PHƯƠNG PHÁP TÌM TIN KHÔNG PHẢI LẬP CÂU HỎI MÀ DỰA VÀO THỨ TỰ CÁC YẾU TỐ NỘI DUNG, DÙNG CHO CÁC KHO LƯU TRỮ TIN TỨC ĐA PHƯƠNG TIỆN[10]

Daisuke Kitayama và Kazutoshi Sumiya
Trường Khoa học Nhân văn và Môi trường, Đại học Hyogo

Nội dung tin tức dưới dạng hình ảnh và văn bản mới đây đã được truyền phát trên truyền hình, báo chí và Internet. Mặc dù, nội dung hình ảnh về các tin tức đã lỗi thời ít có giá trị để xem, nhưng vẫn có thể được coi là có giá trị khi so sánh với nội dung liên quan. Những tin tức lặp đi lặp lại, đặc biệt cần so sánh, thí dụ, các đại hội thể thao Ôlympic và các cuộc triển lãm quốc tế. Chúng tôi xin đưa ra một phương pháp tìm nội dung so sánh dựa vào thứ tự các yếu tố tin tức. Phương pháp này gồm 2 phần. Phần đầu phân tích nội dung tin tức mà ai đó đang dò tìm. Phần hai tự động sản sinh ra các câu hỏi để tìm nội dung về các tin tức so sánh.

1. Mở đầu

Thông tin thường được phổ biến qua các chương trình thời sự và tin tức không chỉ trên truyền hình và báo chí mà còn qua Internet. Tuy nhiên, tin tức chỉ được thông báo trên những kênh này trong một thời gian ngắn (chừng một tuần hoặc một tháng). Tính thời sự của tin tức từ những kênh này thường được coi là quan trọng. Thế nhưng, những mục so sánh các sự kiện Ôlympic trong quá khứ với các sự kiện hiện tại lại bao gồm những bài đặc biệt. Những tin tức cũ không ai tìm kiếm không được coi là còn có giá trị. Tuy nhiên, chúng tôi vẫn coi là có giá trị nếu tìm thấy mối quan hệ giữa tin tức cũ và tin tức mới.

Phương pháp mà chúng tôi đưa ra gồm hai quá trình. Một là, các đối tượng hoặc sự vật được thông báo và các hành vi (hành động) trong tin được tách ra dựa vào thứ tự của các yếu tố nội dung, khác nhau tùy theo phương tiện. Hai là, các mục tin được tìm thấy có thể so sánh một cách hiệu quả với các mục tin mà người dùng đang dò tìm. Một người dùng tin có thể tự động nhận được nội dung để nắm bắt các tin tức với phương pháp của chúng tôi, chỉ cần dò tìm các mục tin và lựa chọn câu hỏi so sánh.

Hình 1 phác họa khái niệm cơ bản của phương pháp mà chúng tôi đưa ra.

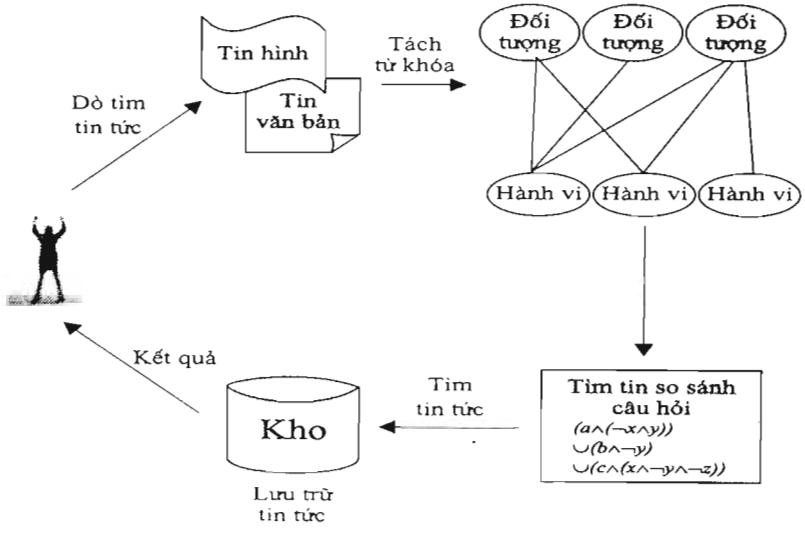
Công trình liên quan

Shin và các cộng sự [1] đã đề nghị tạo ra yêu cầu tin từ những câu hỏi bằng ngôn ngữ tự nhiên. Hệ thống mà họ đưa ra phân tích một cách tự động câu hỏi của người dùng tin sử dụng các từ khóa 5-W 1-H. Phương pháp mà chúng tôi đưa ra không cần phân tích ngữ pháp phức tạp cũng không cần xây dựng từ điển, bởi vì nó không lệ thuộc vào các từ khóa cụ thể. Thay vì như vậy, các từ khóa cấu thành các tin chỉ tách ra mà thôi.

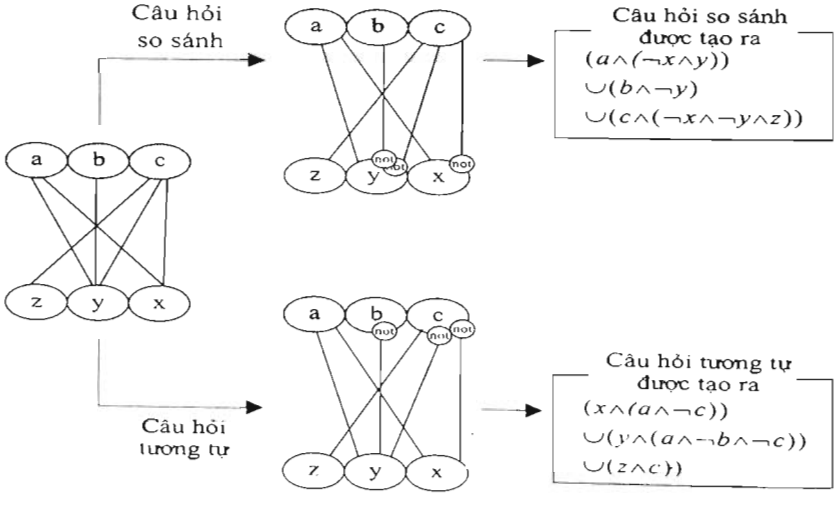
Ohshima và các cộng sự [2] đã đưa ra

[10] Bài viết được trình bày tại Hội nghị Quốc tế Thư viện số châu Á lần thứ X tổ chức tháng 12/2007 tại Hà Nội (BBT)

phương pháp tách các trang đôi. Yumoto và các cộng sự [3] đã đưa ra phương pháp tách các tập hợp trang có quan hệ. Họ đưa ra phương pháp phát hiện các mối quan hệ giữa nội dung, sử dụng mô hình không gian véctơ. Phương pháp của chúng tôi phát hiện các mối quan hệ dựa vào từ khóa mà không dùng mô hình không gian véctơ.



Hình 1. *Khái niệm cơ bản của tìm tin tức so sánh mà không phải lập câu hỏi*



Hình 2. *Tạo ra các câu hỏi so sánh*

3. Tách từ khóa bằng cách sử dụng thứ tự của các yếu tố nội dung

Chúng tôi xác định thứ tự của các yếu tố nội dung như là những đơn vị cơ bản trong thứ tự của nội dung tin tức. Chúng có đặc điểm khác nhau tùy theo phương tiện [4] [5]. Chúng tôi cho rằng, các đối tượng của chủ đề được mô tả trong tin tức thường được diễn đạt bằng danh từ, và hành vi được diễn đạt bằng tập hợp các động từ. Một mục tin có thể được diễn đạt bằng cách sử dụng danh từ cho đối tượng và động từ cho hành vi.

Phương pháp mà chúng tôi đưa ra, tách các đối tượng từ các chủ đề được mô tả trong tin tức bằng cách sử dụng các đơn vị cơ bản. Chúng tôi cho rằng, đối tượng của một chủ đề trong tin hình mô tả chính xác đối tượng được chiếu trên màn hình. Bằng cách này, mức độ quan trọng của từ khóa chỉ đối tượng có thể được tính bằng mật độ từ trong bản chuyển tả tin hình. Mức độ quan trọng của từ khóa chỉ đối tượng, *a*, trong tin hình có thể được tính toán theo công thức:

$$\text{Giá trị đối tượng} = \frac{i}{kc\,(a_1\,,\,a_i\,)} \qquad (1)$$

trong đó, a_i là danh từ a thứ i^{th} trong tin tức.

Hàm số *kc* tính khoảng cách giữa các câu. Khoảng cách giữa các câu được thể hiện bằng một số và có nghĩa là có bao nhiêu câu giữa hai từ khóa. Khoảng cách giữa các câu là 1 khi các từ khóa xuất hiện trong cùng một câu. Chúng tôi cho rằng, các vị trí, nơi mà các đối tượng của chủ đề được mô tả trong tin văn bản, bị phân tán. Mức độ quan trọng của từ khóa chỉ đối tượng, *a*, trong tin văn bản có thể được tính toán theo công thức:

$$\text{Giá trị đối tượng} = min\left(\frac{\sum_{i=1}^{n} dist(s_1, a_i)}{n}, \ldots, \frac{\sum_{i=1}^{n} dist(s_j, a_i)}{n}, \ldots, \frac{\sum_{i=1}^{n} dist(s_m, a_i)}{n}\right) \qquad (2)$$

Trong đó, s_j là câu thứ j^{th} trong tin văn bản. Giá trị tối thiểu của yếu tố này được tách ra bằng cách sử dụng hàm *min* bởi vì không biết kết quả mong đợi.

Phương pháp mà chúng tôi đưa ra tách những hành vi trong tin bằng cách sử dụng thứ tự trình bày nội dung. Chúng tôi coi: kết luận được mô tả ở phần cuối tin hình và động từ chỉ hành động trong kết luận đã diễn đạt những hành vi trong tin. Mức độ quan trọng của từ khóa chỉ

hành vi được tính bằng vị trí mà nó xuất hiện trong phần chuyển tả tin hình. Mức độ quan trọng của từ khóa chỉ hành vi trong tin hình được tính theo công thức:

$$Giá\ trị\ hành\ vi = \sum_{i=1}^{S} \left(\frac{i}{S} \times hàm\ số\ đếm\ (V_i) \right) \qquad (3)$$

Trong đó i là câu thứ i^{th} trong toàn bộ các câu S, và V_i là tập hợp động từ xuất hiện trong câu thứ i^{th}. Hàm đếm (count) tính số lượng động từ phải được tính toán trong V_i. Chúng tôi cho rằng một từ khóa chỉ hành vi trong tin văn bản thường xuất hiện ở phần đầu nơi mà các chi tiết về kết luận được mô tả. Mức độ quan trọng của từ khóa chỉ hành vi trong tin văn bản được tính theo công thức:

$$Giá\ trị\ hành\ vi = \sum_{i=1}^{S} \left(\frac{S - i + i}{S} \times hàm\ số\ đếm\ (V_i) \right) \qquad (4)$$

4. Tạo câu hỏi để tìm các mục so sánh

Mục so sánh là những mục có thể được so sánh bằng cách chú trọng vào các tin tức dò tìm. Chúng tôi xác định các tin tức, nơi mà tâm điểm chú ý là một đối tượng coi như một mục so sánh, và nơi mà tâm điểm chú ý là một hành vi coi như một mục tương tự. Câu hỏi được tạo ra dựa vào một đồ thị, ở đó yếu tố nội dung được mô tả. Đồ thị yếu tố nội dung là một đồ thị lưỡng phân bao gồm từ khóa chỉ đối tượng và từ khóa chỉ hành vi. Đường liên kết cho thấy mối quan hệ giữa đối tượng và hành vi. Chúng tôi cho rằng, mối quan hệ giữa từ khóa chỉ đối tượng và từ khóa chỉ hành vi trong tin hình được xác định bởi một khoảng, nơi có mật độ từ cao của từ khóa chỉ đối tượng. Tuy nhiên, quan hệ giữa từ khóa chỉ đối tượng và từ khóa chỉ hành vi trong tin văn bản được xác định bằng cách sử dụng cùng một mẩu tin hay đoạn văn.

Các câu hỏi so sánh được tự động tạo ra để tìm các mục so sánh liên quan đến chủ đề mà người dùng tin đang dò tìm. Người dùng có thể khẳng định tình hình có nhất quán hay không qua từng thời kỳ. Phần trên của Hình 2 cho thấy một câu hỏi so sánh được tạo ra như thế nào.

Các câu hỏi tương tự được tự động tạo ra để tìm các mục tương tự liên quan đến chủ đề mà người dùng tin đang dò tìm. Người dùng có thể hiểu được hành vi một cách chi tiết. Phần dưới của Hình 2 cho thấy một câu hỏi tương tự được tạo ra như thế nào.

5. Đánh giá

Chúng tôi đã làm thử nghiệm để đánh giá phương pháp mà mình đưa ra bằng

cách đánh giá những kết quả tìm được nhằm tạo ra những câu hỏi, có sử dụng đồ thị các yếu tố tin tức. Những tập hợp dữ liệu cho mỗi câu hỏi được tạo ra trong thử nghiệm này có khoảng 180 mục tin trong hồ sơ lưu trữ tin tức[11]. Tin hình và tin văn bản được đưa vào trong những tập hợp câu dữ liệu này. Những chủ đề thử nghiệm đã tách ra những tập hợp trả lời đúng từ các dữ liệu khi dò tìm tin tức và đã tạo ra những câu hỏi rồi sau đó so sánh các câu hỏi. Có ba chủ đề thử nghiệm. Một tập hợp trả lời đúng là một tập hợp các mục mà hai chủ đề hoặc nhiều hơn đã tách ra. Chúng tôi đánh giá phương pháp đã đưa ra bằng các tỷ lệ chính xác, đầy đủ (mức độ truy hồi), và số đo tần số, được tính toán có dùng các tập hợp câu trả lời đúng. Kết quả được liệt kê trong Bảng 1.

Số đo tần số của câu hỏi được tạo ra từ tin hình cao hơn số đo tần số của câu hỏi được tạo ra từ tin văn bản. Chúng tôi cho rằng các phương tiện khác nhau không xử lý như nhau trong phương pháp đã được đưa ra. Do đó, chúng tôi cần phải cải tiến thuật toán.

6. Những nhận xét kết luận

Báo cáo trình bày đồ thị yếu tố nội dung với mức độ quan trọng dựa vào thứ tự các yếu tố nội dung; dùng đồ thị này để giới thiệu việc tạo ra các câu hỏi nhằm so sánh các mục đã tìm được. Chúng tôi cũng đánh giá kết quả tìm được bằng cách sử dụng các câu hỏi so sánh, được tạo ra bằng phương pháp đã giới thiệu. Trong công việc sắp tới, chúng tôi dự kiến so sánh các phương pháp tính toán từ khóa trong các thử nghiệm với mức độ quan trọng qui ước, đồng thời cải tiến việc tạo ra những câu hỏi dựa vào quan hệ giữa các từ khóa riêng lẻ.

Vũ Văn Sơn *dịch*

Tài liệu gốc: *"Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers", pp.* 216-219

Tài liệu tham khảo

1. Shin, S.E., Seo, Y.H. Query Generation Using Semantic Features. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 234-243. Springer, Heidelberg (2006)

2. Ohshima, H., Oyama, S., Tanaka, K: Sibling Page Search by Page Examples. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 244-253. Springer, Heidelberg (2006)

3. Yumoto, T., Tanaka, K.: Page Sets as Web Search Answers . In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 254-263. Springer, Heidelberg (2006)

4. Wikinews: Style guide. http://en.wikinews.org/wiki/Wikinews: Style guide

5. Analyzing News. http://akasaka.cool.ne.jp/kakeru3/bs3. html

[11] Khoảng 180 mục tin tức đã được giả định và 8 mục tin đã được sử dụng hàng tháng trong vòng 18 tháng và khoảng 40 mục tin tức đã