2006

NGHIÊN CỨU CÁC PHƯƠNG PHÁP CHỈ SỐ HOÁ VÀ TÌM KIẾM THÔNG TIN VĂN BẢN ỨNG DỤNG TRONG THƯ VIỆN SỐ

Lưu, Thị Thúy Liên Đỗ, Quang Vinh

UEH University

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỖ QUANG VINH

NGHIÊN CỨU CÁC PHƯƠNG PHÁP CHỈ SỐ HOÁ VÀ TÌM KIẾM THÔNG TIN VĂN BẢN ỨNG DỤNG TRONG THƯ VIỆN SỐ

Chuyên ngành: Đảm bảo toán học cho máy tính và hệ thống tính toán
Mã số: 1.01.10

TÓM TẮT LUẬN ÁN TIẾN SỸ TOÁN HỌC

HÀ NỘI - 2006

Công trình được hoàn thành tại:
Trường Đại học Bách khoa Hà Nội

Người hướng dẫn khoa học:
1.TS. QUÁCH TUẤN NGỌC
2. PGS. PHƯƠNG XUÂN NHÀN

Phản biện 1: PGS.TS. HỒ THUẦN
Viện Công nghệ Thông tin

Phản biện 2: PGS.TS. ĐỖ TRUNG TUẤN
Đại học Quốc gia Hà Nội

Phản biện 3: TSKH. NGUYỄN MINH HẢI
Học viện Công nghệ Bưu chính Viễn thông

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp nhà nước họp tại: Trường Đại học Bách khoa Hà Nội
vào hồi giờ ngày tháng năm 2006.

Có thể tìm hiểu luận án tại thư viện:
1. Thư viện Quốc gia Việt Nam.
2. Thư viện Trường Đại học Bách khoa Hà Nội.

MỞ ĐẦU

1. NHIỆM VỤ VÀ PHƯƠNG PHÁP NGHIÊN CỨU

♦ Tính cấp thiết, ý nghĩa lý thuyết và thực tiễn của đề tài

Ngày nay, World Wide Web đã xâm nhập vào cuộc sống hàng ngày, đồng thời, qua một số năm giao diện cho Web tiến triển từ duyệt đến tìm kiếm. Hàng triệu người trên thế giới thực hiện tìm kiếm Web hàng ngày, nhưng công nghệ tìm kiếm cơ sở dữ liệu tài liệu lớn ít thay đổi từ những năm 1980. Sự nhận thức chung về Net tạo ra một cuộc cách mạng mới về công nghệ tìm kiếm thông tin trong thư viện số (DL), diễn ra theo cuộc cách mạng phần cứng ở máy tính cá nhân.

Hiện nay, DL là một trong những hướng nghiên cứu chính về công nghệ thông tin trên thế giới.

♦ **Nhiệm vụ của luận án**: Nghiên cứu các phương pháp chỉ số hoá và tìm kiếm thông tin văn bản ứng dụng trong thư viện số.

♦ **Các phương pháp nghiên cứu**: Hệ cơ sở dữ liệu Multimedia; các phương pháp chỉ mục; các phương pháp mã hoá; các phương pháp nén dữ liệu; các phương pháp tìm kiếm thông tin; các phương pháp xác suất và thống kê toán học.

2. CẤU TRÚC LUẬN ÁN

- Phần mở đầu: trình bày nhiệm vụ, đối tượng, phương pháp nghiên cứu và tóm tắt các đóng góp chính của luận án.
- Chương 1 trình bày tổng quan về thư viện số, đề xuất một mô hình hình thức cho thư viện số dựa vào đại số hiện đại.
- Chương 2 trình bày hai phương pháp chính chỉ mục tài liệu văn bản trong thư viện số, phân tích chi tiết phương pháp chỉ mục tệp đảo IFID, các mô hình nén toàn cục và mô hình nén

cục bộ hyperbol IFID, đề xuất mô hình nén cục bộ Bernoulli và nén nội suy IFID.

- Chương 3 trình bày mô hình tìm kiếm thông tin kinh điển: mô hình truy vấn Boole BQ, đề xuất một mô hình truy vấn xếp hạng tài liệu RQ trong thư viện số, đánh giá hiệu suất tìm kiếm dựa vào hai tham số: độ chính xác P và độ phục hồi R.
- Chương 4 trình bày các giải thuật kinh điển: đảo dựa vào bộ nhớ, đảo dựa vào sắp xếp, đề xuất các giải thuật trộn nhiều đường tại chỗ dựa vào sắp xếp và giải thuật phân chia dựa vào văn bản, so sánh các giải thuật đảo, trình bày bài toán chỉ mục CSDL động.
- Phần kết luận: trình bày các kết luận của luận án và các hướng nghiên cứu tiếp theo.

CHƯƠNG 1 - TỔNG QUAN VỀ THƯ VIỆN SỐ

1.1 MỞ ĐẦU

Định nghĩa 1.1 (Arms W.Y.) [31]: Thư viện số là một kho thông tin có tổ chức với các dịch vụ liên kết, trong đó thông tin được lưu trữ ở dạng số và có thể truy cập qua một mạng.

Định nghĩa 1.2 (Chen H., Houston A.L.) [43]: Thư viện số là một thực thể liên quan tới sự tạo ra các nguồn tin và sự hoạt động thông tin qua các mạng toàn cầu. DL là một kho thông tin số có tổ chức.

Định nghĩa 1.3 (Reddy R., Wladawsky-Berger I.) [121]: Thư viện số là các kho dữ liệu mạng về tài liệu văn bản số, ảnh, âm thanh, dữ liệu khoa học và phần mềm là lõi của Internet hiện nay và các kho dữ liệu số có thể truy cập phổ biến về tất cả tri thức của loài người trong tương lai.

Định nghĩa 1.4 (Sun Microsystems) [135]: Thư viện số là sự mở rộng điện tử về các chức năng điển hình NSD thực hiện và các tài nguyên NSD truy cập trong thư viện truyền thống. Các tài nguyên thông tin được chuyển thành dạng số, lưu trữ trong các kho multimedia và làm cho sẵn có thông qua các dịch vụ Web.

Định nghĩa 1.5 (Witten I.H., Bainbridge D.) [154]: Thư viện số là các kho đối tượng số, bao gồm văn bản, video và audio cùng với các phương pháp truy cập và tìm kiếm, lựa chọn, tổ chức và bảo trì.

Tóm lại, thư viện số là một kho thông tin số khổng lồ có tổ chức với các dịch vụ liên kết qua mạng.

1.2 CÁC KHÁI NIỆM CƠ BẢN

Tác giả trình bày các khái niệm cơ bản trong DL: Cơ sở dữ liệu tài liệu, máy tính và mạng.

1.3 NGHIÊN CỨU TIN HỌC TRONG THƯ VIỆN SỐ

Tác giả trình bày các chủ đề nghiên cứu tin học chính trong DL: Mô hình đối tượng, giao diện người sử dụng, tìm kiếm thông tin, quản trị và bảo trì CSDL, tính liên tác.

1.4 MÔ HÌNH HÌNH THỨC CHO THƯ VIỆN SỐ

1.4.1 Cơ sở toán học

Tác giả xét cơ sở toán học cần thiết để phát triển mô hình hình thức cho DL. Các khái niệm bao gồm tập hợp, quan hệ, hàm, dãy, bộ, xâu, đồ thị và văn phạm [1], [3], [4], [7], [8], [9], [13], [144], [147], [150].

1.4.2 Dòng

Định nghĩa 1.14: Một *dòng* là một dãy có miền giá trị là một tập không rỗng.

1.4.3 Cấu trúc

Định nghĩa 1.15: Một *cấu trúc* là một bộ (G,L,F), trong đó G=(V,E) là một dồ thị có hướng với tập đỉnh V và tập cạnh E, L là một tập giá trị nhãn và F là một hàm gán nhãn $F:(V\cup E)\rightarrow L$.

1.4.4 Không gian

Định nghĩa 1.23: Một *không gian* là một không gian đo được, không gian độ đo, không gian xác suất, không gian vectơ hoặc một không gian topo.

1.4.5 Kịch bản

Định nghĩa 1.26: Một *kịch bản* là một dãy sự kiện chuyển trạng thái liên quan (e_1, e_2, ... , e_n) trên tập trạng thái S sao cho $e_k = (s_k, s_{k+1})$ đối với $1 \le k \le n$.

1.4.6 Cộng đồng

Định nghĩa 1.29: Một *cộng đồng* là một bộ (C, R), trong đó:

1. $C = \{c_1 , c_2, ... , c_n\}$ là một tập của các cộng đồng khái niệm, mỗi một cộng đồng quy về một tập cá thể có cùng lớp hoặc kiểu;

2. $R = \{r_1 , r_2, ... , r_n\}$ là một tập quan hệ, mỗi một quan hệ là một bộ $r_j = (e_j, i_j)$ trong đó e_j là một tích Đề các $c_{k_1} \times c_{k_2} \times ... \times c_{k_{n_j}}$, $1 \le k_1 < k_2 < ... < k_{n_j} \le n$, định rõ các cộng đồng bị dính vào quan hệ và i_j là một hoạt động (xem định nghĩa 1.26) mô tả tương tác hoặc truyền thông giữa các cá thể.

1.4.7 Định nghĩa hình thức thư viện số

Ở đây, tác giả tiếp cận bài toán bằng cách định nghĩa một thư viện số "tối thiểu", nghĩa là, tập tối thiểu các thành phần tạo nên một thư viện số.

Định nghĩa 1.35: Cho C là một CSDL với bộ điều khiển H. Một *mục lục siêu dữ liệu* MC đối với C là một tập cặp {(h,

$\{mc_1, mc_2, ..., mc_{k_n}\})\}$, trong đó $h \in H$ và mc_i là siêu dữ liệu mô tả.

Định nghĩa 1.36: Cho C là một CSDL với bộ điều khiển H. Một *kho* là một bộ (R, gt, st, dl), trong đó $R \subset 2^C$ là một họ CSDL (bao gồm $\overline{C}$) và các hàm gt, sto và dl thỏa mãn:

1. gt:H→C ánh xạ một bộ điều khiển h đến một đối tượng số gt(h).
2. sto:CxR→R ánh xạ (do, $\overline{C}$) đến CSDL mở rộng {do}∪$\overline{C}$.
3. dl:HxR→R ánh xạ (h, $\overline{C}$) đến CSDL nhỏ hơn $\overline{C}$-{gt(h)}.

Định nghĩa 1.37: Một *chỉ mục* $I : T \rightarrow 2^H$ là một hàm, trong đó T là một không gian chỉ mục trên một tập thuật ngữ chỉ mục và H là một tập bộ điều khiển. Một *dịch vụ chỉ mục* cài đặt một chỉ mục.

Định nghĩa 1.38: Cho Q là một tập các nhu cầu thông tin của NSD, thường được gọi là *truy vấn*. Cho $M_I : Q \times C \rightarrow R$ là một hàm so khớp, định nghĩa bởi một chỉ mục I, liên kết một số thực với một truy vấn $q \in Q$ và một đối tượng số $do \in C$, chỉ thị đại diện truy vấn so khớp với đối tượng số tốt như thế nào, cả hai theo cấu trúc và nội dung. Một *dịch vụ tìm kiếm* là một tập kịch bản tìm kiếm $\{sc_1, sc_2, ..., sc_t\}$, trong đó đối với mỗi một truy vấn $q \in Q$ có một kịch bản tìm kiếm $sc_k = <e_0, ..., e_n>$ sao cho e_0 là sự kiện bắt đầu gây ra bởi một truy vấn q và sự kiện e_n là sự kiện cuối cùng trả lại các giá trị hàm so khớp $M_I(q, d)$ đối với mọi $d \in C$.

Định nghĩa 1.40: Một *dịch vụ duyệt* là một tập kịch bản $\{sc_1, ..., sc_n\}$ trên một siêu văn bản (nghĩa là các sự kiện được định nghĩa bởi các cạnh của đồ thị siêu văn bản (V_H,E_H)), sao cho sự kiện liên kết duyệt e_i được liên kết với một hàm $TL : V_H \times E_H$

→ ND, cho trước một nút và một liên kết tìm kiếm nội dung của nút đích, nghĩa là, $TL(v_k,e_{k_i})=P(v_t)$ đối với $e_{k_i}=(v_k,v_t)\in E_H$.

Định nghĩa 1.41: Một ***thư viện số*** là một bộ bốn (R, MC, DV, XH) , trong đó:

- R là một kho;
- MC là một mục lục siêu dữ liệu;
- DV là một tập dịch vụ chứa tối thiểu các dịch vụ chỉ mục, tìm kiếm và duyệt;
- XH là một cộng đồng NSD thư viện số.

Kết luận chương 1

- Trình bày tổng quan về DL và các định nghĩa không hình thức về DL của các tác giả khác nhau trên thế giới.
- Đề xuất một mô hình hình thức cho DL dựa vào đại số hiện đại: một thư viện số là bộ bốn (R, MC, DV, XH).

CHƯƠNG 2 - CHỈ MỤC TÀI LIỆU VĂN BẢN

2.1 MỞ ĐẦU

Đối với DL, chúng ta đang nói về dữ liệu lớn, hàng triệu trang văn bản ít có cấu trúc. Nếu không có một chỉ mục có sẵn, chính xác và đầy đủ, việc tìm kiếm thông tin hầu như thất bại.

Tác giả thử nghiệm trên CSDL TREC (Text REtrieval Conference). Đây là một CSDL tài liệu rất lớn, có tổng cộng hơn 2070.29 MB văn bản và 741856 tài liệu.

2.2 CHỈ MỤC TỆP ĐẢO IFID

Định nghĩa 2.2 (Đỗ Trung Tuấn [17]): *Chỉ mục/Chỉ số* là bảng dữ liệu hay cấu trúc dữ liệu dùng để xác định vị trí của các dòng trong tệp theo điều kiện nào đó.

Định nghĩa 2.3 (Folk M.J., Zoellick B., Riccardi G. [6]): *Chỉ mục* là một cách tìm kiếm thông tin.

Định nghĩa 2.4: *Chỉ mục* là một cơ chế nhằm định vị thuật ngữ cho trước trong văn bản [22].

Ở các ứng dụng văn bản, cấu trúc phù hợp đơn giản nhất là *tệp đảo* (IF)/ *tệp mục lục*.

Định nghĩa 2.5 (chỉ mục tệp đảo IFID): Đối với mỗi một thuật ngữ trong từ điển, một IF chứa một *danh sách đảo* (IL) lưu trữ một danh sách con trỏ tới tất cả xuất hiện của thuật ngữ đó trong văn bản chính, trong đó mỗi một con trỏ trong thực tế là số tài liệu mà thuật ngữ đó xuất hiện. IL đôi khi được coi là một *danh sách mục lục* và các con trỏ là *mục lục*.

Bảng 2.2 - Văn bản mẫu; mỗi dòng là một tài liệu.

TÀI LIỆU	VĂN BẢN
1	Information retrieval is searching and indexing
2	Indexing is building an index
3	An inverted file is an index
4	Building an inverted file is indexing

Bảng 2.3 - IF đối với văn bản của bảng 2.2

Số	Thuật ngữ	IL(tài liệu; vị trí)
1	an	(2;4), (3;1), (3;5), (4;2)
2	and	(1;5)
3	building	(2;3), (4;1)
4	file	(3;3), (4;4)
5	index	(2;5), (3;6)
6	indexing	(1;6), (2;1), (4;6)
7	information	(1;1)
8	inverted	(3;2), (4;3)
9	is	(1;3), (2;2), (3;4), (4;5)
10	retrieval	(1;2)
11	searching	(1;4)

2.3 CHỈ MỤC TỆP KÝ SỐ SFID

SFID là phương pháp chỉ mục khác.

- Tệp ký số: SF là một phương pháp xác suất để chỉ mục văn bản. Mỗi một tài liệu có một ký số liên kết/ hoặc bộ mô tả, một xâu bit bắt nội dung tài liệu theo một nghĩa nào đó.
- Tệp ký số bitslice: Sự truy cập SF có thể được tăng nhanh hơn bằng cách dùng kỹ thuật bitslicing, tức là kỹ thuật chuyển vị ma trận bit

2.4 SO SÁNH CÁC PHƯƠNG PHÁP CHỈ MỤC

Tác giả so sánh hai phương pháp chỉ mục chính tài liệu trong DL: chỉ mục tệp đảo IFID và chỉ mục tệp ký số SFID. Từ đó, tác giả rút ra quy luật chỉ mục tài liệu trong DL là: Ở hầu hết ứng dụng, IF thực hiện tốt hơn SF trong phạm vi của cả hai kích thước chỉ mục và tốc độ truy vấn. IF nén chắc chắn là phương pháp chỉ mục hữu ích nhất một CSDL lớn các tài liệu văn bản có độ dài có thể thay đổi.

2.5 CÁC MÔ HÌNH NÉN IFID

2.5.1 Đặt vấn đề

IF nén là phương pháp chỉ mục hữu ích nhất một CSDL lớn các tài liệu văn bản có độ dài có thể thay đổi trong DL. Kích thước của một IF được giảm xuống đáng kể bằng cách nén. Ở đây, tác giả khảo sát các mô hình và phương pháp mã hoá để nén IFID CSDL tài liệu trong DL.

Chìa khoá của bài toán nén là nhận xét mỗi một IL có thể được lưu trữ như một dãy số nguyên tăng dần, không mất tính tổng quát.

2.5.2 Các mô hình nén toàn cục

2.5.2.1 Mô hình không tham số

Mã toàn cục đơn giản nhất là biểu diễn cố định của các số nguyên dương. Mối quan hệ của Shannon giữa độ dài mã lý tưởng l_x và xác suất $\mathcal{P}[x]$ như sau [144]:

$$l_x = - \log \mathcal{P}[x] \qquad (2.3)$$

cho phép phân bố xác suất hàm ý bởi phương pháp mã hoá riêng biệt được xác định.

2.5.2.2 Mô hình Bernoulli toàn cục

Một cách hiển nhiên tham số hoá mô hình và có thể nhận được nén tốt hơn là sử dụng mật độ thực của con trỏ trong IF. Giả thiết tổng số con trỏ *f* được lưu trữ biết trước. Chia *f* cho số thuật ngữ chỉ mục và sau đó cho số tài liệu, coi một xác suất của *f* /(N.n) là bất kỳ tài liệu lựa chọn ngẫu nhiên chứa bất kỳ thuật ngữ lựa chọn ngẫu nhiên. Sau đó, sự xuất hiện con trỏ có thể được mô hình hoá như một quá trình Bernoulli với xác suất này, bằng giả thiết các con trỏ *f* của IF được lựa chọn ngẫu nhiên từ n.N cặp tài liệu-từ có thể trong CSDL.

2.5.3 Các mô hình nén cục bộ

2.5.3.1 Mô hình hyperbol cục bộ

Xác suất khác đối với một mô hình cục bộ là sử dụng một phân bố hyperbol [124], trong đó xác suất $\mathcal{P}[x]$ của một gap x là

$$\mathcal{P}[x] = \mu / x, \text{ đối với } x = 1, 2, ..., m. \qquad (2.10)$$

Phương pháp điển hình cho hiệu năng tốt hơn so với mô hình Bernoulli nhưng cài đặt phức tạp hơn và yêu cầu sử dụng mã hoá số học, như vậy, nó không đưa ra hiệu năng giải mã như nhau [124].

2.5.3.2 Mô hình Bernoulli cục bộ

Nếu tần suất f_t của thuật ngữ t biết trước, một mô hình Bernoulli trên mỗi một IL riêng biệt có thể được sử dụng. Mã Golomb lại được đòi hỏi ít khắt khe hơn về mặt tính toán so với mã hoá số học và cho nén tương tự.

Để khai thác mô hình, cần lưu trữ tham số f_t với mỗi một IL, sao cho giá trị chính xác của b có thể được dùng trong khi giải

mã. Tổng giá thực hiện nhỏ. Mỗi một IL nén dễ dàng được tiếp đầu ngữ với một mã γ đối với f_t – mã γ là một lựa chọn tốt bởi vì hầu hết tần suất có thể được mong đợi nhỏ.

2.5.3.3 Mô hình Bernoulli lệch

Như mã γ, vectơ đối với mã Golomb là V_G = <b, b, b, ...> và bởi vì kích thước *bucket* đều đã sử dụng, một lượng lớn đối xứng lệch của phân bố γ bị mất. Vì vậy, mã Golomb cục bộ chỉ thực hiện ở mép tốt hơn so với mã γ và δ toàn cục.

2.5.3.4 Mô hình nén nội suy

Mặc dù được thúc đẩy như một cơ chế đương đầu với gom nhóm xuất hiện từ, mã V_T vẫn là một mã tĩnh và tương đương với một mô hình bậc 0 đối với d-gap. Sử dụng một mô hình bậc cao hơn cũng cho phép nén nhạy với gom nhóm vì một dãy d-gap nhỏ là bằng chứng rõ ràng d-gap tiếp theo cũng nhỏ. Một cơ chế được giả thiết tham số b đã dùng đối với mỗi một d-gap bằng trung bình của số nào đó của d-gap đã giải mã trước đây. Trong khi hấp dẫn về lý thuyết, lợi ích nén phụ thường nhỏ và bởi vì có nhiều trường hợp hơn được điều khiển, sự cài đặt phức tạp hơn. Một cách tinh tế hơn trong đó có thể nén mỗi một IL nhạy với phân nhóm.

2.5.4 Hiệu năng của các mô hình nén chỉ mục

Các mô hình cục bộ có xu hướng thực hiện nén tốt hơn mô hình toàn cục và không hiệu quả hơn về thời gian xử lý đòi hỏi trong khi giải mã, vì chúng có xu hướng cài đặt phức tạp hơn. Đối với mục đích thực hành, mô hình nén chỉ mục phù hợp nhất là mô hình Bernoulli cục bộ, cài đặt dùng kỹ thuật mã hoá Golomb.

Bảng 2.7 - Nén IF bằng số bit/con trỏ đối với TREC

Mô hình	Số bit/con trỏ
Các mô hình toàn cục	
Đơn nguyên	1918
Nhị phân	20.00
Bernoulli	12.30
γ	6.63
δ	6.38
Các mô hình cục bộ	
Hyperbol	5.89
Bernoulli	5.84
Bernoulli lệch	5.44
Nội suy	5.18

2.6 CÁC HIỆU ỨNG

Tác giả xét các hiệu ứng ảnh hưởng đến chỉ mục tài liệu văn bản trong DL: Gộp dạng chữ, truy gốc từ, từ bỏ qua [31], [94], [102], [154].

Kết luận chương 2

- Phân tích chi tiết hai phương pháp chính chỉ mục tài liệu văn bản trong DL: chỉ mục tệp đảo IFID và chỉ mục ký số SFID
- So sánh 2 phương pháp chỉ mục IFID và SFID, từ đó, rút ra quy luật chỉ mục tài liệu trong DL.
- Phân tích hai mô hình nén tòan cục: mô hình nén không tham số và mô hình nén toàn cục Bernoulli. Tiếp theo, luận án phân tích chi tiết mô hình nén hyperbol cục bộ, từ đó đề xuất các mô hình nén cục bộ Bernoulli và nén nội suy đối với IFID.
- Phân tích các hiệu ứng ảnh hưởng đến kích thước chỉ mục tệp đảo IFID: Gộp dạng chữ, truy gốc từ, từ bỏ qua.

CHƯƠNG 3 - TÌM KIẾM THÔNG TIN

3.1 MỞ ĐẦU

Tác giả khảo sát hai kiểu truy vấn. Thứ nhất là *truy vấn Boole* (BQ) truyền thống. Thứ hai là *truy vấn xếp hạng* (RQ).

3.2 TRUY VẤN BOOLE

Kiểu truy vấn đơn giản nhất là BQ, trong đó các thuật ngữ được tổ hợp với các phép toán AND, OR và NOT [31], [45], [48], [74], [82], [83], [86], [102], [126], [130], [145], [154], [159]. Quá trình truy vấn dùng một IFID là tương đối trực tiếp. Từ vựng được tìm kiếm đối với mỗi một thuật ngữ; mỗi một IL được tìm kiếm và giải mã; và các danh sách được trộn, lấy giao, hợp hoặc bù như thích hợp. Cuối cùng, các tài liệu chỉ mục như vậy được tìm kiếm và hiển thị với NSD như danh sách câu trả lời.

3.2.1 Truy vấn BQ hội

Tác giả khảo sát chi tiết quá trình BQ hội. Giả sử truy vấn là một phép hội, bao gồm các thuật ngữ kết nối với phép toán AND như sau: t_1 AND t_2 AND ... AND t_r

và một BQ hội có r thuật ngữ đang được xử lý.

3.2.2 Truy vấn BQ không hội

Cho đến nay, tác giả chỉ xét kiểu BQ hội. Dạng phổ biến khác là một phép hội của các phép tuyển, trong đó một số lựa chọn được định rõ đối với mỗi một thành phần của nó về cơ bản là một BQ hội: (*text* OR *data* OR *information*) AND

(*search* OR *seek*) AND

(*retrieval* OR *indexing*)

3.3 TRUY VẤN XẾP HẠNG RQ

Cho đến nay, hầu hết các hệ thống tìm kiếm thông tin IR hiện có trong thư viện sử dụng truy vấn Boole BQ, nhưng xử lý không chính xác truy vấn Boole không hội, phức tạp. BQ không phải là phương pháp tìm kiếm thông tin duy nhất. Nếu tập con tài liệu chính xác nào đó đang được tìm kiếm biết trước thì BQ chắc chắn thích hợp, đó là nguyên nhân BQ thành công ở các hệ

thống tìm kiếm thư mục. Tuy nhiên, yêu cầu thông tin thường biết ít chính xác hơn. Vì vậy, nó đôi khi hữu ích có khả năng định rõ một danh sách thuật ngữ chỉ thị tốt các tài liệu có liên quan, dù chúng không cần tất cả có mặt trong tìm kiếm tài liệu. Ở đây, tác giả nghiên cứu gán một độ tương tự cho mỗi một tài liệu theo cách đòi hỏi phải so khớp sát một truy vấn.

3.3.1 So khớp toạ độ

Một cách đưa ra tính linh động hơn so với một câu trả lời có-hoặc-không nhị phân đơn giản là đếm số thuật ngữ truy vấn xuất hiện trong mỗi một tài liệu. Càng nhiều thuật ngữ xuất hiện hơn, càng có nhiều khả năng hơn tài liệu là có liên quan. Cách tiếp cận được gọi là *so khớp toạ độ*. Truy vấn thành một truy vấn lai, trung gian giữa một truy vấn hội AND và một truy vấn tuyển OR: một tài liệu chứa bất kỳ trong số thuật ngữ được xem như một câu trả lời tiềm năng, nhưng sự ưu tiên được cho các tài liệu chứa tất cả hoặc hầu hết chúng. Tất cả thông tin cần thiết nằm trong IF và cài đặt tương đối dễ.

3.3.2 Tích trong độ tương tự

Quá trình được hình thức hoá bằng một tích trong của một vectơ truy vấn với một tập vectơ tài liệu.

Độ tương tự của truy vấn Q với tài liệu D_d được biểu diễn như sau:

$$S(Q, D_d) = Q \cdot D_d \qquad (3.1)$$

trong đó phép toán · là phép tích trong.

Bảng 3.1 – Các vectơ đối với tính toán tích trong:

(a) Vectơ tài liệu; (b) Vectơ truy vấn.

(a)	d	Vectơ tài liệu $W_{d,t}$							
		inf	ret	sea	indexing	bui	index	inv	file
	1	1	1	1	1	0	0	0	0
	2	0	0	0	1	1	1	0	0

	3	0	0	0	0	0	1	1	1
	4	0	0	0	1	1	0	1	1
(b)	searching	0	0	1	0	0	0	0	0
	indexing	0	0	0	1	0	0	0	0

Bài toán thứ nhất có thể được giải quyết bằng cách thay thế đánh giá "có" hoặc "không" nhị phân bằng một số nguyên chỉ thị thuật ngữ xuất hiện bao nhiêu lần trong tài liệu. Số đếm xuất hiện này được gọi là *tần suất bên trong tài liệu* của thuật ngữ $f_{d,t}$

Tổng quát hơn, thuật ngữ t trong tài liệu d có thể được gán một *trọng số tài liệu-thuật ngữ*, ký hiệu là $w_{d,t}$ và trọng số khác $w_{q,t}$ trong vectơ truy vấn. Độ tương tự là tích trong của hai trọng số $w_{d,t}$ và $w_{q,t}$ – lấy tổng của tích các trọng số của các thuật ngữ truy vấn và thuật ngữ tài liệu tương ứng:

$$S(Q, D_d) = Q \cdot D_d = \sum_{t=1}^{n} w_{q,t} \cdot w_{d,t} \qquad (3.3)$$

Bài toán thứ hai không nhấn mạnh đến các thuật ngữ khó tìm. Thực vậy, một tài liệu với đủ lần xuất hiện của một thuật ngữ phổ biến luôn được xếp hạng đầu tiên nếu truy vấn chứa thuật ngữ đó, không kể các từ khác. Điều này có thể được thực hiện bằng cách lấy trọng số thuật ngữ tuân theo *tần suất tài liệu đảo* (IDF) của nó. Giả thiết nhất quán với các quan sát của Zipf. [82], [83]. Zipf quan sát tần suất của một mục có xu hướng là tỉ lệ nghịch với hạng của nó. Tức là, nếu hạng được coi là một độ đo tầm quan trọng thì trọng số w_t của một thuật ngữ t được tính như sau:

$$w_t = \frac{1}{f_t} \qquad (3.5)$$

trong đó: f_t là số tài liệu chứa thuật ngữ t.

Sau đó, các vectơ tài liệu được tính như sau:

$$w_{d,t} = r_{d,t} \qquad (3.8)$$

hoặc $w_{d,t} = r_{d,t} \cdot w_t$ (TF x IDF)

Phương pháp sau nhằm gán các trọng số tài liệu-thuật ngữ được gọi là luật TF x IDF: tần suất thuật ngữ nhân tần suất tài liệu đảo.

Các trọng số truy vấn-thuật ngữ $w_{q,t}$ được tính tương tự.

Giả sử tài liệu và các vectơ truy vấn được mô tả bằng

$$w_t = \log_e(1 + N / f_t)$$

$$r_{d,t} = 1 + \log_e f_{d,t} \qquad r_{q,t} = 1 \qquad (3.9)$$

$$w_{d,t} = r_{d,t} \qquad w_{q,t} = r_{q,t} \cdot w_t$$

Vì vậy, người ta thường dựa vào một nhân tử *chuẩn hoá* để không kể đến phần đóng góp của các tài liệu dài. Do đó, biến thể khác của luật tích trong đánh giá độ tương tự bằng

$$S(Q, D_d) = \frac{\sum_{t \in Q} w_{q,t} \cdot w_{d,t}}{|D_d|} \qquad (3.10)$$

trong đó $|D_d| = \sum_i f_{d,i}$ là *độ dài* của tài liệu D_d nhận được bằng cách đếm số thuật ngữ chỉ mục.

3.3.3 Mô hình không gian vectơ

Bất kỳ trọng số thuật ngữ w_t và các tần suất thuật ngữ tương đối $r_{d,t}$ và tài liệu $r_{q,t}$ được gán và bất kỳ trong số tài liệu-thuật ngữ $w_{d,t}$ và trọng số truy vấn-thuật ngữ $w_{q,t}$ phát sinh do sự gán này, kết quả là giống nhau – mỗi một tài liệu được biểu diễn bởi một vectơ trong không gian n-chiều và truy vấn cũng được biểu diễn bằng một vectơ n-chiều.

Độ tương tự đối với một cặp vectơ là khoảng cách Euclide:

$$S(Q, D_d) = \sqrt{\sum_{t=1}^{n} |w_{q,t} - w_{d,t}|^2} \qquad (3.11)$$

Điều thực sự quan tâm là *hướng* chỉ thị bởi hai vectơ hoặc chính xác hơn sự khác nhau về hướng, không kể độ dài.

Góc θ được tính từ

$$\cos\ \theta\ =\ \frac{X\ \cdot Y}{|X\ \|Y\ |} \qquad (3.14)$$

Luật cosin đối với xếp hạng:

$$\cos(Q, D_d) = \frac{Q \cdot D_d}{|Q \| D_d|} = \frac{1}{W_q W_d} \sum_{t=1}^{n} w_{q,t} \cdot w_{d,t} \qquad (3.15)$$

trong đó: W_d là độ dài Euclide – trọng số – của tài liệu d;

W_q là trọng số của truy vấn.

Có thể sử dụng luật này với bất kỳ phương pháp lấy trọng số thuật ngữ mô tả ở trên. Chẳng hạn, giả sử biến thể mô tả ở phương trình (3.9) được sử dụng. Sau đó, tính độ tương tự được mô tả bằng (3.18):

$$\cos(Q, D_d) = \frac{1}{W_d W_q} \sum_{t \in Q \cap D_d} (1 + \log_e f_{d,t}) \cdot \log_e \left(1 + \frac{N}{f_t}\right)$$

3.4 ĐÁNH GIÁ HIỆU SUẤT TÌM KIẾM

3.4.1 Độ chính xác và độ phục hồi

Đánh giá hiệu suất tìm kiếm dựa vào hai tham số chính sau đây [45], [82], [83], [86], [122], [145], [159].

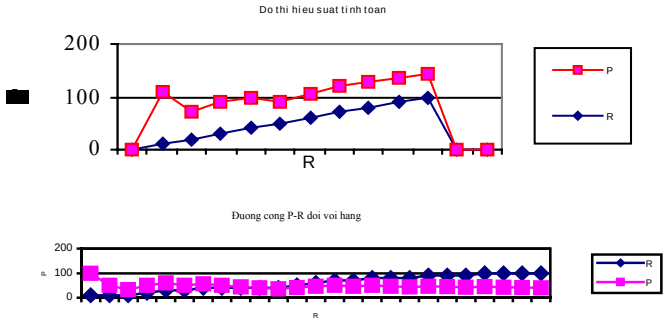
Độ chính xác (precision) P của một phương pháp xếp hạng đối với điểm cắt nào đó r là một phần trong số tài liệu xếp hạng cao nhất r có liên quan đến truy vấn:

$$P = \frac{\text{so tai lieu tim kiem duoc co lien quan}}{\text{tong so tai lieu tim kiem duoc}} \qquad (3.19)$$

Độ phục hồi (recall) R của một phương pháp tại giá trị r nào đó là tỷ lệ của tổng số tài liệu có liên quan được tìm kiếm trong r cao nhất:

$$R = \frac{\text{so tai lieu tim kiem duoc co lien quan}}{\text{tong so tai lieu co lien quan}} \qquad (3.20)$$

3.4.2 Đường cong độ phục hồi-độ chính xác



Hình 3.1 – Đường cong P-R đối với hạng của bảng 3.2

3.5 ĐỘ ĐO COSIN

Tác giả khảo sát độ đo cosin. Rõ ràng, nhiều thông tin hơn được yêu cầu so với xử lý BQ và thực hiện các quyết định về thông tin này nên được cấu trúc như thế nào để làm cho xử lý xếp hạng có hiệu quả trong giới hạn thời gian và bộ nhớ yêu cầu. Các kỹ thuật phát triển ở đây cho phép các RQ được đánh giá trên CSDL lớn dùng không nhiều hơn không gian bộ nhớ và thời gian CPU so với yêu cầu bởi đánh giá BQ.

3.5.1 Tần suất bên trong tài liệu

3.5.2 Tính độ đo cosin

Tác giả đánh giá độ đo cosin dùng luật lấy trọng số TFxIDF. Chiến lược đơn giản nhất là đọc mỗi một tài liệu của CSDL, tính một giá trị cosin cho nó và duy trì một danh sách đã sắp

xếp của các giá trị cosin r cao nhất tìm được tới chừng mực cùng với văn bản của tài liệu tương ứng.

3.5.3 Bộ nhớ dành cho trọng số tài liệu

3.5.4 Sắp xếp

Thành phần cuối cùng của quá trình xếp hạng là sắp xếp.

Kết luận chương 3

- Phân tích chi tiết mô hình tìm kiếm thông tin kinh điển dựa vào truy vấn Boole BQ hiện đang được sử dụng trong hầu hết các hệ thư viện, chỉ ra nhược điểm của truy vấn BQ
- Đề xuất một mô hình tìm kiếm văn bản dựa vào truy vấn xếp hạng RQ có đánh giá hiệu suất dựa vào độ chính xác P và độ phục hồi R.
- Khảo sát chi tiết về độ đo cosin.

CHƯƠNG 4 - GIẢI THUẬT XÂY DỰNG IFID

4.1 MỞ ĐẦU

Tác giả khảo sát bài toán xây dựng chỉ mục tệp đảo IFID, vì đây là dạng chỉ mục thiết thực nhất đối với cả hai truy vấn BQ và RQ.

Bảng 4.1 - Ma trận tần suất đối với văn bản của bảng 2.2

	Thuật ngữ							
	inf	ret	sea	ind	bui	index	inv	fil
1	1	1	-	1	-	-	-	-
2	-	-	-	1	1	1	-	-
3	-	-	-	-	-	1	1	1
4	-	-	-	1	1	-	1	1

Bảng 4.2 - Chuyển vị tương đương của ma trận tần suất

Số	Thuật ngữ	Tài liệu			
		1	2	3	4
1	information	1	-	-	-
2	retrieval	1	-	-	-
3	searching	-	-	-	-
4	indexing	1	1	-	1

5	building	-	1	-	1
6	index	-	1	1	-
7	inverted	-	-	1	1
8	file	-	-	1	1

4.2 GIẢI THUẬT ĐẢO DANH SÁCH MÓC NỐI

Thực tế, một tham khảo chéo chỉ là tên khác đối với một chỉ mục đảo, trong đó mỗi một thuật ngữ của văn bản nào đó được liệt kê theo thứ tự ABC, cùng với một danh sách số dòng xuất hiện trong đó. Thời gian đảo T là:

$T = Bt_r + Ft_p +$ (đọc và phân tích cú pháp văn bản)

$I(t_d + t_r)$ (ghi IF nén)

giây, trong đó các ký hiệu được định nghĩa ở bảng 4.3.

Đối với các CSDL cỡ GB, cách tiếp cận danh sách móc nối là không thích hợp bởi vì nó đòi hỏi hoặc quá nhiều bộ nhớ hoặc quá nhiều thời gian. Tuy nhiên, nó là phương pháp tốt nhất đối với các CSDL nhỏ.

4.3 GIẢI THUẬT ĐẢO DỰA VÀO SẮP XẾP

Vấn đề chính với giải thuật thảo luận ở trên là đòi hỏi quá nhiều bộ nhớ và sử dụng một dãy truy cập dữ liệu chủ yếu là ngẫu nhiên, ngăn cản một ánh xạ hiệu quả từ bộ nhớ lên đĩa. Sự truy cập tuần tự là phương thức xử lý hiệu quả duy nhất đối với các tệp đĩa lớn vì tốc độ truyền thường cao và tìm kiếm ngẫu nhiên mất thời gian. Hơn nữa, sự sử dụng đĩa dường như không thể tránh được đối với lượng dữ liệu đang được xem xét và như vậy, giải thuật đảo nên thực hiện xử lý tuần tự trên bất kỳ tệp đĩa được yêu cầu. Sự xem xét dẫn đến một giải thuật đảo dựa vào sắp xếp [4], [10], [29], [81].

Thời gian thực hiện là:

$T = Bt_r + Ft_p + 10ft_r +$ (đọc và phân tích cú pháp, ghi tệp)

$20ft_r + R(1.2k \log k)t_c +$ (sắp xếp các chương trình)

$[\log R]\,(20ft_r + ft_c)+$ (trộn các chương trình)

$10ft_r + I(t_d + t_r)$ (ghi IF nén)

Yêu cầu không gian đĩa khổng lồ, nghĩa là dù cho phép đảo dựa vào sắp xếp đơn giản là giải thuật tốt nhất đối với CSDL trung bình cỡ khoảng 10÷100 MB, không phù hợp đối với CSDL thực sự lớn cỡ GB.

4.4 GIẢI THUẬT NÉN CHỈ MỤC TRỰC TIẾP

4.4.1 Giải thuật trộn nhiều đường

Bây giờ, quá trình trộn là hướng bộ xử lý hơn so với hướng đĩa và sự giảm hơn nữa về thời gian có thể đạt được bằng cách sử dụng trộn nhiều đường, dẫn đến giải thuật trộn nhiều đường dựa vào sắp xếp được khảo sát bởi Moffat và Bell [108].

Cách tiếp cận có thể được thực hiện sâu hơn. Giả sử tất cả chương trình R được ghi vào tệp tạm thời, tiếp theo nó thực hiện trộn đơn R-đường. Thời gian thực hiện:

$T = Bt_r + Ft_p +$ (đọc và phân tích cú pháp)

$R(1.2k \log k)t_c + I'(t_r + t_d) +$ (sắp xếp, nén và ghi)

$f\,[\log R]tc + I'(t_a/b + t_r + t_d) +$ (trộn)

$I(t_r + t_d)$ (nén lại)

giây, trong đó $b \le M/R$ là kích thước của bộ đệm nhập được cấp phát cho mỗi một chương trình và k, R và I' như ở trên.

4.4.2 Giải thuật trộn nhiều đường tại chỗ

Trong khi phép trộn R-đường mô tả ở trên, 1 bloc b B từ mỗi một chương trình có trong bộ nhớ, cung cấp dự tuyển vào trong heap. Khi bắt đầu trộn, bloc đầu tiên từ mỗi một chương trình được đọc. Mỗi khi bộ ba cuối cùng từ bất kỳ bloc riêng biệt được đưa vào heap, một bloc thay thế được đọc. Giả sử bloc cuối cùng ở mỗi một chương trình được nhồi đến nỗi nó là quá chính xác dài b B. Đệm làm tăng nhẹ kích thước của tệp tạm

thời nhưng nghĩa là mỗi một chương trình nén chiếm một số bloc nguyên; như chúng ta sẽ nhận thấy ngay, điều này cho phép tiết kiệm không gian đáng kể ở chỗ khác.

Thời gian thực hiện là:

$T = Bt_r + Ft_p +$ (đọc và phân tích cú pháp)

$R(1.2k \log k)t_c + I'(t_r + t_d) +$ (sắp xếp, nén và ghi)

$f[\log R]t_c + (I' + I)(t_s/b + t_r + t_d) +$ (trộn và mã hoá lại)

$2I'(t_s/b + t_r)$ (hoán vị)

giây, trong đó $k = (M - L)/10$, $R = [f / k]$, $b < M / (R + 1)$ và I' là kích thước lớn nhất của IF, giả sử $I' = 1.35\ I$.

4.5 GIẢI THUẬT ĐẢO NÉN BÊN TRONG BỘ NHỚ

4.5.1 Giải thuật đảo bộ nhớ lớn

Giả sử một máy có bộ nhớ chính rất lớn. Nếu đối với mỗi một thuật ngữ t tần suất tài liệu f_t là biết rõ khi bắt đầu đảo, một mảng bên trong bộ nhớ lớn có thể được cấp phát chính xác kích thước thích hợp để lưu trữ danh sách của số tài liệu d và tần suất $f_{d,t}$. Thời gian đảo là:

$T = Bt_r + Ft_p +$ (lượt thứ nhất, đọc và phân tích cú pháp)

$Bt_r + Ft_p + 2I'\ t_d + I(t_r + t_d) +$ (lượt thứ hai, đảo)

4.5.2 Giải thuật phân chia dựa vào từ vựng

Giống như giải thuật đảo dựa vào sắp xếp đơn giản, giải thuật "bộ nhớ lớn" chỉ thích hợp đối với các CSDL có kích thước trung bình. Thời gian đòi hỏi là:

$T = Bt_r + Ft_p +$ (đọc và phân tích cú pháp)

$l(Bt_r + Ft_p) + 2I't_d + I(t_r + t_d)$ (xử lý tải)

giây, trong đó l là số tải và $I' = 1.05I$.

4.5.3 Giải thuật phân chia dựa vào văn bản

Cơ sở cho chia nhỏ công việc, giả sử văn bản tự phân chia đúng hơn từ vựng. Thứ nhất, một IF được tạo ra đối với một

chùm tài liệu ban đầu, sau đó, đối với chùm tài liệu thứ hai và v.v, trộn tất cả các IF riêng phần thành một IF cuối cùng. Tác giả nhận thấy một trường hợp có thể thực hiện trộn tại chỗ và ở đây có một ứng dụng tương tự trong đó chiến lược trộn tại chỗ khác có thể được sử dụng. Thời gian thực hiện là:

$T = Bt_r + Ft_p +$ (đọc và phân tích cú pháp)

$Bt_r + Ft_p + 3I't_d + 2cI'(t_s/b + t_r)$ (đảo tại chỗ)

$(I' + I)(t_s/b + t_r + t_d)$ (kết đặc)

giây, trong đó $c = I'/(M - L/3)$ là số chùm văn bản bị cắt thành và như trước đây, $I' \approx 1.05I$ và b là một kích thước bloc phù hợp.

4.6 SO SÁNH CÁC GIẢI THUẬT ĐẢO

Các giải thuật xử lý tốt nhất với một CSDL lớn là giải thuật dựa vào sắp xếp, nhiều đường, trộn, tại chỗ ở mục 4.4.2 và giải thuật phân chia dựa vào văn bản ở mục 4.5.3.

4.7 CƠ SỞ DỮ LIỆU ĐỘNG

Ở trên, tác giả khảo sát các giải thuật chỉ mục với giả thiết CSDL là tĩnh. Tuy nhiên, đối với một CSDL hiếm khi thực sự tĩnh. Vì vậy, bài toán về CSDL động không thể bị bỏ qua. Một CSDL có thể động theo một trong hai cách: mở rộng văn bản hoặc mở rộng chỉ mục.

Kết luận chương 4

- Phân tích chi tiết các giải thuật kinh điển: giải thuật đảo danh sách móc nối và giải thuật đảo dựa vào sắp xếp, từ đó chỉ ra hạn chế của chúng là chỉ thích hợp với các CSDL tài liệu văn bản cỡ nhỏ và vừa.
- Đề xuất hai giải thuật trộn nhiều đường tại chỗ dựa vào sắp xếp và giải thuật phân chia dựa vào văn bản.
- So sánh các giải thuật đảo, từ đó rút ra kết luận hai giải thuật trộn nhiều đường tại chỗ dựa vào sắp xếp và giải thuật

phân chia dựa vào văn bản phù hợp với CSDL tài liệu văn bản cỡ lớn trong thư viện số.

- Khảo sát bài toán CSDL động theo hai cách: mở rộng văn bản và mở rộng chỉ mục.

KẾT LUẬN

Các kết luận được rút ra từ luận án bao gồm:

1. Luận án đề xuất một mô hình hình thức cho thư viện số dựa vào đại số hiện đại: Một thư viện số là một bộ bốn (R, MC, DV, XH) , trong đó:

- R là một kho;
- MC là một mục lục siêu dữ liệu;
- DV là một tập dịch vụ chứa tối thiểu các dịch vụ chỉ mục, tìm kiếm và duyệt;
- XH là một cộng đồng NSD thư viện số.

2. Luận án phân tích chi tiết các phương pháp chỉ mục tài liệu văn bản trong thư viện số: phương pháp chỉ mục tệp đảo IFID và phương pháp chỉ mục ký số SFID, so sánh hai phương pháp chỉ mục, rút ra quy luật chỉ mục tài liệu trong thư viện số là: Ở hầu hết ứng dụng, IF thực hiện tốt hơn SF trong phạm vi của cả hai kích thước chỉ mục và tốc độ truy vấn. IF nén chắc chắn là phương pháp chỉ mục hữu ích nhất một CSDL lớn các tài liệu văn bản có độ dài có thể thay đổi. Luận án phân tích các mô hình nén toàn cục và mô hình nén cục bộ hyperbol, từ đó, đề xuất mô hình nén cục bộ Bernoulli và nén nội suy đối với IFID dựa vào các phương pháp xác suất và thống kê toán học, phương pháp mã hóa, phương pháp nén dữ liệu.

3. Luận án phân tích chi tiết mô hình tìm kiếm thông tin kinh điển dựa vào truy vấn Boole BQ hiện đang được sử dụng

trong hầu hết các hệ thư viện, chỉ ra nhược điểm của truy vấn BQ. Từ đó, luận án đề xuất một mô hình tìm kiếm văn bản dựa vào truy vấn xếp hạng RQ có đánh giá hiệu suất dựa vào độ chính xác P và độ phục hồi R.

4. Luận án phân tích chi tiết các giải thuật kinh điển: giải thuật đảo danh sách móc nối và giải thuật đảo dựa vào sắp xếp, chỉ ra hạn chế của chúng là chỉ thích hợp với các CSDL tài liệu văn bản cỡ nhỏ và vừa. Từ đó, luận án đề xuất hai giải thuật trộn nhiều đường tại chỗ dựa vào sắp xếp và giải thuật phân chia dựa vào văn bản phù hợp với CSDL tài liệu văn bản cỡ lớn trong thư viện số.

Các định hướng nghiên cứu tiếp theo

Tác giả dự định nghiên cứu tiếp theo trong tương lai:

1. Nghiên cứu các phương pháp chỉ mục và tìm kiếm ảnh;
2. Nghiên cứu các phương pháp chỉ mục và tìm kiếm video;
3. Nghiên cứu bài toán tóm tắt và trích rút tài liệu văn bản trong thư viện số.

DANH MỤC CÔNG TRÌNH

1. Đỗ Quang Vinh, Quách Tuấn Ngọc (2001), "Một mô hình dữ liệu hướng đối tượng thời gian đối với tài liệu cấu trúc", *Tạp chí Bưu chính viễn thông & Công nghệ thông tin*, 160(6), tr. 29-32.
2. Đỗ Quang Vinh (2005), "Mô hình nén chỉ mục tệp đảo trong thư viện số", *Kỷ yếu Hội thảo Quốc gia một số vấn đề chọn lọc của công nghệ thông tin và truyền thông lần thứ VIII*, Hải Phòng, tr. 666-674.
3. Đỗ Quang Vinh (2005), "Phương pháp chỉ mục tài liệu trong thư viện số", *Tạp chí Bưu chính viễn thông & Công nghệ thông tin*, 265, tr. 40-47.
4. Đỗ Quang Vinh (2005), "Tóm tắt và trích rút tài liệu văn bản trong thư viện số", *Tạp chí Khoa học và Công nghệ* - Viện Khoa học và Công nghệ Việt Nam, tập 43, số 4, tr.6-14.
5. Đỗ Quang Vinh (2006), "Một phương pháp tìm kiếm thông tin dựa vào mã BCH trong thư viện số", *Tạp chí Khoa học và Công nghệ* - Viện Khoa học và Công nghệ Việt Nam, tập 44, số 1, tr.11-18.
6. Đỗ Quang Vinh (2006), "Truy vấn xếp hạng tài liệu văn bản trong thư viện số", *Kỷ yếu Hội thảo Quốc gia một số vấn đề chọn lọc của công nghệ thông tin và truyền thông lần thứ IX*, Đà Lạt.