THỬ ĐỀ XUẤT QUY TRÌNH TỰ ĐỘNG TÓM TẮT VĂN BẢN KHOA HỌC

PGS. TS. VƯƠNG TOÀN
Viện Thông tin Khoa học Xã hội

1. Trong thời đại bùng nổ thông tin, thời gian luôn có hạn mà với mỗi người, ngày càng nhiều việc phải làm hơn. Hoạt động thông tin – thư viện đã có nhiều đổi mới, trong đó có việc sử dụng những thành quả của công nghệ thông tin để đáp ứng phần nào nhu cầu tham khảo những công trình khoa học, đã được công bố không chỉ trong nước mà cả ngoài nước, không chỉ bằng tiếng mẹ đẻ mà cả bằng các ngoại ngữ phổ biến, lúc đầu chỉ cần là xác định danh mục và địa chỉ tư liệu cần tim đọc...

Thế nhưng việc xác định được tài liệu thật sự cần tham khảo không phải lúc nào cũng dễ dàng _ nhất là đối với các nhà nghiên cứu trẻ, thâm niên chưa nhiều - nếu chỉ qua tên sách, tên bài trong các CSDL thư mục. Chẳng hạn; không mấy thông tin được phản ánh qua các tên bài/sách; ví như: *Vết nứt trong ứng dụng* (Ấn phẩm: TẠP CHÍ *THẾ GIỚI VI TÍNH* A tháng 3/2006, trang 11; Thực hiện: PC World Mỹ**,** 03/2006), *Nỗi oan* thì, mà, là của Nguyễn Đức Dân (TP HCM, Nxb Trẻ, 2001; tái bản, 2003), *Dire et ne pas dire* (Nói và không nói) của O. Ducrot (P., 1972),...

Và thế là một trong các hoạt động xử lý thông tin xuất hiện được nhiều người dùng tin quan tâm, đó là các dạng tóm tắt văn bản, với nội dung có phần khác biệt – nên không hẳn đã có sự tương ứng về thuật ngữ giữa các ngôn ngữ. Chẳng hạn, tiếng Việt có: *tóm tắt, giới thiệu sách, điểm sách, lược thuật, bình thuật,..*(tạp chí *Thông tin khoa học xã hội* luôn có mục *Giới thiệu sách nhập về Thư viện...*) ; tiếng Pháp có *résumé, lecture (de livre); compte-rendu* (tạp chí *Bulletin de la Sociéte de la Linguistique de Paris* ra mỗi năm 2 số thì số thứ 2 luôn dành điểm lại các công trình ngôn ngữ học trên thế giới mà Toà soạn tiếp cận được), *annotation* (trong các *bulletin signalétique*),... ; tiếng Anh có: *summary, abstract, book review*,…(tạp chí *Vietnam Social Science* luôn có mục Book review); tiếng Nga có *referat* (Viện Thông tin KHXH Nga có bộ *referativnyi jurnal*),…

Các dạng tóm tắt này đều do con người xử lý, nghĩa là do những người có hiểu biết tốt về chuyên ngành (và ngoại ngữ) đọc rồi tóm tắt, nên đảm bảo được tính mạch lạc của văn bản; song cũng vì thế không khỏi mang dấu ấn chủ quan của người xử lý; trong khi đặc điểm của văn bản khoa học là trong mỗi văn bản, tác giả – nhà khoa học – luôn mong muốn trình bày, thậm chí là khẳng định một ý tưởng khoa học, cần được trình bày lại dù là dưới dạng tóm tắt một cách hết sức khách quan..

Với những tiến bộ của công nghệ thông tin, để có bản tóm tắt ngắn gọn nhưng vẫn đủ ý, và đặc biệt là hết sức trung thành với văn bản gốc, chúng ta hoàn toàn có thể nghĩ tới việc tự động tóm tắt văn bản khoa học, với những quy trình riêng thích hợp.

2. Khác với việc (người) đọc rồi diễn đạt lại một cách tóm tắt như lâu nay các nhà tư liệu học (documentalistes) ở các cơ quan thông tin – thư viện hoặc các biên tập viên trên các phương tiện thông tin đại chúng thường làm, ở đây chúng tôi muốn đề cập đến một quy trình cho phép máy tính có thể *tự động tóm tắt văn bản khoa học* (dưới đây, xin được gọi tắt là *văn bản*) tương đối chính xác nhất.

Trong điều kiện hiện nay của khả năng kỹ thuật, chúng tôi cho rằng trước mắt, chúng ta mới chỉ có thể xây dựng được quy trình cho phép máy làm được việc rút trích văn bản. Nghĩa là máy chưa làm được việc sắp xếp lại văn bản, thêm từ nối liên kết,…nhằm Việt hoá tối đa văn bản tóm tắt cuối cùng có được.

2. 1. Để xác định rõ xuất xứ của văn bản (thuận tiện cho việc tiếp cận khai thác về sau), việc đầu tiên máy có thể làm là:

Lệnh cho máy tự động nhận dạng các yếu tố thư mục, rồi miêu tả văn bản lần lượt theo các nguyên tắc biên mục, nhờ sự trợ giúp của một phần mềm hỗ trợ thư viện).

Đó là: tên tác giả, tên sách, nơi xuất bản, nhà xuất bản, năm xuất bản, số trang, đối với sách.

Với bài viết, là tên bài, sau đó là tên tạp chí, số, năm, trang (từ… đến…).

2. 2. Sau đó, ta cho máy tiến hành tự động tóm tắt văn bản. Có 2 khả năng xảy ra: sẵn có và phải rút trích.

2. 2. 1. Cho máy đọc lướt nhằm phân loại văn bản, chủ yếu xem ngay trong văn bản đã *có sẵn* (những) đoạn văn mang tính chất "tóm tắt" hay không? Nếu có, thì máy sẽ tự động "cắt" lấy ngay các đoạn này.

Để phát hiện (những) đoạn văn mang tính chất "tóm tắt", máy sẽ phải nhận dạng xem có *phần mào đầu* (chapeau) hay không?

Hay trong bài có các từ tổng quan hay Tóm lại (là) không?

Hoặc khi máy phát hiện trong bài viết có:

- Bài viết này giới thiệu…
- Chủ đề của bài viết này là….
- Tổng quan
- Bài viết cũng
- Kết luận rút ra được là…
- Trên đây, tôi đã giới thiệu…
- Hi vọng bài viết

Thì máy sẽ tự động lấy hết câu có các từ trên hay đoạn này.

Tự động lược bỏ phần cuối trước hai chấm. Ví dụ:

…có các tính năng cơ bản sau:

Những đoạn chữ không bình thường trong văn bản (như được in đậm, hoặc in nghiêng) cần được lưu ý vì đó là những chỗ cần nhấn mạnh. Với văn bản có đánh số (numbering) trong bài [Ví dụ, bài: Hồ Tất Thành Bí mật về bộ lọc vật liệu. Thế giới vi tính A tháng 4/2006, tr. 125] thì máy sẽ tự động cắt lấy các tiêu đề (lớn nhỏ này), trước đó, tự động thêm vào đoạn câu sau:

- Bài viết này gồm các đoạn/phần sau:

2.2.2. Nếu văn bản không thuộc loại trên [nghĩa là không có (những) đoạn văn mang tính chất "tóm tắt"], thì sẽ phải tiến hành các bước theo quy trình sau:

- Định chủ đề, xác định 4-5 tiêu đề đề mục (subject headings) hoặc từ khoá (key words), rồi để máy tự động chọn lưu tất cả những câu có các từ khoá đó / Hoặc để máy tự động xoá tất cả các câu còn lại. Cũng cần lưu ý rằng không ít trường hợp, từ khoá không có ngay trong tiêu đề. Ví dụ bài: Vết nứt trong ứng dụng (**PC World Mỹ,** 03/2006)

*

* *

Có thể minh họa quy trình trên qua các bước như các cách có thể thực hiện trong ví dụ sau:

Tên bài cần tóm tắt: Tổng quan Home Networking và ứng dụng

Ấn phẩm: Tạp chí Bưu chính, viễn thông & Công nghệ thông tin. 12:00, 23/02/2006 **Thực hiện:** ***Ths. Nguyễn La Giang*** 16 tr.

Yêu cầu; Bài tóm tắt đủ ý chính, song chỉ dài tối đa là 1,6 tr.

Có hai cách, tuỳ theo việc có dựa hoặc không dựa vào từ khoá (TK).

Cách a: **Dựa vào từ khoá**

Giữ lại các dòng có các TK: mạng trong nhà: 27, Home Networking: 14.

Giữ lại các dòng có đánh số ở đầu, được in đậm và nghiêng

Tự động lược bỏ phần cuối trước dấu hai chấm (:), ở đây là: có các tính năng cơ bản sau:

Đánh giá: Chưa đạt yêu cầu vì Bài tóm tắt có được dài tới 3,5 trang.

Do đó, cần bổ sung thêm:

Cách a1:

1. Giữ lại toàn bộ những câu có:

tổng quan: 1 (không kể ở đề bài)

Bài viết này: 1

Bài viết cũng : 1

2. Giữ lại toàn bộ những câu trong phần Kết luận

3. Sau đó, giữ lại các dòng có TK mạng trong nhà: 27, Home Networking: 14.

4. Chỉ giữ lại các dòng có đánh số ở đầu, được in đậm và nghiêng

Tự động lược bỏ phần cuối trước dấu hai chấm (:), ở đây là: có các tính năng cơ bản sau:

Cách a2:

1. Giữ lại những câu có các từ:

- tổng quan: 1 (không kể ở đề bài)

- Bài viết này: 1

- Bài viết cũng : 1

2. Và giữ lại các dòng có các TK: mạng trong nhà: 27, Home Networking: 14. Sau đó chỉ giữ lại những câu có các TK trên trong:

- những tiểu mục/câu được in đậm.Tự động lược bỏ phần cuối trước hai chấm: có các tính năng cơ bản sau:

- Kết luận

Đánh giá cả a1 và a2: Đạt yêu cầu, vì Bài tóm tắt có được chỉ dài 1,5 trang.

Cách b: **Không dựa vào từ khoá**

1. Giữ lại những câu có:

- tổng quan: 1 (không kể ở đề bài)

- Bài viết này: 1

- Bài viết cũng : 1

- Kết luận

2. Và những câu được in đậm và nghiêng. Tự động lược bỏ phần cuối trước hai chấm: có các tính năng cơ bản sau:

Đánh giá: Đạt yêu cầu, vì Bài tóm tắt có được chỉ dài 1,5 trang.

3. Trên đây, chúng tôi đã thử đề xuất quy trình làm việc để máy tính có thể tiến hành tự động tóm tắt văn bản khoa học (tỉ lệ 1/10). Công việc đã được thử nghiệm ở một số ví dụ kiểu như trên cho thấy kết quả là đáng khích lệ.

Tuy nhiên, trong một số trường hợp tương tự, do những câu được cắt tự động rất có thể trở thành câu cụt hoặc thừa từ, song với nhà nghiên cứu chuyên ngành thì vẫn có thể đọc hiểu… Và cũng do vậy, chúng tôi rất hi vọng sẽ có được sự trợ giúp của phần mềm *Grammatical autocorrect for Vietnamese,* sao cho khôi phục tính liên kết vản bản, phù hợp với tính mạch lạc của tiếng Việt.

Quy trình này đã được góp ý kiến và trao đổi trong phạm vi một nhóm nghiên cứu mà chúng tôi có điều kiện tham gia. Rất mong được quý đồng nghiệp cho thêm ý kiến để chúng ta sớm có (những) mô hình tối ưu tự động tóm tắt văn bản khoa học khả thi, đáp ứng tốt hơn nhu cầu của người dùng tin.

Hà Nội, tháng 2 năm 2007

PGS. TS. Vương Toàn (thứ hai từ trái sang) tại Hội nghị IFLA, Bangkok