần của tập các câu và tập các mục từ G(T, S, W). Trong đó:

$T = \{t_1, \ldots, t_n\}$là tập các mục từ, tức là danh sách mục từ chính đã nói ở các bước trên.

S $\{s_1, \ldots, s_m\}$ là tập các câu trong cụm đang xét.

W là ma trận trọng số chỉ sự liên kết giữa các mục từ với các câu, giá trị của mỗi thành phần trong ma trận chính là trọng số $w_{i,j}$ của 1 mục từ i trên câu j.

Với mỗi mục từ t_i và mỗi câu s_j ta tính chỉ số trội của chúng tương ứng là $u(t_i)$ và $v(s_j)$.

Ta tập hợp các chỉ số trội của các mục từ và các câu tương ứng vào trong 2 vec-tơ u và v được cho như sau:

$$u = \frac{1}{\sigma} W v, \quad v = \frac{1}{\sigma} W^T u$$

Trong đó σ là hằng số tỷ lệ. Cách tính $\{u, v, \sigma)$ được cho như sau:

Chọn giá trị khởi đầu cho v là vec-tơ với tất cả giá trị thành phần là 1.

Lặp lại giữa 2 bước sau cho tới khi hội tụ (khi u và v không có thay đổi nữa)

1) Tính và chuẩn hóa u:

$$u = Wv, \quad u = u/\|u\|$$

2) Tính và chuẩn hóa v:

$$v = W^T u, \quad v = v/\|v\|$$

[1] Trọng tâm của cụm được cho bởi vec-tơ có các trọng số là các trọng số trung bình của tất cả các từ trong biểu diễn trên tất cả các câu trong cụm.

Tính $\sigma = u^T Wv$ khi đã hội tụ.

Dựa vào v ta biết được câu trội nhất trong cụm đang xét.

Sau khi đã tính độ trội của từng câu trong cụm như trên, ta gọi $f(S_i)$ là độ trội của câu S_i trong cụm của nó dựa trên thông số thống kê từ, và dùng chung trong cả hai cách 1 và 2 ở trên. Từ đây, ta có thể bổ sung thêm các thông số khác, chẳng hạn thông số ưu tiên câu mở đầu đoạn văn, ưu tiên câu có nhiều danh từ riêng, ... Ở đây ta bổ sung thêm thông số ưu tiên câu mở đầu đoạn.

4.4. Bổ sung tham số ưu tiên câu mở đầu đoạn vào hàm tính độ trội của câu

Theo tác giả Loxeva (1980), được nêu trong [2], và theo chứng minh của tác giả Zhang [19] thì đa số câu mở đầu đoạn là câu có thể cho nội dung cơ bản của cả đoạn. Do vậy, ở đây ta bổ sung thêm thông số ưu tiên câu mở đầu đoạn để tăng hiệu quả tóm tắt.

Ta gọi $f_1(S_i) = h(S_i).f(S_i)$ là hàm tính độ trội của câu trong cụm sau khi bổ sung thêm thông số ưu tiên câu mở đầu đoạn, trong đó $h(S_i)$ là hàm tính độ trội của câu theo thông số ưu tiên câu mở đầu đoạn.Do vẫn còn những trường hợp khác mà câu đầu đoạn không phải là câu chủ đề của đoạn hoặc không chứa nội dung cơ bản trong văn bản, ta không sử dụng ưu tiên này một cách trực tiếp, tức là mọi câu mở đầu đoạn đều được chọn là câu trội. Ta bổ sung thêm thông số ưu tiên câu mở đầu đoạn vào công thức tính câu trội có kèm theo hệ số, và hệ số này có thể được điều chỉnh cho phù hợp, có thể tùy vào loại văn bản. Độ trội của câu S_i trong cụm có bổ sung thêm thông số ưu tiên câu mở đầu đoạn được tính theo hai trường hợp như sau:

- Nếu S_i không phải câu mở đầu đoạn văn, $f_1(S_i) = f(S_i)$
- Nếu S_i là câu mở đầu đoạn văn, $f_1(S_i) = (1 + a.f_{max})f(S_i)$, trong đó f_{max} là giá trị trội lớn nhất trong cụm sau khi tính f, $0 \le a \le 1$.

Tuy nhiên, thông tin thống kê từ (là thông tin chính để tính f (.) cho các câu trong cụm theo cả hai cách 1 và 2 ở trên) vẫn có một mức độ quan trọng nhất định. Do đó, ta qui định thêm, câu được chọn khi bổ sung tham số ưu tiên câu mở đầu đoạn phải có độ trội về thống kê từ lớn hơn độ trội về thống kê từ trung bình của toàn cụm.

Độ trội của câu S_i trong cụm có bổ sung thêm thông số ưu tiên câu mở đầu đoạn được tính lại như sau:

- Nếu S_i là câu mở đầu đoạn văn và $f(S_i) \ge f_{avg}$, $f_1(S_i) = (1+a(f_{max}- f_{avg}))f(S_i)$, trong đó f_{max} là giá trị trội lớn nhất trong cụm sau khi tính f, f_{avg} là giá trị trội trung bình trong cụm sau khi tính f, $0 \le a \le 1$.
- Ngược lại, $f_1(S_i) = f(S_i)$

Việc điều chỉnh hệ số a sẽ có ảnh hưởng đến kết quả tóm tắt. Nếu a càng lớn thì thông tin ưu tiên câu mở đầu đoạn càng chiếm ưu thế, và nếu a nhỏ thì thông tin thống kê từ chiếm ưu thế. Khi $a = 0$ thì thông tin ưu tiên câu mở đầu đoạn không còn ảnh hưởng đến việc tính độ trội của câu trong cụm, và khi $a = 1$ thì các câu mở đầu đoạn trong cụm sẽ trội nhất, trường hợp có nhiều câu như vậy trong 1 cụm thì câu trội nhất là một trong số các câu mở đầu đoạn đó mà có độ trội về thống kê từ cao nhất.

Việc chọn một giá trị phù hợp cho a sẽ được thực hiện qua việc thử nghiệm chương trình với nhiều giá trị a khác nhau trên nhiều văn bản.

5. ĐÁNH GIÁ KẾT QUẢ TÓM TẮT

Việc đánh giá kết quả tóm tắt của một hệ thống tóm tắt văn bản, và trang Web, tiếng Việt là vấn đề khó khăn vì hiện tại ta không có sẵn một kho tóm tắt nào, và hơn nữa là có chất lượng tốt, để có thể làm dữ liệu nền để từ đó so sánh và đánh giá các bản tóm tắt do hệ thống tự động tạo ra. Trái lại, đối với tiếng Anh, người ta đã có một kho tóm tắt mẫu DMOZ, của Open Directory Project, được xây dựng bằng tay bởi các chuyên gia làm dữ liệu nền để đánh giá các kết quả tóm tắt do các hệ thống tự động tạo ra. Sau đây là một số phương pháp đánh giá hiện có trên thế giới.

5.1.Độ đo mức chính xác - mức bao phủ theo câu

Độ đo *mức chính xác* và *mức bao phủ theo câu* [11],[12] đã được sử dụng rộng rãi để đánh giá chất lượng của một hệ thống tóm tắt. *Mức chính xác theo câu* là phần trăm các câu trong bản tóm tắt mà trùng khớp với bản tóm tắt chuẩn, chất lượng cao. Mặt khác, *mức bao phủ theo câu* là phần trăm các câu trong bản tóm tắt chuẩn mà đã được chứa trong bản tóm tắt cần được đánh giá.

5.2.Độ đo dựa trên nội dung

Độ đo *dựa trên nội dung* [11],[12] đo mức độ tương tự ở cấp độ từ vựng. Việc đánh giá được thực hiện bằng cách tạo các vec-tơ tần số xuất hiện của các từ (hay cụm từ) cho cả bản tóm tắt cần đánh giá và bản tóm tắt chuẩn, sau đó đo độ tương tự cosin giữa hai vec-tơ này. Và độ tương tự giữa hai vec-tơ càng cao thì chất lượng của bản tóm tắt càng cao.

5.3.Đánh giá tự động dùng các *Đồ thị văn bản*

Các tiếp cận trong đánh giá tóm tắt *dựa trên nội dung* bỏ qua các mối quan hệ giữa các từ khóa được thể hiện trong văn bản. Trong công trình của nhóm tác giả Jr. Santos Eugen [12] đã giới thiệu phương pháp đánh giá dùng *đồ thị văn bản*. Trong đó, họ đo độ tương tự giữa hai bản tóm tắt hoặc giữa bản tóm tắt và tài liệu được đo dựa trên các mối quan hệ (giữa các từ khóa). Trong tiếp cận này, mỗi tài liệu hay bản tóm tắt được biểu diễn bởi một *đồ thị văn bản*, là một đồ thị có hướng của các khái niệm hay thực thể và các mối quan hệ giữa chúng.

6. THỬ NGHỆM

Kết quả xử lý trên một số trang Web, được chọn loại có cấu trúc trang Web tương đối phức tạp, và phần nội dung tương đối lớn. Ngoài ra, các kết quả này cũng được đối chiếu với kết quả tương ứng có từ chương trình MS Word, và cho thấy kết quả của chương trình chứa thông tin đầy đủ hơn của Word và không mắc lỗi như của Word như đưa ra thông tin trùng lắp, cũng như ưu việt hơn nhờ bổ sung các thông tin ngữ nghĩa tiếng Việt. Thử nghiệm được tiến hành như sau:

Văn bản nguồn

Với một mạng lưới bao gồm 20 viện, 4 phân viện và 3.638 cán bộ công nhân viên; trong đó có 250 tiến sĩ và 1.700 kỹ sư, lực lượng KHCN của ngành CN hiện nay rất hùng hậu. . . . . Nhiều công trình nghiên cứu đã được nhận các giải thưởng KHCN như: giải thưởng VIFOTEC, giải thưởng quốc tế WIPO. Sử dụng hợp lí, tiết kiệm tài nguyên, năng lượng, phát triển nguồn nguyên liệu trong nước thay thế nguồn nguyên liệu ngoại nhập; chế tạo các thiết bị, dây chuyền ở trong nước... là những chuyển biến cơ bản nhất của hoạt động KHCN trong ngành CN.

Trước tiên, Bộ CN đã thành công trong việc thực hiện thí điểm chuyển đổi Viện Máy và Dụng cụ CN (IMI) hoạt động theo mô hình công ty mẹ - công ty con. Qua thực tế triển khai, doanh thu của IMI đã tăng từ 168 tỷ đồng (năm 2000) lên 300 tỷ đồng (năm 2003). . . . . TCty đã tự chế tạo máy biến áp 110kV, đang chế tạo loại 220kV; nhận và chuyển giao công nghệ sản xuất dây và cáp điện và đã xuất khẩu sản phẩm trị giá khoảng 200 triệu USD trong năm 2003.

TCty Thép Việt Nam đang chuẩn bị đầu tư xây dựng Nhà máy luyện cán thép ở Quảng Ninh, công suất 500.000 tấn/năm thép phôi và 300.000 tấn thép cán/năm. Dây chuyền nhập khẩu toàn bộ với công nghệ hiện đại là cơ hội để nhận chuyển giao, học hỏi, làm chủ những tiến bộ mới nhất của công nghệ luyện cán thép thế giới. . . . . .

Trong lĩnh vực gia công cơ khí và chế tạo thiết bị phục vụ các ngành kinh tế, nhiều kết quả nghiên cứu đã được áp dụng thành công và khẳng định khả năng sản xuất ở trong nước. Đó là dây chuyền thiết bị sản xuất phân NPK theo công nghệ vê viên bằng hơi nước và bao viên theo các màu khác nhau, công suất 4 vạn tấn/năm. . . . . .

Tuy nhiên, ngoài những kết quả đạt được, hoạt động KHCN của ngành CN trong những năm qua còn chưa tạo ra nhiều sản phẩm có hàm lượng khoa học cao và làm nên chuyển biến mang tính đột phá trong sản xuất CN, tốc độ đổi mới công nghệ còn chậm. Một số ngành có biểu hiện lúng túng trong phương án đầu tư đổi mới công nghệ. Vì thế định hướng hoạt động KHCN đến năm 2005 của Bộ CN là tập trung vào tiếp nhận, làm chủ công nghệ mới, đặc biệt là công nghệ cao sử dụng trong sản xuất và nâng cao trình độ sản phẩm lên tương đương với các nước trong khu vực.

Bản tóm tắt đi kèm

Phát triển khoa học công nghệ (KHCN) là một trong những chương trình hành động 5 năm 2001 - 2005 của ngành công nghiệp (CN) triển khai thực hiện Nghị quyết Đại hội Đảng toàn quốc lần thứ IX. Đây cũng là mục tiêu góp phần nâng cao giá trị sản xuất của toàn ngành CN trong năm 2003 tăng lên 16%. Tạo bước phát triển, lựa chọn công nghệ thích hợp để sản xuất các sản phẩm CN xuất khẩu chính là những thành tựu mà hoạt động KHCN đã và đang đem lại.

Kết quả tóm tắt của chương trình

Lựa chọn: Phương pháp so sánh cụm: CLD; ngưỡng tương tự: 0.2; phương pháp chọn câu trội: tương hỗ; hệ số ưu tiên câu mở đầu đoạn **a = 1/10**; rút gọn: **10%**.

Sử dụng hợp lí, tiết kiệm tài nguyên, năng lượng, phát triển nguồn nguyên liệu trong nước thay thế nguồn nguyên liệu ngoại nhập; chế tạo các thiết bị, dây chuyền ở trong nước... là những chuyển biến cơ bản nhất của hoạt động KHCN trong ngành CN.

Trong lĩnh vực gia công cơ khí và chế tạo thiết bị phục vụ các ngành kinh tế, nhiều kết quả nghiên cứu đã được áp dụng thành công và khẳng định khả năng sản xuất ở trong nước.

Tuy nhiên, ngoài những kết quả đạt được, hoạt động KHCN của ngành CN trong những năm qua còn chưa tạo ra nhiều sản phẩm có hàm lượng khoa học cao và làm nên chuyển biến mang tính đột phá trong sản xuất CN, tốc độ đổi mới công nghệ còn chậm.

Kết quả tóm tắt của MS Word: (độ nén 10%)

TCty đã tự chế tạo máy biến áp 110kV, đang chế tạo loại 220kV; nhận và chuyển giao công nghệ sản xuất dây và cáp điện và đã xuất khẩu sản phẩm trị giá khoảng 200 triệu USD trong năm 2003.

Công ty Thép Miền Nam đã nghiên cứu áp dụng các tiến bộ kỹ thuật mới trong sản xuất. Trong lĩnh vực gia công cơ khí và chế tạo thiết bị phục vụ các ngành kinh tế, nhiều kết quả nghiên cứu đã được áp dụng thành công và khẳng định khả năng sản xuất ở trong nước.

Nhận xét

Kết quả của chương trình khá đủ ý. Trong khi đó, đa số các nội dung được bao quát trong bản tóm tắt bằng Word đều thuộc phần đầu của văn bản gốc.

Trong phần rút câu trội, ta đã đề cập đến việc bổ sung thông số ưu tiên câu mở đầu đoạn văn vào công thức tính độ trội của câu trong cụm. Trong đó, *a* là hệ số tham gia của thành phần ưu tiên câu mở đầu đoạn vào công thức tính đó. Việc điều chỉnh hệ số *a* sẽ có ảnh hưởng khác nhau lên kết quả tóm tắt khi cách viết văn trong các văn bản thay đổi. Đối với các văn bản mà trong đó các đoạn được viết theo dạng diễn dịch, tức câu mở đầu mang nội dung chính thì khi hệ số *a* càng lớn thì chương trình cho kết quả càng tốt vì với giá trị *a* lớn các câu mở đầu sẽ dễ dàng được chọn. Ngược lại, với văn bản có các đoạn văn không được viết theo dạng diễn dịch, câu chủ đề có thể nằm giữa đoạn hay cuối đoạn thì *a* càng lớn chương trình càng có kết quả kém.

Hình 1: **Giao diện của chương trình.**

7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi xây dựng hệ thống tạo tóm tắt trang Web tiếng Việt dựa trên việc trích các câu từ tài liệu gốc để tạo tóm tắt. Hệ thống này được phát triển theo hướng thống kê từ kết hợp với sử dụng các tri thức về văn bản như ưu tiên câu mở đầu đoạn, chiều dài câu, câu thuộc đoạn đầu tiên, hay đoạn cuối cùng trong văn bản. So với các đề tài trước đây chỉ xử lý dựa trên thống kê các dãy từ phổ biến, giải pháp đề xuất đã áp dụng các kết quả xử lý ngôn ngữ tự nhiên đã có (như tách từ, gán nhãn từ loại), từ đó thống kê trên các thành phần chính trên câu thay vì là các dãy từ. Việc phát hiện các cụm từ cho phép chúng tôi bước đầu giải quyết bài toán ngữ nghĩa trong quá trình tạo tóm tắt văn bản. Chúng tôi đang xây dựng từ điển đồng nghĩa nhằm phát hiện các từ đồng nghĩa trong văn bản và nhờ đó có thể tăng hiệu quả của việc trích câu trội bằng việc điều chỉnh thành phần của vector đặc trưng câu dựa trên các từ đồng nghĩa trong bước so sánh vector đặc trưng của hai câu.

EXTRACTING AND SUMMARIZING THE CONTENT OF VIETNAMESE WEB PAGES

Do Phuc, Ho Anh Thu
Center for Information Technology Development (CITD) – VNU-HCM

ABSTRACT: *Document summarization is to make an abridge version of document or a set of documents. The problem of document summarization has a long history of development. The first work in this research direction has belonged to Luhn since 1958. With the increase of volume of information in the Internet especially the Web pages in English or in Vietnamese language Web pages, the problem of developing the summarization techniques which can help to summarize the content of web pages or documents has been the interest of researchers. In this paper, we would like to present the results of building a summarization system for summarizing the content of Vietnamese web pages based on the extraction of salience sentences from original document. We use the natural language processing such as word segmentation, POS tagging, compound noun extracting for increasing the efficiency of document summarization and opening a solution for semantic text summarization.*

TÀI LIỆU THAM KHẢO

[1]. Lại Thị Hạnh, *Trích cụm danh từ tiếng Việt nhằm phục vụ cho các hệ thống tra cứu thông tin đa ngôn ngữ,* Luận văn thạc sĩ Tin học, Thư viện Cao học, Khoa Công nghệ thông tin, Trường Đại Học Khoa học Tự nhiên, ĐH Quốc gia TPHCM, 2002.

[2]. Võ Lý Hòa, *Tìm hiểu văn bản tóm tắt và phương pháp tóm tắt văn bản*, Luận án tiến sĩ Ngữ văn, Đại học Khoa học Xã hội và Nhân văn, TP.HCM, 2004

[3]. Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương, "Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt", *Kỷ yếu Hội thảo ICT.rda'03*, Hà Nội, 2003.

[4]. Đỗ Phúc, Hoàng Kiếm, "Rút trích ý chính từ văn bản tiếng Việt hỗ trợ tạo tóm tắt nội dung", *Tạp chí Các công trình nghiên cứu - triển khai viễn thông và công nghệ thông tin*, số 13, trang 59-63, 2004.

[5]. Đồng Thị Bích Thủy, Hồ Bảo Quốc, Ứng dụng xử lý ngôn ngữ tự nhiên trong hệ thống tìm kiếm thông tin trên văn bản tiếng Việt. ĐHKHTN, TP. Hồ Chí Minh, 2001.

[6]. Aone C., M. E. Okurowski, J. Gorlinsky, and B. Larsen, "A scalable summarization system using robust nlp", *Proceeding of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, p.66-73, 1997.

[7]. Barzilay R., and M. Elhadad, "Using lexical chains for text summarization", *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, ACL, Madrid, Spain, 1997.

[8]. Delort -Y. J., B. Bouchon-Meunier, and M. Rifqi, "Enhanced Web Document Summarization Using Hyperlinks", *under submission*, 2003.

[9]. Dinh Dien, Hoang Kiem, Nguyen Van Toan, "Vietnamese Word Segmentation", *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPR2001)*, p. 749-756, Tokyo, 2001.

[10]. Goldstein J., M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing text documents: Sentence selection and evaluation metrics", *Proceedings of SIGIR*, 1999.

[11]. Hassel Martin, Automatic text summarization evaluation, *Term Paper*, Royal Institute of Technology.

[12]. Jr. Santos Eugen, Ahmed A. Mohamed, and Qunhua Zhao, "Automatic Evaluation of Summaries Using Document Graphs", *ACL*, 2004.

[13]. Luhn H. P., "The Automatic Creation of Literature Abstracts", *IBM Journal of Research Development*, 2(2), p. 159-165, 1958.

[14]. Mallet Daniel, *Text Summarization: An Annotated Bibliography*, (Last compiled June 24), 2003.

[15]. Nguyen Thi Minh Huyen, Laurent Romany , Xuan Luong Vu, "A Case Study in POS Tagging of Vietnamese Texts", *TALN 2003*, Batz-sur-Mer, 2003.

[16]. Oard Douglas W.; "The Vector Space Model", *LBSC 708A/CMSC*, 838L, session 3, 2001.

[17]. Radev Dragomir R., Eduard Hovy and Kathleen McKeown, "Introduction to the special issue on summarization", *Computational Linguistics*, 28(4), p.399-408, 2002.

[18]. Zha H., "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering", *SIGIR'02*, p. 113-120, 2002.

[19]. Zhang Y., N. Zincir-Heywood, Evangelos Milios, "World Wide Web Site Summarization", *Technical Report CS-2002-08*, Faculty of Computer Science, Dalhousie University, 2002.