MỘT MÔ HÌNH TÌM KIẾM THÔNG TIN VĂN BẢN TRONG THƯ VIỆN SỐ

ĐỖ QUANG VINH

Bộ môn Công nghệ Thông tin –Trường Đại học Văn hóa Hà Nội

Tóm tắt: Xếp hạng thường không xảy ra ngoại tuyến và không được chú ý đến. Như với động cơ tìm kiếm trên World Wide Web, các tài liệu hoặc tóm tắt tài liệu được hiển thị trên một máy trạm và khi xuất hiện, người dùng tin có thể chấp nhận các tài liệu liên quan và không chấp nhận tài liệu không liên quan. Động cơ tìm kiếm có thể đánh giá lại xếp hạng định kỳ, hoặc thậm chí sau mỗi một quyết định của người dùng tin, nâng hạng tài liệu giống với các tài liệu được chấp nhận và giảm hạng tài liệu giống với các tài liệu không được chấp nhận. Ở đây, chúng tôi khảo sát một mô hình tìm kiếm thông tin văn bản xác suất trong thư viện số. Nội dung chính của bài báo: 1. Đặt vấn đề; 2. Mô hình tìm kiếm thông tin xác suất ; 3. Sự phản hồi liên quan; 4. Hiệu suất tìm kiếm.

Từ khóa: tìm kiếm thông tin**,** thư viện số.

1. ĐẶT VẤN ĐỀ

Xếp hạng thường không xảy ra ngoại tuyến và không được chú ý đến. Như với động cơ tìm kiếm trên World Wide Web, các tài liệu hoặc tóm tắt tài liệu được hiển thị trên một máy trạm và khi xuất hiện, người dùng tin có thể chấp nhận các tài liệu liên quan và không chấp nhận tài liệu không liên quan. Động cơ tìm kiếm có thể đánh giá lại xếp hạng định kỳ, hoặc thậm chí sau mỗi một quyết định của người dùng tin, nâng hạng tài liệu giống với các tài liệu được chấp nhận và giảm hạng tài liệu giống với các tài liệu không được chấp nhận. Ở đây, chúng tôi khảo sát một mô hình tìm kiếm thông tin văn bản xác suất trong thư viện số.

2. MÔ HÌNH TÌM KIẾM THÔNG TIN XÁC SUẤT

Tìm kiếm thông tin IR đề cập đến tổ chức, lưu trữ, tìm kiếm và đánh giá thông tin có liên quan tới nhu cầu thông tin của người sử dụng.

Mô hình IR tổng quát là một cặp bao gồm các đối tượng và một ánh xạ liên kết ("tìm kiếm") một số đối tượng với một đối tượng đại diện cho một truy vấn.

Cho

$$D = \{d_1, d_2, ..., d_M\}, M \geq 2 \qquad (1)$$

là một tập hữu hạn không rỗng đối tượng.

Chú ý: trường hợp M = 1 có thể được xem xét nhưng nó là tầm thường. Các đối tượng tiêu biểu là đại diện.

Cho $\Re$ là một ánh xạ tìm kiếm từ D vào trong lực lượng của nó $\rho(D)$, nghĩa là,

$$\Re : D \rightarrow \rho(D). \qquad (2)$$

Bằng cách kết hợp tập đối tượng D và ánh xạ tìm kiếm $\Re$, chúng tôi định nghĩa cấu trúc tìm kiếm thông tin như sau:

Định nghĩa 1 (cấu trúc tìm kiếm thông tin):

Cấu trúc tìm kiếm thông tin SIR là một bộ 2 $S = <D, \Re>$. (3)

Định nghĩa 1 là một định nghĩa tổng quát: nó không đề cập đến về các dạng riêng biệt của ánh xạ tìm kiếm $\Re$ và đối tượng D. Từ đó, các mô hình IR riêng biệt khác nhau có thể nhận được bằng cách đặc tả D và $\Re$.

Chúng tôi trình bày một định nghĩa thống nhất đối với các mô hình IR dùng SIR.

Định nghĩa 2 (mô hình tìm kiếm thông tin MIR):

Mô hình tìm kiếm thông tin MIR là một SIR S = <D, $\Re$> với 2 thuộc tính sau đây:

(i) $q = \delta \Rightarrow \mu\tilde{a}_i(q, \delta) = 1 \ \forall i, q, \delta$ (tính phản xạ); (4)

(ii) $\Re^1(q) = \{\delta \in D | \ \mu\tilde{a}_i(q, \delta) = \max \mu\tilde{a}_k(q, \delta_k)\} \cap a\alpha_i$, i cố định tùy ý.

trong đó:

+ $T = \{t_1, t_2, ..., t_N\}$ là một tập hữu hạn thuật ngữ chỉ mục, $N \geq 1$;

+ $O = \{o_1, o_2, ..., o_U\}$ là một tập hữu hạn đối tượng, $U \geq 2$;

+ $(D_j)_{j \in J = \{1, 2, ..., M\}}$ là một họ cluster đối tượng, $D_j \in \rho(O)$, $M \geq 2$;

+ $D = \{\delta_j | \ j \in J\}$ là một tập tài liệu, trong đó tập mờ đã chuẩn hóa $\delta_j = \{(t_k , \mu\delta_j(t_k))| \ t_k \in T, k = 1, ..., N\}$, $j = 1, ..., M$, $\mu\delta_j : T \rightarrow S \subseteq [0, 1] \subset \mathbf{R}$ là đại diện cluster của cluster đối tượng D_j. Chẳng hạn, O có thể bao gồm các bài báo, mỗi một trong chúng tự bản thân là một cluster và mỗi một đại diện mờ của cluster là một tài liệu. Ở trường hợp này, nếu tập mờ là một tập chính xác, tài liệu là duy nhất đối với biểu diễn vectơ nhị phân kinh điển. Ví dụ khác, một cluster có thể là một bộ sưu tập bài báo được coi là liên quan với nhau, trong đó đại diện cluster hoặc tài liệu là một đại diện mờ của một trong những bài báo hoặc của một giả-bài báo (trọng tâm, nghĩa là, nó thực sự không phải là một trong những bài báo trong cluster nhưng mô tả tốt toàn bộ nội dung của cluster). Như vậy, về mặt hình thức mô hình IR cluster truyền thống được bao hàm ở đây là một trường hợp đặc biệt.

+ $A = \{\tilde{a}_1, ..., \tilde{a}_C\}$ là một tập hữu hạn tiêu chuẩn, $C \geq 1$, trong đó $\tilde{a}_i = \{((q, \delta_j), \mu\tilde{a}_i(q, \delta_j)) \ | \ \delta_j \in D, j = 1, ..., M\}$, $i = 1, ..., C$ là một quan hệ mờ chuẩn hóa, $\mu\tilde{a}_i : D \times D \rightarrow [0, 1] \subset \mathbf{R}$, $q \in D$ cố định tùy ý. Theo truyền thống, IR kinh điển có thuộc tính phân đôi (lưỡng cực) trong đó có 2 tiêu chuẩn rõ ràng:

(i) có mặt và không có mặt;

(ii) tìm kiếm được thực hiện dựa vào (i).

Chúng ta giả thiết rằng có thể có 1, 2 hoặc nhiều hơn tiêu chuẩn (nghĩa là, liên quan, không liên quan, không thể quyết định được) với mỗi một mức độ khác nhau. Từ đó, chúng ta bắt buộc phải chấp nhận tiêu chuẩn là quan hệ mờ.

+ $a\alpha_i = \{\delta \in D | \ \mu\tilde{a}_i(q, \delta) > \alpha_i\}$, $i = 1, ..., C$ là một α_i-lát cắt tiêu chuẩn mạnh $\tilde{a}_i$, $\alpha_i \geq 0$, $q \in D$ cố định tùy ý;

+ $\Re : D \rightarrow \rho(D)$ là một ánh xạ tìm kiếm. Về mặt hình thức, tìm kiếm nghĩa là liên kết một tập con tài liệu với một truy vấn nếu chúng liên quan với nhau – tuân theo một tiêu chuẩn lựa chọn - đủ mạnh. Từ đó, chúng ta bắt buộc phải xem truy vấn là một tài liệu và tìm kiếm được định nghĩa dùng α-lát cắt.

***Định nghĩa 3 (mô hình tìm kiếm thông tin xác suất PIR)**:*

Mô hình tìm kiếm thông tin xác suất PIR là một MIR S = <D, $\Re$> thỏa mãn điều kiện sau đây:

$$C = 2 \qquad (5)$$

Chúng ta lấy C = 2 là vì ở mô hình IR xác suất truyền thống có 2 tiêu chuẩn: có liên quan và không liên quan.

Định nghĩa 4 (PIR): định nghĩa 3 có thể được định nghĩa lại như sau:

Mô hình tìm kiếm thông tin xác suất PIR là một MIR S = <D, $\Re$>

trong đó: $C = 2$ và $\Re(q) = \{\delta | \mu\tilde{a}_i(q,\delta) \geq \mu\tilde{a}_j(q, \delta)\}$, $j = i + (-1)^{i+1}$, $\mu\tilde{a}_i(q, \delta) > \alpha_i\}$. (6)

***Định nghĩa 5 (mô hình tìm kiếm thông tin xác suất kinh điển)**:*

Cho D là một tập tài liệu, $q \in D$ một truy vấn và P(R|(q, d)) xác suất tài liệu $d \in D$ là có liên quan /không liên quan với truy vấn q tương ứng. Cho $\Re(q)$ là tập tài liệu tìm kiếm đáp ứng truy vấn q. Một tài liệu d được lựa chọn đáp ứng một truy vấn q nếu

$$P(R|(q, d)) \geq P(I|(q, d)) \qquad \text{(Luật quyết định Bayes)} \qquad (7)$$

nghĩa là,

$$\Re(q) = \{d | \ P(R|(q, d)) \geq P(I|(q, d))\} \qquad (8)$$

Chính xác hơn, P(R|(q, d)) và P(I|(q, d)) là xác suất liên đới tới d khi nó được xét có liên quan và không liên quan tới q tương ứng.

Các tài liệu đã lựa chọn có thể được xếp hạng giảm dần của độ liên quan của chúng (nguyên lý xếp hạng theo xác suất). Một giá trị ngưỡng thường được sử dụng.

Đánh giá P(R|(q, d)) và P(I|(q, d)) dựa vào công thức Bayes.

Từ quan điểm toán học, chúng ta chú ý: đối với xác suất có điều kiện P(R|(q, d)) có ý nghĩa R và (q, d) là các thực thể đồng nhất, nghĩa là, chúng là các sự kiện có σ-đại số giống nhau trên một trường sự kiện Ω (khái niệm xác suất của Kolmogoroff vì xác suất này có bản chất thống kê). Tượng tự đối với P(I|(q, d)). Dù ký hiệu P(R|(q, d)) hoặc P(I|(q, d)) được gọi là xác suất liên quan R hoặc không liên quan I của tài liệu d đối với truy vấn q, thực chất nó là xác suất được gán cho tài liệu d để biểu thị độ liên quan hoặc không liên quan tới truy vấn q.

Cho D là một tập đối tượng, một đối tượng cố định bất kỳ $q \in D$ và hai tiêu chuẩn $\tilde{a}_1$ và $\tilde{a}_2$ là liên quan và không liên quan tương ứng. Cho $\mu\tilde{a}_i(q, \delta)$, i = 1, 2 là mức độ mà một đối tượng bất kỳ $d \in D$ thỏa mãn tiêu chuẩn $\tilde{a}_i$ liên quan tới q.

Định nghĩa 6 PIR

Mô hình tìm kiếm thông tin xác suất PIR là một MIR S = <D, ℜ>

trong đó: $\Re(q) = \{\delta | \mu\tilde{a}_1(q,\delta) \geq \mu\tilde{a}_2(q, \delta)\}, \mu\tilde{a}_1(q, \delta) > \alpha_1\}$. (9)

PIR là một trường hợp đặc biệt của MIR (ở định nghĩa 4, ta lấy i =1).

S. Dominich đã chứng minh PIR ở định nghĩa 6 và mô hình tìm kiếm thông tin xác suất kinh điển ở định nghĩa 5 là tương đương [4].

Bookstein và Swanson đã đề xuất một mô hình tìm kiếm, trong đó một số tài liệu được nhận dạng một lần là có liên quan với truy vấn. Ở mô hình xác suất, sự xuất hiện của một thuật ngữ riêng biệt trong một tài liệu được hiểu hoặc là một bằng chứng tài liệu có liên quan hoặc không liên quan. Để thiết lập một trọng số đối với mỗi một thuật ngữ, các xác suất có điều kiện về "có liên quan tới truy vấn, căn cứ vào thuật ngữ xuất hiện" và "không liên quan tới truy vấn, căn cứ vào thuật ngữ xuất hiện" được đánh giá dựa trên một số xét đoán liên quan đã biết. Ở một CSDL có N tài liệu, R của nó có liên quan, giả sử R_t của các tài liệu liên quan chứa thuật ngữ t và thuật ngữ t xuất hiện ở f_t tài liệu. Ở đây, N, f_t và R là các giá trị đối với tập tài liệu huấn luyện nào đó mà đối với nó các xét đoán liên quan đã được quyết định. Chẳng hạn, chúng có thể do trình bày với một NSD một số tài liệu xếp hạng cao nhất từ một truy vấn vòng đầu đã đánh giá dùng cơ chế tìm kiếm khác như phương pháp cosin [18].

Bảng 1 – Các xác suất có điều kiện.

	Số tài liệu		
	Có liên quan	Không liên quan	Tổng
Thuật ngữ t có mặt	R_t	$f_t - R_t$	f_t
Thuật ngữ t vắng mặt	$R - R_t$	$N - f_t - (R - R_t)$	$N - f_t$
Tổng	R	N – R	N

Các xác suất có điều kiện có thể được đánh giá từ bảng 1. Chẳng hạn,

$$P[\text{có liên quan} \mid \text{thuật ngữ t có mặt}] = R_t / f_t \quad (10)$$

và $P[\text{không liên quan} \mid \text{thuật ngữ t có mặt}] = (f_t - R_t) / f_t$

Tương tự,

$$P[\text{thuật ngữ t có mặt} \mid \text{có liên quan}] = R_t / R \quad (11)$$

và $P[\text{thuật ngữ t có mặt} \mid \text{không liên quan}] = (f_t - R_t) / (N - R)$

Từ đó, một trọng số w_t đối với thuật ngữ t nhận được dùng công thức Bayes:

$$w_t = \frac{R_t /(R - R_t)}{(f_t - R_t)/(N - f_t - (R - R_t))} \quad (12)$$

trong đó các giá trị lớn hơn 1 chỉ thị sự xuất hiện của thuật ngữ t nên được lấy như là trợ giúp cho giả thuyết tài liệu là có liên quan, và các giá trị nhỏ hơn 1 chỉ thị sự xuất hiện của thuật ngữ giả thiết tài liệu là không liên quan. Giả sử N = 20 tài liệu được xem xét; R = 13 có liên quan và thuật ngữ t xuất hiện ở R_t = 11 trong số tài liệu liên quan và ở một trong số tài liệu không liên quan; tức là, f_t = 12. Sau đó, trọng số w_t gán cho thuật ngữ này là

$$w_t = \frac{11/(13-11)}{(12-11)/(20-12-(13-11))} = \frac{5.5}{0.17} = 33$$

và thuật ngữ biểu thị rõ sự liên quan vì nó xuất hiện thường xuyên ở các tài liệu liên quan và hiếm gặp ở các tài liệu không liên quan. Tuy nhiên, giả sử để thay thế $R_t = 4$ và $f_t = 7$. Sau đó, $w_t = (4/9)/(3/4) = 0.59$ và tập huấn luyện giả thiết sự xuất hiện của thuật ngữ t ở một tài liệu bị coi là bất lợi nhỏ cho tài liệu đó có liên quan với truy vấn. Một giá trị $w_t = 1$ chỉ thị thuật ngữ là trung lập và xuất hiện ngẫu nhiên qua các tài liệu có liên quan và không liên quan.

Có thể giả thiết sự xuất hiện của các thuật ngữ ở các tài liệu là độc lập, thì trọng số đối với một tài liệu D_d được tính bằng cách nhân trọng số của các thuật ngữ:

$$w(D_d) = \prod_{t \in D_d} w_t \qquad (13)$$

Các tài liệu với trọng số cao được lựa chọn như câu trả lời với truy vấn. Vì tất cả được yêu cầu là thứ tự tài liệu, không phải là giá trị số chính xác của trọng số, thường biểu diễn là một tổng logarit:

$$\sum_{t \in D_d} \log w_t = \sum_{t \in D_d} \log \frac{R_t/(R-R_t)}{(f_t - R_t)/(N - f_t - (R-R_t))} \qquad (14)$$

Ở đây, một kết quả âm chỉ thị tài liệu được dự báo là không liên quan. Một tổng trọng số của 0 chỉ thị có nhiều bằng chứng chống lại sự liên quan là phải bị phạt và tài liệu nên sinh ra bởi một quá trình ngẫu nhiên nào đó.

3. SỰ PHẢN HỒI LIÊN QUAN

Sự phản hồi liên quan là quá trình sửa đổi truy vấn để nâng cao hiệu suất tìm kiếm. Giả sử một truy vấn Q_0 được đưa ra với một hệ tìm kiếm và một số tài liệu được trả lại. Sau đó, NSD khảo sát một số hoặc tất cả chúng và quyết định là chúng có hoặc không liên quan. Trong một môi trường xử lý theo lô, đây là điểm cuối của quá trình – hệ thống cho phép chỉ định các tài liệu có liên quan và sau đó, không thực sự nghi ngờ sự lựa chọn này, NSD làm việc với tập con các tài liệu này. Nhưng nó không cần kết thúc ở đó. Giả sử NSD lựa chọn một số tài liệu và chỉ thị cho hệ thống, ”Tôi thích các tài liệu này, tìm cho tôi các tài liệu tương tự” và lựa chọn các tài liệu khác hoặc “Không, có nhiều tài liệu lạc đề, tôi không muốn thấy bất kỳ tài liệu nào như thế”. Đây là truy vấn tương tác, tiếp tục cho đến khi NSD được thỏa mãn với tập câu trả lời. Nó yêu cầu truy vấn được thích nghi, nhấn mạnh một số thuật ngữ, không nhấn mạnh các thuật ngữ khác và có thể đưa vào một số thuật ngữ mới hoàn toàn đã trích lọc ngoài các tài liệu ưa thích. Một dãy truy vấn $<Q_i>$ được thực hiện hiệu quả, trong đó Q_{i+1} được mong đợi gần hơn với truy vấn “tối ưu” so với Q_i .

Salton, Buckley và Harman đề xuất phương pháp lặp lại truy vấn. Tất cả sử dụng biểu diễn vectơ mô tả ở trên, trong đó tài liệu D_d và truy vấn Q đều được coi là n-vectơ trọng số, trong đó n là số thuật ngữ truy vấn riêng biệt. Chiến lược đơn giản nhất như sau:

$$Q_{i+1} = Q_i - D_n + \sum_{d \in R} D_d \qquad (15)$$

trong đó:

D_n là tài liệu xếp hạng cao nhất không liên quan;

R là tập tài liệu có liên quan.

Chỉ một tài liệu không liên quan được phép phủ định các thuật ngữ trong truy vấn, nhưng tất cả tài liệu liên quan được phép trợ giúp các thuật ngữ mà chúng chứa. Ba quyết định phải được thực hiện.

Thứ nhất, thường hạn chế phép tính trừ vectơ sao cho không một thuật ngữ nào nhận được một trọng số nhỏ hơn 0 – tài liệu không liên quan không được phép cho bất kỳ thuật ngữ có trọng số âm.

Thứ hai, các tài liệu có xu hướng có trọng số khác 0 nhiều hơn nữa so với truy vấn ban đầu Q_0 , như vậy, biểu thức này có thể tạo lập một truy vấn mới với hàng trăm hoặc hàng nghìn thuật ngữ, sẽ là đánh giá đắt. Do đó, nó thường sắp xếp các thuật ngữ trong các tài liệu liên quan theo trọng số giảm dần và chỉ lựa chọn một tập con trong chúng ảnh hưởng đến truy vấn tăng lên Q_{i+1}.

Thứ ba, ở đây mỗi một trong ba thành phần có thể được lấy trọng số để cho có xu hướng Q_{i+1} hoặc gần với Q_i hoặc gần hơn tới các tài liệu liên quan.

Các biểu thức phản hồi tổng quát hơn cho phép một số lớn hơn trong những tài liệu không liên quan ảnh hưởng đến truy vấn mới và bao hàm dự trữ sẵn cho truy vấn ban đầu nhằm ảnh hưởng đến tất cả truy vấn tiếp theo:

$$Q_{i+1} = \pi Q_0 + \omega Q_i + \lambda \sum_{d \in R} D_d + \eta \sum_{d \in I} D_d \qquad (16)$$

trong đó:

π, ω, λ và η là các hằng trọng số (với $\eta \leq 0$);

R là tập con tài liệu có liên quan;

I là một tập con tài liệu không liên quan bởi vì đáp ứng của NSD với các phép lặp truy vấn.

Sự đánh giá các kỹ thuật phản hồi liên quan là phức tạp bởi vì xếp hạng đã duyệt lại, độ chính xác sẽ cao bởi vì hệ thống sẵn sàng cho biết một số tài liệu liên quan và không liên quan. Để đơn giản, thường giả thiết các tài liệu xem xét bởi NSD đơn giản không có mặt trong CSDL ở vòng đánh giá thứ hai và các truy vấn duyệt lại được thực hiện chống lại một CSDL đã duyệt lại. Nhưng bởi vì các tài liệu đã xoá được xếp hạng cao ở vòng thứ nhất, chúng có thể có liên quan và ở vòng thứ hai sự vắng mặt của các tài liệu này chắc chắn nghĩa là độ chính xác giảm dần. Do đó, một truy vấn lặp có thể được báo cáo như là có hiệu năng tìm kiếm kém hơn tại mỗi một đánh giá, mặc dù sự phản hồi đang hoạt động tốt để đưa vào các tài liệu liên quan mới. Sự lựa chọn là bỏ CSDL không đề cập đến khi đánh giá hiệu suất tìm kiếm trong các vòng hai và tiếp theo về thực hiện truy vấn. Ở trường hợp này, hiệu suất tìm kiếm tăng mạnh, chỉ vì các tài liệu được xem xét có liên quan ở một vòng được đẩy lên đỉnh của xếp hạng ở vòng tiếp theo vì nội dung của chúng được bao hàm trong truy vấn duyệt lại.

Kinh nghiệm chỉ ra một vòng của phản hồi thường đem đến truy vấn tốt hơn đáng kể và một vòng hai đem đến lợi ích phụ thêm nhỏ (Harman). Tuy nhiên, có nhiều thay đổi được lựa chọn, bao gồm số tài liệu nên trình bày và các nhân tử trọng số khác nhau bao hàm trong các công thức ở trên và không có một hướng dẫn rõ ràng như về các kỹ thuật với các tham số gì, là tốt nhất đối với bất kỳ tình huống đã cho. Như với lựa chọn độ đo tương tự ở vị trí đầu tiên, các luật phản hồi khác nhau hoá ra là có hiệu quả đối với các loại CSDL khác nhau và các kiểu truy vấn khác nhau.

Ở một mức đơn giản hơn có các sơ đồ phản hồi thực dụng hơn trong đó hệ thống tính toán một danh sách thuật ngữ tuân theo công thức trọng số là quan trọng ở các tài liệu liên quan và sau đó, trình bày với NSD theo thứ tự trọng số. Sau đó, NSD tự do lựa chọn trong số thuật ngữ này, mở rộng truy vấn gốc để bao gồm các từ có thể bị bỏ sót tại thời điểm truy vấn ban đầu được tạo thành.

Tất cả lựa chọn này giả thiết ít nhất một tài liệu liên quan được trích lọc trong khi xử lý truy vấn ban đầu Q_0. Tuy nhiên, dù không có tài liệu nào được tìm thấy, vẫn có một số kỹ thuật có thể được áp dụng để mở rộng truy vấn. Đơn giản nhất là báo cáo trong khi không có câu trả lời nào với truy vấn gốc, NSD có thể được lợi bằng cách thử lại với một mô tả lựa chọn và các từ khác. Chúng ta sử dụng một từ điển đồng nghĩa trực tuyến có ích hơn, hoặc hiển thị một danh sách các từ đồng nghĩa đối với mỗi một thuật ngữ truy vấn và yêu cầu NSD lựa chọn các từ bổ sung được thêm vào truy vấn hoặc tự động mở rộng truy vấn không kiểm tra với NSD.

4. HIỆU SUẤT TÌM KIẾM

Ở đây, chúng tôi trình bày định nghĩa hai độ đo quan trọng về hiệu suất: độ phục hồi và độ chính xác. Cách thông thường nhất mô tả hiệu suất tìm kiếm là tính số tài liệu có liên quan tìm kiếm được và chúng được liệt kê theo hạng như thế nào [3], [6], [7], [11].

4.1 Độ phục hồi và độ chính xác

Định nghĩa 7 (độ chính xác P):

$$P = \frac{N_R}{\kappa} \qquad (17)$$

Định nghĩa 8 (độ phục hồi R):

$$R = \frac{N_R}{N_T} \qquad (18)$$

trong đó:

N_T là tổng số tài liệu có liên quan tới một truy vấn q, $N_T \neq 0$;

$| \Re(q) | = \kappa$ là số tài liệu tìm kiếm được đáp ứng q, $\kappa \neq 0$;

N_R là số tài liệu có liên quan tìm kiếm được.

Chẳng hạn, nếu 50 tài liệu được tìm kiếm trong câu trả lời về truy vấn nào đó và 35 trong chúng có liên quan thì độ chính xác tại 50 là P = 70%.

Nếu ở truy vấn tương tự như trước đó, có 70 tài liệu liên quan thì độ phục hồi tại 50 là R=50%, vì 35/70 của tài liệu liên quan được lựa chọn bên trong 50 tài liệu cao nhất. Độ phục hồi đánh giá sự mở rộng tới tìm kiếm là vét cạn và định lượng mức độ phủ của tập câu trả lời.

Định đề: Tỉ số giữa độ phục hồi và độ chính xác R / P thay đổi tuyến tính đối với κ.

Chứng minh:

$$N_R = R\, N_T = P\, \kappa \Rightarrow R / P = \kappa / N_T \quad \text{(đpcm)}. \qquad (19)$$

Van Rijsbergen đưa ra một tổ hợp có trọng số của độ phục hồi và độ chính xác như sau [14]:

$$1 - ((a \,.\, P \,.\, R)/ (b \,.\, P + R)) \qquad (20)$$

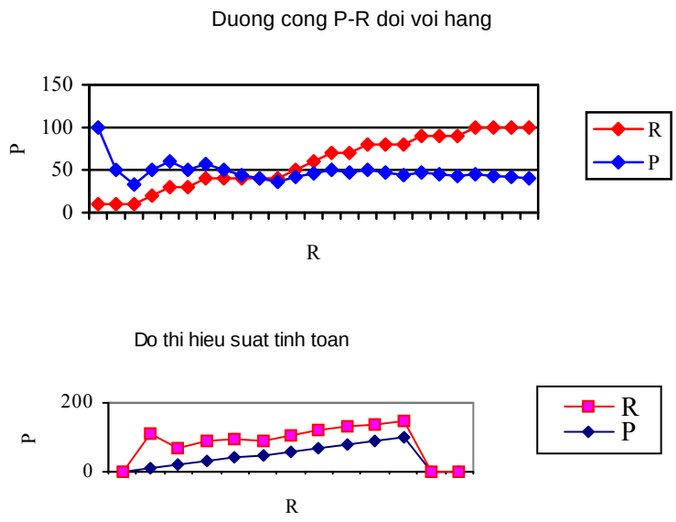
Bảng 2 – Độ phục hồi và độ chính xác.

(a) Hạng;

(a)	r	R (%)	P (%)
	1	10	100
	2	10	50
	3	10	33
	4	20	50
	5	30	60
	6	30	50
	7	40	57
	8	40	50
	9	40	44
	10	40	40
	11	40	36
	12	50	42
	13	60	46
	14	70	50
	15	70	47
	16	80	50
	17	80	47
	18	80	44
	19	90	47
	20	90	45
	21	90	43
	22	100	45
	23	100	43
	24	100	42
	25	100	40

(b) Hiệu suất tính toán.

(b)	R (%)	P (%)
	0	-
	10	100
	20	50
	30	60
	40	57
	50	42
	60	46
	70	50
	80	50
	90	47
	100	45
	TB 3-điểm	53
	TB 11-điểm	61



Hình – Đường cong P-R đối với hạng của bảng 2.

Bảng 2a trình bày một mẫu của tính toán này áp dụng vào một xếp hạng đáp ứng truy vấn nào đó. Cột thứ nhất trình bày thứ tự hạng của tài liệu và cột thứ hai chỉ thị liệu có tài liệu liên quan đến truy vấn không. Đối với bảng 2, giả sử có 10 tài liệu liên quan trong toàn bộ CSDL và 25 tài liệu được tìm kiếm và hiển thị. Dĩ nhiên, tại thời điểm giải thuật được yêu cầu đáp ứng xếp hạng, sự liên quan là không biết; mặt khác, giải thuật có thể đơn giản loại bỏ các tài liệu không liên quan và không bao giờ trình bày chúng. Sự liên quan là một quyết định thực hiện sau khi có đánh giá bởi một hoặc nhiều người. Không nên giả thiết sự liên quan là tuyệt đối. Một người đánh giá có thể đánh giá một tài liệu có liên quan, trong khi người khác đánh giá không liên quan. Người thiết kế các thử nghiệm IR lớn phải xem xét tất cả bài toán và thiết lập một cách thức thử nghiệm hợp lý, không tầm thường.

Cột thứ hai trình bày độ phục hồi - một phần trong số tài liệu liên quan được trả lại. Theo định nghĩa, độ phục hồi không giảm như danh sách hạng được xử lý. Cột cuối cùng ở bảng 2a trình bày độ chính xác tại điểm đó – một phần trong số tài liệu đã tìm kiếm có liên quan. Vì tài liệu thứ nhất có liên quan, độ chính xác tại điểm đó là 100%.

Bảng 2b trình bày xếp hạng của bảng 2a được báo cáo như các giá trị độ phục hồi - độ chính xác chuẩn hoá như thế nào. Cột thứ nhất trình bày 11 điểm độ phục hồi chuẩn từ 0% đến 100%. Đối với mỗi một điểm, cột thứ hai trình bày giá trị độ chính xác tương ứng, đánh giá tại số tài liệu yêu cầu để đạt được mức độ phục hồi đó.

Cuối cùng, 11 giá trị độ chính xác thường được kết hợp thành một tổng giá trị đơn giản đối với hiệu suất tìm kiếm. Có hai cách thực hiện: Thứ nhất, lấy trung bình độ chính xác tại các giá trị phục hồi 20%, 50% và 80%, cho một hiệu suất 3-điểm ở mẫu là 53%. Thứ hai, sử dụng một trung bình 11-điểm, trong đó mức 0% cũng được bao hàm, cho một hiệu suất 11-điểm ở mẫu là 61%.

4.2 Đường cong P-R

Vì độ phục hồi là một hàm không giảm của hạng, độ chính xác có thể được coi là một hàm của độ phục hồi đúng hơn là hàm của hạng. Thật vậy, hiệu suất tính toán được trình bày ở bảng 2b là hiệu quả. Quan hệ được tạo thành ở một đồ thị đã biết như một đường cong P-R, vẽ đồ thị độ chính xác là một hàm của độ phục hồi. Bởi vì độ chính xác thường cao tại các mức độ phục hồi thấp và thấp tại các mức độ phục hồi cao, đường cong nói chung giảm dần. Đường cong P-R đối với mẫu ở bảng 2 được trình bày ở hình trên.

Nếu một giải thuật xếp hạng hoàn chỉnh được phát triển, tất cả tài liệu liên quan nên được xếp hạng trước trong số tất cả tài liệu không liên quan. Ở trường hợp này, độ chính xác bằng 100% tại tất cả mức độ phục hồi và đường cong P-R là một đường nằm ngang tại 100%. Điều này trợ giúp so sánh hai

giải thuật xếp hạng: vẽ đồ thị các đường cong P-R của chúng và nếu một đường cong nằm hoàn toàn phía trên đường cong khác thì giải thuật đó tốt hơn. Không may, tình huống đơn giản này là trường hợp hiếm khi xảy ra và các đường cong thường cắt nhau, có thể vài lần.

TÀI LIỆU THAM KHẢO

[1] W. Abramowicz, *Knowledge-based Information Retrieval and Filtering from Web*, Kluwer Academic Publishers, Boston, 2003.

[2] W.Y. Arms, *Digital Libraries*, MIT Press, Cambridge, 2003.

[3] G.G. Chowdhury., *Introduction to Modern Information Retrieval*, Library Assocciation Publishing, London, 1999.

[4] S. Dominich, *Information Retrieval*, University of Veszprém, Budapest, 2005.

[5] E.A. Fox, *Advanced Digital Libraries*, Virginia Polytechnic Institue and State University, 2000.

[6] R.A. Korfhage, *Information Storage and Retrieval*, John Wiley, New York, 1997.

[7] G. Kowalski, *Information Retrieval Systems*, Kluwer Academic Publishers, Boston, 1997.

[8] A. Large, L.A. Tedd, R.J. Hartley, *Information Seeking in the Online Age*, K.G. Saur Verlag, Munchen, 2001.

[9] W. Mendelhall, T. Sincich, *Statistics for the Engineering and Computer Science*, 2nd Edition, Collier Macmillan, London, 1989.

[10] C.T. Meadow, *Text Information Retrieval Systems*, Academic Press, San Diego, 1992.

[11] S.E. Robertson, M. Beaulieu, *Research and Evaluation in Information Retrieval*, Journal of Documentation, 53(1), 1997, pp. 51-57.

[12] S.M. Ross, *Probability Models for Computer Science*, Harcourt Academic Press, San Diego, 2002.

[13] B.R. Schatz, *Information Retrieval in Digital Libraries*, Science 275, 1997, pp. 327-334.

[14] C.J. Van Rijsbergen, *Information Retrieval*, 2nd Edition, Butterworths, London, 1979.

[15] I.H. Witten, D. Bainbridge, *How to Build a Digital Library*, Morgan Kaufmann, San Francisco, 2003.

[16] W. Wu, H. Xiong, S. Shekhar, *Clustering and Information Retrieval*, Kluwer Academic Publishers, Boston, 2004.

[17] C.T. Yu, W. Meng, *Principles of Database Query Processing for Advanced Applications*, Morgan Kaufmann, San Francisco, 1998.

[18] Đỗ Quang Vinh, *Truy vấn xếp hạng tài liệu văn bản trong thư viện số*, Báo cáo tại Hội thảo Quốc gia một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông lần thứ IX, Đà Lạt, 2006.