Tạp chí Tin học và Điều khiển học, T.24, S.1 (2008), 87–96

ĐẠI SỐ GIA TỬ VÀ BÀI TOÁN SẮP XẾP MỜ TÀI LIỆU TIẾNG VIỆT

TRẦN THÁI SƠN, LÊ QUỐC THÁI, NGUYỄN VĂN NAM

Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

Abstract. This paper presents the solution to problems of fuzzy ranking Vietnamese documents using approach of Hedge Algebra based on envaluating an accordance of found documents by heuristic criteria. These criteria firstly are evaluated by degree of importance (weight) and then are combined following some aggregation operator to define final evaluation unique number, that will be used for sorting the documents. The experimental results are presented to show the effectiveness of the proposed solution.

Tóm tắt. Bài báo trình bày một giải pháp cho bài toán sắp xếp mờ tài liệu tiếng Việt theo hướng tiếp cận lý thuyết về Đại số gia tử trên cơ sở đánh giá mức độ đáp ứng của các tài liệu tìm thấy theo một số tiêu chí heuristics. Những tiêu chí này được đánh giá theo mức độ quan trọng (trọng số) và được tổng hợp theo một phép kết nhập để cho ra con số đánh gia cuối cùng - làm căn cứ để sắp xếp tài liệu. Các kết quả thí nghiệm đã được trình bày để thấy rõ hiệu quả của giải pháp.

MỞ ĐẦU

Trong thời đại bùng nổ thông tin hiện nay, việc tìm kiếm các văn bản tài liệu theo những nội dung, chủ đề hay các tiêu chí nhất định trở nên khó khăn, không chỉ trên kho dữ liệu khổng lồ Internet mà nó còn là vấn đề ngay cả khi ta tiến hành tìm kiếm trên các cơ sở dữ liệu nội bộ. Thời gian tốn kém cho việc tìm và tra cứu các văn bản này đòi hỏi phải có những phương pháp hữu hiệu để xử lý. Các phần mềm tìm kiếm (Search Engine) nối tiếp nhau ra đời đã phần nào giải quyết được vấn đề nêu trên. Tuy nhiên, phần lớn các công cụ tìm kiếm này chỉ dùng cho việc tìm kiếm tài liệu bằng tiếng Anh, khi dùng cho tiếng Việt bộc lộ những khiếm khuyết dễ thấy. Trước hết và chủ yếu là do trong tiếng Anh mỗi âm tiết riêng biệt đều có ý nghĩa nhất định tạo thành một từ, còn tiếng Việt có khi một tập hợp hai ba thậm chí bốn âm tiết mới hình thành một từ có nghĩa (nông trường, kiến thức...). Do đó, khi tìm văn bản tiếng Việt, các công cụ tìm kiếm theo kiểu tiếng Anh vẫn tìm theo từng âm tiết đơn lẻ và cho ra kết quả là những văn bản không liên quan gì đến nội dung cần tìm. Thí dụ khi tìm các văn bản có nội dung liên quan đến từ "sinh vật học" (là một ngành nghiên cứu), công cụ sẽ cho ra cả các văn bản có câu kiểu như "Hai học sinh thi đấu vật trong hội thao". Việc này làm tăng số lượng các kết quả tìm kiếm lên gấp nhiều lần mà hoàn toàn không phục vụ cho mục đích tìm kiếm. Ngoài ra, các công cụ này cũng không cho phép tìm kiếm với các văn bản tiếng Việt không dùng phông UNICODE.

Trong bài báo này, chúng tôi trình bày một giải pháp xây dựng công cụ tìm kiếm dành cho tiếng Việt. Công cụ này chỉ giới hạn trong việc xử lý câu hỏi tìm kiếm và sắp xếp các kết

quả tìm được theo một trật tự ưu tiên có chủ đích mà không đề cập đến các phần thu thập và đánh chỉ số các tài liệu, những phần quan trọng của một bộ công cụ tìm kiếm. Tư tưởng của giải pháp được đưa ra là kết hợp các đánh giá heuristic và thuật toán sắp xếp mờ dựa trên kết quả nghiên cứu của đại số gia tử. Đầu tiên chúng tôi sẽ trình bày bài toán sắp xếp mờ các văn bản tiếng Việt. Sau đó để tiện theo dõi chúng tôi sẽ trình bày một số khái niệm cơ bản về Đại số gia tử sẽ dùng trong bài báo. Tiếp theo đó giải pháp cho bài toán sắp xếp mờ tài liệu ứng dụng ĐSGT sẽ được trình bày. Và cuối cùng sẽ là chương trình thử nghiêm.

1. BÀI TOÁN SÁP XẾP MỜ TÀI LIỆU

Như trong phần mở đầu có nêu, việc sắp xếp các tài liệu tiếng Việt khi tìm kiếm thông tin trên các kho dữ liệu lớn là một vấn đề đáng được quan tâm nhưng chưa có nhiều nghiên cứu đề cập đến. Một cách tóm tắt, bài toán đặt ra ở đây là với một yêu cầu của người sử dụng, làm thế nào có thể tìm kiếm và đưa ra một danh sách các tài liệu đã có sắp xếp theo những tiêu chí nhất định để phục vụ tốt hơn yêu cầu này. Có nghĩa các tài liệu tìm được (con số có thể lên hàng nghìn, hàng chục nghìn) phải được sắp xếp theo các tiêu chí mà giải pháp cho là tốt, các tài liệu có khả năng đáp ứng đúng yêu cầu hơn thì phải được xếp trên. Tất nhiên việc cho rằng tài liệu này đáp ứng yêu cầu hơn tài liệu kia chỉ là tương đối vì suy cho cùng, chính người sử dụng cũng khó có thể biết một cách đích xác là trong hai tài liệu tìm được, cái nào đáp ứng yêu cầu của mình hơn. Tuy nhiên có những tiêu chí chung dễ được chấp nhận để đánh giá độ phù hợp của một tài liệu với yêu cầu của người sử dụng mà chúng tôi xin liệt kê sau đây:

1. Nếu các từ tìm kiếm trong yêu cầu tìm kiếm của người sử dụng nằm trong tiêu đề của tài liệu thì khả năng tài liệu phù hợp nhiều với yêu cầu của người sử dụng. Lí do tương đối rõ: các từ nêu trong tiêu đề thường là các từ quan trọng phản ánh nội dung chính của tài liệu.

2. Nếu các từ tìm kiếm nằm càng gần nhau trong văn bản thì khả năng phù hợp của tài liệu càng cao. Ở đây lí do có thể không rõ ràng như tiêu chí trên nhưng thực ra cũng đơn giản: khi đặt yêu cầu tìm kiếm, người sử dụng thường nêu ra các cụm từ xoay quanh nội dung chính của tài liệu cần tìm, và nếu các cụm từ này trong văn bản tìm ra càng nằm gần nhau thì khả năng tài liệu này có nội dung đáp ứng yêu cầu người sử dụng càng cao. Còn nếu các cụm từ này nằm càng xa nhau trong văn bản thì khả năng tài liệu này có nội dung như người sử dụng mong muốn càng thấp. Thí dụ, ta muốn tìm các tài liệu liên quan đến nông thôn Việt Nam thời đổi mới, và ta gõ nguyên cum từ này vào câu hỏi tìm kiếm. Nếu có văn bản có chứa từ "nông thôn Việt Nam" ở phần đầu và đâu đó phần cuối cùng có từ "đổi mới" thì văn bản này có thể sẽ tập trung nói về nông thôn Việt Nam nói chung, còn ở thời kỳ đổi mới chỉ là thêm thắt, do đó sẽ đáp ứng yêu cầu người sử dụng không bằng văn bản mà cụm từ kia đi liền từ đầu.

3. Tần số xuất hiện của những từ khoá trong văn bản cũng phản ánh mức độ đáp ứng của văn bản với yêu cầu tìm kiếm. Hiển nhiên, số lần xuất hiện của từ khoá càng nhiều, mức độ đáp ứng càng cao.

4. Nếu văn bản tìm kiếm trên INTERNET lấy được từ các trang WEB có uy tín trong lĩnh vực tìm kiếm thì khả năng đáp ứng sẽ tốt hơn. Điều này có lẽ là dễ chấp nhận với các văn bản thông thường, tuy vậy với các trường hợp rất đặc biệt (như văn bản "độc" (hiếm)) thì có thể không hoàn toàn chính xác, nhưng dù sao thì với số đông, tiêu chí ta đưa ra vẫn cho kết quả tốt.

5. Nếu trong yêu cầu tìm kiếm có những từ kiểu có thể so sánh được, thí dụ như "con sông dài", thì những văn bản có những cụm từ như "con sông rất dài" hay "con sông tương đối dài" sẽ được đánh giá căn cứ vào các từ nhấn (hay còn gọi là gia tử) "rất", "tương đối" theo kết quả của của lý thuyết về Đại số gia tử (ĐSGT) mà chúng tôi sẽ trình bày ở phần sau.

Như vậy ta có một số tiêu chí để đánh giá mức độ phù hợp của văn bản với yêu cầu của người sử dụng. Các tiêu chí này có thể sẽ được bổ xung trong quá trình hoạt động của công cụ tìm kiếm. Với mỗi tiêu chí ta có một kết quả đánh giá. Bài toán đặt ra là ta phải tổng hợp các đánh giá này lại thành một đánh giá chung, có quan tâm đến mức độ quan trọng khác nhau của từng tiêu chí. Đánh giá cuối cùng này sẽ giúp ta sắp xếp các tài liệu theo thứ tự mức độ phù hợp giảm dần, qua đó giúp người sử dụng chủ động trong việc xem xét các kết quả tìm kiếm.

2. ĐẠI SỐ GIA TỬ VÀ ỨNG DỤNG TRONG BÀI TOÁN SẮP XẾP MỜ

Bài toán sắp xếp (hay đánh giá) có thể nêu một cách tổng quát như sau: căn cứ vào điểm n chuyên gia $J1, J2, ..., Jn$ đánh giá m đối tượng $O1, O2, ..., Om$ theo k tiêu chí $C1, C2, .., Ck$, hãy sắp xếp m đối tượng đó theo thứ tự tăng (hoặc giảm) dần của tiêu chí tổng hợp. Đây là một bài toán rất thường gặp trong thực tế, như đánh giá học sinh, đánh giá các kết quả đấu thầu hay bình chọn các hoa hậu, các quốc gia, các thành phố theo những mục đích khác nhau.. Cũng đã có rất nhiều nghiên cứu được tiến hành trong và ngoài nước, với các cách đặt vấn đề đa dạng: điểm đánh giá là số hoặc ngôn ngữ (tức có yếu tố mờ), tiêu chí có trọng số hoặc không, chuyên gia bình đẳng hay kinh nghiệm khác nhau (dẫn đến trọng số khác nhau), sử dụng hàm kết nhập khác nhau để tổng hợp kết quả... Dựa vào lý thuyết về Đại số gia tử, chúng tôi cũng đã có bài viết về vấn đề này [5]. Trong bài báo này, dựa vào những nghiên cứu mới nhất về ĐSGT, chúng tôi đưa ra một cách giải quyết mới bài toán sắp xếp mờ các tài liệu. Trước hết xin tổng hợp lại một số kiến thức cơ bản về ĐSGT sẽ dùng đến trong bài [1, 2, 4].

2.1. Đại số gia tử

Trong ứng dụng này, các giá trị ngôn ngữ dùng để mô tả dữ liệu đối sánh được với nhau, nên chúng ta giới hạn việc nghiên cứu sử dụng đại số gia tử tuyến tính. Một cách không hình thức, ĐSGT là một cấu trúc đại số được đưa vào tập các giá trị của một biến ngôn ngữ (thí dụ biến "phù hợp"), khi ta coi tập các từ nhấn - gia tử, (thí dụ "rất", "tương đối",...) là các toán tử một ngôi, khi tác động lên các phần tử sinh của biến ngôn ngữ (thí dụ "phù hợp", "không phù hợp") cho ta tập các phần tử của ĐSGT (tập {rất phù hợp, rất không phù hợp, tương đối không phù hợp, tương đối rất không phù hợp, khá phù hợp...}) có thể sắp thứ tự

theo ngữ nghĩa của chúng ("rất rất phù hợp" > "rất phù hợp" > "tương đối phù hợp">.. > "tương đối không phù hợp"> "rất không phug hợp"..). Đại số gia tử tuyến tính được xác định bằng hệ tiên đề trong định nghĩa sau.

Định nghĩa 1. (Đại số gia tử tuyến tính) Một cấu trúc đại số $X = (X, C, H, \leq)$ với $H = H^+ \cup H^-$, C là tập gồm hai phần tử sinh dương và âm được gọi là đại số gia tử tuyến tính nếu thoả các điều kiện sau:

(A0) H^+ và H^- là các tập sắp thứ tự tuyến tính.

(A1) Mỗi gia tử hoặc dương hoặc âm đối với bất kỳ gia tử nào khác kể cả chính nó.

(A2) Nếu $h \leq k$ thì từ ($x \leq hx$ hoặc $x \leq kx$) ta suy ra $hx \leq kx$ và từ ($x \geq hx$ hoặc $x \geq kx$) ta suy ra $hx \geq kx$.

(A3) Nếu u và v độc lập thì với mọi $x \in H(u)$ ta có $x \notin H(v)$ hay $H(u) \cap H(v) = \emptyset$.

(A4) Nếu $h \neq k,\ hx = kx$ thì x là điểm dừng.

Nếu $h \neq k,\ hx < kx$ thì $h'hx < k'kx$ với mọi $h,\ k$ thuộc H.

Nếu $h \neq k,\ hx \neq x$ thì hx và kx độc lập.

(A5) Nếu $u \notin H(v)$ và $u \leq v (u \geq v)$ thì $u \leq h'v (u \geq h'v),\ h' \in H$.

Hàm đo trong đại số gia tử tuyến tính

Trong việc giải quyết bài toán sắp xếp, chỉ sử dụng cấu trúc thứ tự của ĐSGT là chưa đủ. Ta phải xem xét tiếp khái niệm ánh xạ định lượng ngữ nghĩa là khái niệm để lượng hoá các phần tử của ĐSGT.

Ánh xạ định lượng ngữ nghĩa được xây dựng dựa trên độ đo tính mờ của gia tử. Gọi $H(x)$ là tập các phần tử của X sinh ra từ x bởi các gia tử. Nghĩa là $H(x)$ bao gồm các khái niệm mờ mà nó phản ánh ý nghĩa nào đó của khái niệm x. Vì vậy, kích thước của tập $H(x)$ có thể biểu diễn tính mờ của x. Từ đó, ta có thể định nghĩa độ đo tính mờ như sau: Độ đo tính mờ của x, ta ký hiệu là $fm(x)$, là đường kính của tập $f(H(x)) = \{f(u) : u \in H(x)\}$.

Từ ý tưởng trực quan trên, chúng ta đưa ra định nghĩa độ đo tính mờ theo kiểu tiên đề hoá như sau.

Định nghĩa 2. Hàm $fm(.) : X \to [0, 1]$ được gọi là hàm độ đo tính mờ trên X nếu nó thoả mãn các điều kiện sau:

i) $fm(c^-) = W > 0$ và $fm(c^+) = 1 - W > 0$ trong đó c^- và c^+ lần lượt là phần tử sinh âm và phần tử sinh dương.

ii) Nếu $H = H^- \cup H^+$ là tập các gia tử đang xét và $H^- = \{h_1, h_2, ..., h_p\}$ với $h_1 > h_2 > ... > h_p, H^+ = \{h_{p+1}, h_{p+2}, ..., h_{p+q}\}$ với $h_{p+1} < h_{p+2} < ... < h_{p+q}$ thì $\sum_{i=1}^{p+q} fm(h_i c) = fm(c),\ c \in \{c^+, c^-\}$.

iii) Với hai phần tử x và y trong X, mọi h trong H, ta có $\dfrac{fm(hx)}{fm(x)} = \dfrac{fm(hy)}{fm(y)}$, tỷ số này không phụ thuộc vào đối số và ký hiệu là $fm(h)$ gọi là độ đo tính mờ của gia tử h.

Mệnh đề 1.1. *Cho fm là hàm độ đo tính mờ trên X. Ta có:*

i) $fm(c^-) + fm(c^+) = 1$.

ii) $\sum_{i=1}^{p+q} fm(h_i c) = fm(c),\ c \in \{c^+, c^-\}$.

iii) $\sum_{i=1}^{p+q} fm(h_i x) = fm(x)$.

iv) *Nếu* $x = h_{im}...h_{i2}h_{i1}$ *thì* $fm(x) = fm(h_{im}...h_{i2}h_{i1}c) = fm(h_{im})...fm(h_{i1})fm(c)$.

v) $\sum_{i=1}^{p} fm(h_i) = \alpha,\ \sum_{i=p+1}^{p+q} fm(h_i) = \beta$ sao cho $\alpha + \beta = 1,\ \alpha, \beta > 0$.

Định nghĩa 3. Hàm dấu $sign : X \to \{-1, 0, 1\}$ được định nghĩa đệ quy như sau:

i) $sign(c^-) = -1$ và nếu $hc^- < c^-$ thì $sign(hc^-) = sign(c^-)$, nếu $hc^- > c^-$ thì $sign(hc^-) = -sign(c^-)$.

$sign(c^+) = +1$ và nếu $hc^+ > c^+$ thì $sign(hc^+) = sign(c^+)$, nếu $hc^+ < c^+$ thì $sign(hc^+) = -sign(c^+)$.

ii) $sign(h'hx) = -sign(hx)$ nếu h' âm đối với h và $h'hx \neq hx$.

iii) $sign(h'hx) = sign(hx)$ nếu h' dương đối với h và $h'hx \neq hx$.

iv) $sign(h'hx) = 0$ nếu $h'hx = hx$.

Mệnh đề 1.2. *Với mọi gia tử* h *và phần tử* $x \in X$ *nếu* $sign(hx) = +1$ *thì* $hx > x$ *và nếu* $sign(hx) = -1$ *thì* $hx < x$.

Định nghĩa 3. $f : X \to [0, 1]$ gọi là hàm định lượng ngữ nghĩa của X nếu:

i) f bảo toàn thứ tự trên X.

ii) Với mọi $h,\ k$ thuộc H^+ hoặc $h,\ k$ thuộc H^- và $x,\ y$ thuộc X:

$$\frac{f(hx) - f(x)}{f(kx) - f(x)} = \frac{f(hy) - f(y)}{f(ky) - f(y)}.$$

Cho fm là hàm độ đo tính mờ trên X. Hàm định lượng ngữ nghĩa $v : X \to [0, 1]$ được xây dựng như sau ($x = h_{im}...h_{i2}h_{i1}$):

i) $v(c^-) = w - \alpha fm(c^-),\ v(c^+) = w + \alpha fm(c^+)$.

ii) $v(h_j x) = v(x) + sign(h_j x)\left[\sum_{i=j}^{p} fm(h_i x) - \frac{1}{2}(1 - sign(h_j x)sign(h_i h_j x)(\beta - \alpha)fm(h_j x)\right]$

nếu $j \leq p$

và $v(h_j x) = v(x) + sign(h_j x)\left[\sum_{i=p+1}^{j} fm(h_i x) - \frac{1}{2}(1 - sign(h_j x)sign(h_i h_j x)(\beta - \alpha)fm(h_j x)\right]$

nếu $j > p$.

Mệnh đề 3.1. *Cho fm là hàm độ đo tính mờ trên X, v là hàm định lượng ngữ nghĩa đã xây dựng trước đây. Khi đó ta có:*

i) $0 \leq v(x) \leq 1,\ \forall x \in X$.

ii) Nếu $x < y$ *thì* $v(x) < v(y)$ *với mọi* $x,\ y$ *thuộc* X.

Từ kết quả trên ta thấy hàm định lượng ngữ nghĩa là hàm đơn điệu tăng trong đoạn [0,1] , do đó tồn tại hàm ngược $v^{-1} : [0, 1] \to X$. Kết quả này sẽ dùng về sau.

2.2. Giải bài toán sắp xếp mờ theo cách tiếp cận ĐSGT

Xin nhắc lại, bài toán sắp xếp (hay đánh giá) có thể nêu một cách tổng quát như sau: căn cứ vào điểm n chuyên gia $J1, J2, .., Jn$ đánh giá m đối tượng $O1, O2, .., Om$ theo k tiêu

chí $C1, C2, .., Ck$, hãy sắp xếp m đối tượng đó theo thứ tự tăng (hoặc giảm) dần của tiêu chí tổng hợp. Giả sử điểm chuyên gia Jh cho đối tượng O_i theo tiêu chí C_t là D_{hit}. D_{hit} ở đây có thể là số (như 5, 7, 9...), có thể mờ (như "khoảng 8 điểm" hay "khá", "rất tốt"...). Trong trường hợp mọi D_{hit} đều là số thực và các trọng số (nếu có) cũng là số thực thì người ta sử dụng một phép kết nhập nào đó như dùng hàm trung bình cộng có trọng số (thường gặp hơn cả) hay lấy MAX, MIN (trong trường hợp đánh giá dựa trên giá trị cực đại hay cực tiểu)... Ở đây sự phức tạp của bài toán nằm ở chỗ đưa ra cách đánh giá phù hợp, tức lựa chọn hàm kết nhập. Ví dụ đơn giản như bài toán đánh giá học sinh, từ trước đến nay người ta căn cứ vào điểm số (có cả điểm đạo đức) để lấy trung bình cộng (có thể có môn có hệ số 2) rồi xếp hạng. Việc này đơn giản về mặt thao tác nhưng chính xác hay không thì còn nhiều tranh cãi. Tuy nhiên trong trường hợp D_{hit} có thể nhận giá trị không chính xác (mờ), thí dụ như "xấp xỉ 7", "không dưới 8" hay "tốt", "rất kém" thì việc xử lý đã trở nên phức tạp hơn nhiều. Theo L. Zadeh, mỗi tập mờ có thể ứng với một hàm số thực từ tập miền trị vào đoạn [0,1] , từ đó việc tính toán trên điểm được chuyển sang tính toán trên từng giá trị của hàm thực được gọi là hàm thuộc (membership function) này. Cách này có những nhược điểm lớn là cách chọn hàm thuộc rất khó, không có quy tắc nào; việc tính toán trên các hàm thuộc thường phức tạp; việc giải mờ từ các kết quả tính toán được là điều gần như không thể cho nên với các đánh giá bằng từ ngữ tự nhiên, kết quả sẽ rất không phù hợp.

Trong [5] chúng tôi đã trình bày giải pháp cho bài toán sắp xếp mờ dựa trên ĐSGT. Giải pháp này đã khắc phục được những nhược điểm kể trên, tuy nhiên phải dưa trên một giả thiết không được hoàn toàn tự nhiên là các gia tử tác động lên các phần tử sinh phải cho ra tập phân bố đều trên trục biểu diễn giá trị (hoặc nếu không ta phải bổ xung thêm các giá trị nhân tạo để có phân bố đều). Với các kết quả nghiên cứu mới gần đây về ĐSGT, cụ thể là các hàm định lượng ngữ nghĩa, ta đã có thể tính được các giá trị của một từ bất kỳ trong tập giá trị của biến ngôn ngữ mà không cần dựa vào giả thiết phân bố đều như trên (với lựa chọn α, β không bằng nhau, sự phân bố đều sẽ không có). Ngoài ra, cũng dựa trên hàm định lượng ngữ nghĩa, có thể chuyển cả các giá trị đánh giá bằng số thực về mờ để tạo nên sự thống nhất trong những tính toán tiếp theo. Việc mờ hoá này hoàn toàn có thể lý giải được do việc cho điểm là do các chuyên gia (tức là con người) xác định dựa trên kinh nghiệm và đánh giá chủ quan của mỗi người, khó có thể chính xác tuyệt đối được. Việc mờ hoá do vậy không chỉ là yêu cầu chuyên môn mà còn là đòi hỏi khách quan. Tóm lại, có thể tiến hành giả bài toán sắp xếp mờ như sau:

Nếu D_{hit} nhận giá trị số, trước hết ta sẽ tiến hành quy chuẩn khoảng xác định của miền giá trị của D_{hit}, thí dụ về đoạn [1,10] (theo cách cho điểm phổ biến hiện nay). Với giá trị mới của D_{hit} trên [1,10], tra theo bảng ta được giá trị ngôn ngữ tương ứng. Lấy giá trị định lượng ngữ nghĩa của giá trị ngôn ngữ này thay cho D_{hit} làm giá trị tính toán.

Nếu D_{hit} nhận giá trị ngôn ngữ, ta lấy giá trị định lượng ngữ nghĩa của D_{hit} làm giá trị tính toán.

Tương tự với các trọng số nếu có.

Bài toán bây giờ trở lại là bài toán sắp xếp với các số thực và ta có thể sử dụng mọi kết quả nghiên cứu đã biết về việc này để giải quyết.

3. GIẢI PHÁP CHO BÀI TOÁN SẮP XẾP MỜ TÀI LIỆU

Giải pháp của chúng tôi nêu ra ở đây cho bài toán sắp xếp mờ tài liệu tiếng Việt gồm hai phần chính.

3.1. Xử lý câu hỏi của người sử dụng

Như đã nêu ở trên, tiếng Việt và tiếng Anh có sự khác biệt về cấu tạo âm tiết. Cho nên, để xử lý đúng yêu cầu của người sử dụng, bước tách từ là một bước cần thiết trong giải pháp. Bước tách từ này nhằm các mục đích: tách được tập các từ khoá, bỏ các từ không cần thiết và tạo cả tập từ tương đương (đồng nghĩa). Vì bài báo tập trung vào vấn đề sắp xếp nên phần này chúng tôi chỉ sử dụng những giải pháp đã biết của các nhà nghiên cứu trong việc tách câu tiếng Việt. Cách đơn giản nhất là duyệt câu từ trái qua phải rồi chọn từ có nhiều âm tiết nhất có nghĩa (tra từ điển) làm từ cần tìm và lặp lại quá trình đó cho đến hết câu. Theo cách này, những câu kiểu "học sinh học sinh học" sẽ không tách được đúng, nhưng để tập trung vào việc sắp xếp nên vấn đề tách từ trong câu tiếng Việt chúng tôi sẽ quay lại trong bài báo khác. Sau khi tách được các từ, ta cần tiến hành loại bỏ tự động các hư từ, tức là các từ mang tính gắn kết, thí dụ "là", "và"... vì các từ này không có ý nghĩa trong việc tìm kiếm. Việc này được tiến hành bằng cách tra từ điển. Cuối cùng, với tập các từ tìm được, ta cũng lưu luôn cả các từ đồng nghĩa với chúng, cũng bằng cách tra từ điển các từ đồng nghĩa.

3.2. Sắp xếp tài liệu theo thứ tự ưu tiên

Sau khi tìm ra được tất cả các văn bản từ kho dữ liệu có chứa tập các từ khoá có trong yêu cầu của người sử dụng, bước tiếp theo là sắp xếp các tài liệu này theo thứ tự mức độ thoả mãn yêu cầu người sử dụng giảm dần. Để làm điều đó, với mỗi tiêu chí đã nêu ở phần trên, thực hiện đánh giá cụ thể như sau: (trong trường hợp này, số lượng chuyên gia sẽ là 1).

- Nếu tất cả các từ khoá trong yêu cầu tìm kiếm đều có mặt trong tiêu đề tài liệu, mức độ phù hợp sẽ là cao nhất (mặc định là "rất tốt", hoặc theo tuỳ chọn của người sử dụng). Nếu thiếu một từ khoá sẽ "tốt"... giảm dần đến "không tốt" khi không có từ khoá nào nằm trong tiêu đề.

- Khoảng cách giữa các từ khoá được đo bằng số âm tiết nằm giữa chúng (bằng số khoảng trống tính được). Dễ dàng thấy định nghĩa này thoả các tiên đề về khoảng cách. Khoảng cách nói chung bằng tổng các khoảng cách nói trên... Khoảng cách nhỏ nhất có đựoc sẽ ứng với đánh giá "rất tốt" (hoặc theo tuỳ chọn của người sử dụng, chỉ cần bảo toàn quan hệ thứ tự), khoảng cách lớn nhất ứng với "không tốt".

- Số lần xuất hiện lớn nhất của các từ khoá trong văn bản ứng với đánh giá "rất tốt", ít nhất ứng với "không tốt".

- Các trang WEB có thể được sắp xếp theo số lượng người truy nhập (có công cụ trên WEB chuyên làm việc này). Vì vậy việc đánh giá tài liệu tìm được theo uy tín của trang WEB là hoàn toàn làm được và cũng sắp xếp từ trang uy tín nhất là "rất tốt" đến trang thứ 10 chẳng hạn là "không tốt" (những trang sau đều là "rất không tốt").

- Nếu có các từ nhấn (gia tử) đi cùng các từ khoá, việc đánh giá tuân theo lý thuyết về ĐSGT: nếu văn bản có từ khoá x được đánh giá là "y" thì văn bản có cụm từ αx sẽ được đánh giá là αy. Chẳng hạn nếu văn bản có từ khoá "nhà to" được đánh giá là "tốt" thì văn bản có chứa cụm từ "nhà tương đối to" sẽ được đánh giá là "tương đối tốt".

Sau khi có các đánh giá, có thể tính các giá trị định lượng ngữ nghĩa của các giá trị của các phần tử của ĐSGT có các phần tử sinh là "tốt" và "không tốt". Có thể xem [3] để thấy một bảng tính cụ thể (trong đó nêu cách tính cho ĐSGT với các phần tử sinh là "small" và "large" nhưng với hai phần tử sinh của ta là "tốt" và "không tốt", cách tính hoàn toàn không khác biệt). Từ kết quả có được, có thể tổng hợp lại để có kết quả cuối cùng theo những giải pháp thông thường, chẳng hạn lấy trung bình cộng có trọng số hay trung bình bình phương (có trọng số)...

4. CHƯƠNG TRÌNH TÌM KIẾM VÀ SẮP XẾP THỰC NGHIÊM

Chúng tôi đã tiến hành viết chương trình thử nghiệm phương pháp nêu trên, dựa trên phần mềm mã nguồn mở nutch (có tại trang WEB http://lucene.apache.org/nutch). Việc đánh giá các trang tài liệu tìm được căn cứ vào các tiêu chí nêu trên và có dùng trọng số. Việc kết nhập các đánh giá riêng lẻ là dùng trung bình cộng (có trọng số). Kết quả cho thấy bước đầu hiệu quả của phương pháp đưa ra: số tài liệu được đưa ra là ít hơn hẳn (do loại các tài liệu không liên quan vì khâu xử lý tách từ) và được sắp xếp tương đối tốt.

4.1. Xử lý tách từ khóa từ câu truy vấn của người sử dụng

Câu truy vấn	Từ khóa tách được	Kết quả khi chưa tách từ	Kết quả sau khi tách từ
học sinh học sinh học	*học sinh, học, sinh học*	1828	48
hợp tác xã nông nghiệp	*hợp tác xã, nông nghiệp*	241	9
máy nghe nhạc cảm biến	*máy nghe nhạc, cảm biến*	33	3
lập trình ngôn ngữ pascal	*lập trình, ngôn ngữ, pascal*	14	11

Ta thấy kết quả sau khi xử lý (tách từ) là số lượng tài liệu tìm thấy ít hơn hẳn do không đưa ra các tài liệu không theo đúng yêu cầu (như các tài liệu chỉ chứa từ "sinh" trong thí dụ đầu, từ "xã" trong thí dụ thứ hai...).

4. 2. Sắp xếp tài liệu dựa trên đánh giá mờ

Các thí dụ sau cho thấy hiệu quả bước đầu của việc ứng dụng bài toán sắp xếp mờ.

Ví dụ 1. Từ khoá "kinh tế đô thị", kết quả tìm kiếm cho ra 286 tài liệu. Từ vị trí thứ tư các tài liệu có thay đổi so trước khi sắp xếp và sau khi sắp xếp (chính xác hơn là chỉ sắp xếp theo phần mềm mã nguồn mở đã có và thêm vào sự sắp xếp theo thuất toán của chúng tôi đưa ra). Tuy nhiên các tài liệu thay thế nhau ở vị trí này cũng không thật khác biệt về mặt đáp ứng câu hỏi người sử dụng.

Ví dụ 2. Từ khoá "kinh tế hàng hoá", kết quả tìm kiếm cho ra 648 tài liệu. Khác biệt giữa

hai danh sách (trước khi sắp xếp và sau khi sắp xếp) bắt đầu từ vị trí thứ 5, nhưng cũng như Ví dụ trên, khác biệt không thật rõ.

Ví dụ 3. Từ khoá "máy tính xách tay Việt Nam", kết quả tìm kiếm cho ra 12 tài liệu. Trước khi sắp xếp, vị trí thứ 2 là tài liệu "Long mạch:Trả lời những câu hỏi" của Bùi Trọng Liễu , sau khi sắp xếp tài liệu này bị đẩy xuống rất sâu trong danh sách. Nội dung bài này nói về long mạch (phong thuỷ), chỉ có một câu có từ khoá: "Còn chuyện mỗi lần nó (tức long mạch) di dời đi đâu, thì có phải di dời theo nó không, theo tôi không có vấn đề : ở thế kỉ 21 này, nó cũng như cái điện thoại di động hay cái máy tính xách tay, đi đâu xách nó theo cũng được mà! ". Rõ ràng nội dung bài báo rất ít liên quan đến từ khoá "máy tính xách tay Việt Nam", nên việc xếp nó xuống phía dưới danh sách là hợp lý.

Ví dụ 4. Từ khoá "giáo trình kỹ thuật điện", kết quả tìm kiếm cho ra 25 tài liệu. Trước khi sắp xếp, vị trí thứ 2 là tài liệu "Giáo trình điện tử", đưa ra danh mục các giáo trình điện tử (tức giáo trình ở dạng lưu trên mạng về đủ các lĩnh vực nhạc hoạ, ngoại ngữ, vi tính... (không liên quan mấy đến giáo trình kỹ thuật điện). Sau khi sắp xếp tài liệu này cũng bị đẩy xuống cuối danh sách.

Ví dụ 5. Từ khoá "cài đặt windows XP", kết quả tìm kiếm cho ra 20 tài liệu. Trước khi sắp xếp, vị trí thứ 3 là tài liệu " Lỗi nghiêm trọng của Windows 2000 khiến người dùng..."lộ hàng", có nội dung tóm tắt như sau: "Một nhóm chuyên gia của Israel tuyên bố vừa phát hiện được một lỗ hổng đặc biệt nghiêm trọng bên trong hệ điều hành Windows 2000 của Microsoft, cho phép hacker theo dõi toàn bộ những dữ liệu dạng text từng được gõ vào máy tính, bao gồm email, mật khẩu và cả số thẻ tín dụng.". Nội dung này rõ ràng không máy liên quan đến từ khoá "cài đặt windows XP" và do đó, việc nó bị xếp xuống dưới sau khi sắp xếp theo thuật toán mờ là điều dễ hiểu.

Ví dụ 6. Từ khoá "phần mềm mã nguồn mở", kết quả tìm kiếm cho ra 52 tài liệu. Trước khi sắp xếp, vị trí thứ 3 là tài liệu " Giới thiệu về mật mã lượng tử" có nội dung tóm tắt là: "Mật mã lượng tử là công nghệ cho phép bảo mật thông tin truyền đi bằng truyền thông quang, qua quang sợi cũng như qua không gian (FSO - Free Space Optical communications). Nó cho phép thông tin được bảo mật "tuyệt đối", không phụ thuộc vào độ mạnh của máy tính, độ tối tân của dụng cụ hay sự xảo quyệt của hacker." Rõ ràng nội dung này rất ít liên quan đến từ khoá "'phần mềm mã nguồn mở". Trong danh sách tài liệu sau khi sắp xếp theo thuật toán sắp xếp mờ, nó cũng bị đẩy xuống dưới.

Qua các ví dụ trên ta thấy hiệu quả bước đầu của thuật toán, giúp sắp xếp các tài liệu tìm thấy theo thứ tự hợp lý khả năng đáp ứng yêu cầu người sử dụng. Thuật toán cũng như chương trình sẽ được tiếp tục hoàn thiên trong thời gian tới.

5. KẾT LUẬN

Bước đầu chương trình thử nghiệm chứng tỏ được ưu điểm của phương pháp đưa ra trong bài báo. Để hoàn thiên, một số các tiêu chí sẽ tiếp tục được nghiên cứu đưa vào chương trình, đồng thời các trọng số sẽ được chỉnh theo các mẫu học để đi tới các tham số chuẩn hơn. Ngoài ra, khâu tách từ cũng sẽ được hoàn thiện để có một phần mềm có thể đưa vào

sử dụng trong thực tế.

TÀI LIỆU THAM KHẢO

[1] N. C. Ho and W. Wechler, Hedge algebras, An algebraic approach to structures of sets of linguistic domains of linguistic truth variable, *Fuzzy Sets and Systems* **35** (3) (1990) 281–293.

[2] N. Cat Ho and W. Wechler, Extended hedge algebras and their application to Fuzzy logic, *Fuzzy Sets and Systems* **52** (1992) 259–281.

[3] N. C. Ho, T. T. Son, T. D. Khang and L. X. Viet, Fuzzyness measure, quantified semantic mapping and interpolative method of approximate reasoning in medical expert systems, *Journal of Informatics and Cybernetics* **18** (3) (2002) 237–252.

[4] Nguyễn Cát Hồ, Trần Thái Sơn, Logic mờ và quyết định mờ dựa trên cấu trúc thứ tự của giá trị ngôn ngữ, *Tạp chí Tin học và Điều khiển học* **4** (1983).

[5] Nguyễn Cát Hồ, Trần Thái Sơn, Về khoảng cách giữa các giá trị của biến ngôn ngữ trong Đại số gia tử và bài toán sắp xếp mờ, *Tạp chí Tin học và Điều khiển học* (1) (1995) 10–20.

Nhận bài ngày 3 - 12 - 2007
Nhận lại sau sửa ngày 24 - 1 -2008