MỘT PHƯƠNG PHÁP TÌM KIẾM THÔNG TIN DỰA VÀO MÃ BCH TRONG THƯ VIỆN SỐ

ĐỖ QUANG VINH

I - MỞ ĐẦU

Tìm kiếm thông tin là một chủ đề chính đối với thư viện số. Người sử dụng tìm kiếm tài liệu trong các cơ sở dữ liệu (CSDL) của thư viện số dùng bất kỳ thuật ngữ xuất hiện ở bản ghi và không cần thiết am hiểu về cấu trúc bản ghi hoặc các qui tắc tạo lập bản ghi. Gần đây, các nghiên cứu về thư viện số tập trung vào tìm kiếm thông tin được phân tán trên nhiều máy tính qua mạng [1].

Mục vào trong tệp chỉ mục đối với CSDL tài liệu điển hình bao gồm một bộ nhận dạng tài liệu, cùng với một danh sách bộ mô tả hoặc các thuộc tính mô tả tài liệu riêng biệt. Để thống nhất, các bộ mô tả thường được chọn từ một từ điển lý thuyết của các bộ mô tả chấp nhận được và một cận trên được đặt trên số bộ mô tả có thể được chọn để mô tả bất kỳ tài liệu đơn. Một truy vấn tới một CSDL như thế lại là một danh sách về các bộ mô tả hoặc thuộc tính, nghĩa là, tìm kiếm tất cả tài liệu

a> xuất bản sau năm 1990;

b> xuất bản về tìm kiếm thông tin;

c> xuất bản về lý thuyết mã đại số;

d> xuất bản về mã BCH (Bose-Chaudhari- Hocquenqhem)

trong đó: a, b, c, d nằm trong từ điển.

Để tự động hoá quá trình tìm kiếm, cần mã hoá cả hai dữ liệu tài liệu và truy vấn ở dạng phù hợp đối với xử lý. Ở đây, chúng tôi đề xuất một phương pháp tìm kiếm thông tin nhận được từ cấu trúc đại số của mã sửa lỗi tuyến tính. Nó có ưu điểm ngắn gọn hơn một số phương pháp đã có trước đây và dễ xử lý hơn.

II - MÃ BCH (BOSE - CHAUDHARI - HOCQUENQHEM)

Phát biểu hình thức bài toán như sau:

Cho V là một tập hữu hạn và D là một tập con hữu hạn của V sao cho $\mathrm{Max}_{d\in D}|d| = t$, trong đó |d| ký hiệu số phần tử thuộc d. Cho q là tập con phân biệt của V sao cho $|q| \le t$. Tìm một ký hiệu và một giải thuật phù hợp để xử lý tự động cho phép một trong

a> biểu diễn các phần tử thuộc D;

b> quyết định, đối với mỗi một $d \in D$, liệu $d \supseteq q$ hay không.

Chú ý: ở dạng này, bài toán mô tả một lớp rộng hơn các trạng thái xử lý dữ liệu thực so với chỉ bài toán tìm kiếm thông tin mô tả trước đó.

Chúng tôi đề xuất biểu diễn các phần tử thuộc V bằng các r-bộ nhị phân và các phần tử thuộc D bằng mod 2 tổ hợp tuyến tính của r-bộ này. Vì tập tất cả r-bộ như thế là một không gian vector trên GF(2), như tổ hợp tuyến tính. Để đảm bảo có thể biểu diễn mỗi một tập $d \in D$ rõ ràng, chúng tôi yêu cầu không có hai tổ hợp tuyến tính như thế riêng biệt là bằng nhau. Hình thức hơn: một tập K phần tử của một không gian vector được coi là một *mã chống tuyến tính giải mã được* có độ sâu t (DLS) nếu không có hai tổ hợp tuyến tính riêng biệt của b hoặc có phần tử ít hơn bằng nhau.

Quan hệ giữa các yêu cầu áp đặt lên K bằng định nghĩa này và khái niệm phụ thuộc tuyến tính quen thuộc hơn là khái niệm trong lý thuyết mã, nghĩa là, K là một DSL nếu và chỉ nếu mọi tập con gồm có 2t hoặc ít hơn vector là độc lập tuyến tính.

Tiếp theo, chúng tôi trở lại câu hỏi về tìm kiếm DLS. Các thuộc tính của các tập như thế ở trường hợp tổng quát trong đó các vector là m-bộ phần tử từ một trường hữu hạn tuỳ ý được xem xét tại độ dài bởi Tallini và Segre, nhưng một ít kết quả của họ có thể được áp dụng vào trường hợp nhị phân GF(2). Tuy nhiên, các phương pháp tồn tại nhằm sinh ra các họ lớn của các tập như thế và xuất hiện do lí thuyết mã sửa lỗi tuyến tính.

Một mã tuyến tính nhị phân là một không gian con trong không gian vector n-bộ trên GF(2) và vì vậy là một nhóm con của nhóm cộng tính n-bộ. Vì vậy, nó có thể mở rộng toàn bộ nhóm trong lớp modulo thành mã lớn hơn. Hơn nữa, vì sự lựa chọn một đầu lớp là tuỳ ý bên trong mỗi một lớp, thông thường lựa chọn một phần tử có trọng số cực tiểu như đầu lớp, vì điều này có thể được chỉ ra nhằm cực tiểu hoá xác suất giải mã đúng khi giải mã được dựa trên sự mở rộng lớp. Với quy ước này, đầu lớp phù hợp duy nhất với các lỗi có thể được hiệu chính bằng mã và một mã sửa t-lỗi chính xác ở trường hợp tất cả mẫu có trọng số t hoặc nhỏ hơn xuất hiện như các đầu lớp.

Bây giờ, bất kỳ ma trận kiểm tra tính chẵn lẻ đối với một mã định rõ tính đồng cấu từ các lớp của mã trên không gian vector có (n - k)-bộ trên GF(2), với nhân bằng mã. Như vậy, nếu H là một ma trận kiểm tra tính chẵn lẻ và y là một (n - k)-bộ thì tập tất cả vector x sao cho $xH^T = y$ là một lớp. Quan trọng hơn, nếu x_1 và x_2 là các đầu lớp duy nhất do quy ước trọng số cực tiểu thì $x_1H^T = x_2H^T \Leftrightarrow x_1 = x_2$ Như vậy, không có hai đầu lớp riêng biệt duy nhất x_1 , x_2 sao cho $(x_1 \oplus x_2)H^T = 0$; nghĩa là, nhân của H (tức là mã) không chứa vector có trọng số nhỏ hơn 2t + 1 nếu mã sửa t lỗi. Điều này tạo ra kết quả là vì các vector trong không gian n chiều là các hàm đặc trưng đối với tổ hợp cột của H, không tập có 2t hoặc ít hơn cột của H là phụ thuộc tuyến tính. Vì vậy, các cột của bất kỳ ma trận kiểm tra tính chẵn lẻ đối với một mã sửa t lỗi là một mã chồng tuyến tính giải mã được có độ sâu t.

Như vậy, dựa vào lực lượng của V và lực lượng cực đại của bất kỳ tập con thuộc D, chúng ta có thể tìm được một mã sửa lỗi sao cho các cột của một ma trận kiểm tra tính chẵn lẻ đối với mã đó là đủ để biểu diễn các phần tử của V và các tổ hợp tuyến tính của chúng biểu diễn các phần tử của D. Tính duy nhất của các đầu lớp do quy ước trọng số cực tiểu bảo đảm rằng nó có thể quyết định chỉ từ các đại diện của q và d, dù $d \supseteq q$ hay không. Thực tế, có một phép thử mạnh có thể áp dụng vào mục đích này; phép thử sẽ được mô tả sau đây.

Ví dụ: giả sử $|V| = 16000$ $\quad \underset{d \in D}{\text{Max}} |d| = 10$

Điều này có thể tương ứng với tập tài liệu, mỗi một trong chúng được chỉ mục bằng cách lựa chọn không nhiều hơn 10 thuật ngữ chỉ mục từ một từ vựng có 16000 thuật ngữ. Chúng ta lựa chọn biểu diễn các phần tử của V bằng các cột của một ma trận kiểm tra tính chẵn lẻ đối với một mã BCH [14]. Tham chiếu các bảng về đa thức tối giản ở [14], chúng ta tìm thấy có một mã BCH có độ dài 16383 và trọng số 21 yêu cầu 140 bit kiểm tra tính chẵn lẻ. Như vậy, các cột của ma trận kiểm tra tính chẵn lẻ đối với một mã như thế có độ dài 140 bit và bất kỳ tập lên tới 16383 của các cột này là một DLS 10.

Bây giờ, chúng ta đặt câu hỏi liệu $q \subseteq d$ đối với $d \in D$ đại diện cho một tập con các phần tử cho trước của V sao cho $|q| \leq t$. Theo quan niệm thảo luận trước đó, có thể nhận thấy câu hỏi là liệu đầu lớp tương ứng với q được phủ bởi đầu lớp tương ứng với d hay không. Chúng ta chứng tỏ có thể trả lời câu hỏi này, nhưng ở đây chúng ta yêu cầu một câu trả lời cài đặt nhanh, hiệu quả. Dựa vào (n - k) bộ s_q đại diện cho q, chúng ta biết bất kỳ phần tử y trong lớp tương ứng với s_q là một nghiệm thoả mãn $yH^T = s_q$. Dễ dàng sinh ra một nghiệm thoả mãn hệ này có (n - k) phương trình n ẩn, nhưng điều chúng ta yêu cầu là đầu lớp c_q , là nghiệm có trọng số cực tiểu. Một cách có thể tìm được nghiệm này bằng cách sinh ra lớp từ bất kỳ nghiệm ban đầu, nghĩa là, tạo ra tổng nghiệm ban đầu lần lượt với mỗi một vector của mã và sau đó, tìm kiếm phần tử có trọng số cực tiểu - nhưng đây hầu như không phải là một thủ tục hiệu quả. Hơn nữa, một cách phải thực hiện tương tự đối với mỗi một s_d và sau đó, so sánh mỗi một c_d và c_q . Chúng tôi đề xuất một cách tiếp cận trực tiếp hơn dựa vào bổ đề sau đây.

Bổ đề 1

Cho c_d và c_q là các đầu lớp có trọng số $\le t$ đối với mã sửa lỗi nào đó. Cho w(x) là một hàm ánh xạ toàn không gian vector lên số nguyên như sau: w(x) = trọng số của đầu lớp (một phần tử có trọng số cực tiểu) trong lớp chứa x. Sau đó, $c_d \supseteq c_q$ nếu và chỉ nếu:

$$w(c_d \oplus c_q) + w(c_q) = w(c_d) \qquad (1)$$

Chứng minh:

Điều kiện cần là hiển nhiên.

Đối với điều kiện đủ, giả sử $w(c_q)$ = trọng số của $c_q = x$ và $w(c_d)$ = trọng số của $c_d = y$. Giả sử thêm $c_q \not\subset c_d$ nhưng (1) có hiệu lực. Sau đó, lớp chứa $(c_d \oplus c_q)$ được dẫn bởi c_0 nào đó có trọng số (y - x). Nhưng bây giờ $((c_d \oplus c_q) \oplus c_0)$ là một vector mã $\neq 0$ và hơn nữa

$$\text{trọng số của } ((c_d \oplus c_q) \oplus c_0) \le (x + y) + (y - x)$$
$$\le 2(y)$$
$$\le 2t$$

là không thể có vì mã chứa các vector $\neq 0$ có trọng số $< (2t + 1)$. Như vậy, c_0 không thể có trọng số (y - x) và bổ đề được chứng minh.

Bây giờ, không có các đầu lớp c_d và c_q sẵn có cho chúng ta kiểm thử. Chúng ta có syndrome s_d và s_q thay thế và vì $(c_d \oplus c_q)H^T = (c_d H^T) \oplus (c_q H^T)$ syndrome của $(c_d \oplus c_q)$ mang tên $(s_d \oplus s_q)$. Tuy nhiên, không có vấn đề gì vì từ đó sự tương ứng giữa syndrome và lớp là 1-1, đó là vấn đề trực tiếp định nghĩa một hàm $f(s_x) = w(x)$. Điều này dẫn đến một dạng tương đương của (1):

$$f(s_d \oplus s_q) + f(s_q) = f(s_d).$$

III - THỦ TỤC TÌM KIẾM

Từ bổ đề 1, rõ ràng bài toán tính toán chính trong tìm kiếm là xác định liệu $f(s_{d \oplus q}) = f(s_d) - f(s_q)$. Về hai đại lượng, $f(s_d)$ và $f(s_q)$ có thể được tính dễ dàng (xem phần IV) và thường nhận được bằng cách đưa vào một thẻ ngắn là phần đại diện của tài liệu và truy vấn. Tuy nhiên, sự tính toán về $f(s_{d \oplus q})$ phức tạp hơn.

Theo thuật ngữ mã, tính toán $f(s_{d \oplus q})$ tương đương với tìm kiếm trọng số đầu lớp của một mảng chuẩn từ syndrome của lớp. Zieler và Prange đề xuất một phương pháp dựa vào lý thuyết nhóm. Về nguyên lý, phương pháp của họ làm việc đối với bất kỳ mã tuyến tính, nhưng khối lượng tính toán liên quan phụ thuộc nhiều vào các thuộc tính của mã riêng biệt khi xem xét. Các thủ tục hiệu quả tìm được đối với số mã có độ dài trung bình, bao hàm mã Golay.

Như lựa chọn, chúng tôi coi bài toán tính toán $f(s_{d \oplus q})$ là một bài toán con của bài toán sửa lỗi. Cách tiếp cận rất hiệu quả đối với mã BCH, trong đó một lượng lớn các thủ tục sửa lỗi đã biết [2], [14], [18]. Một phương pháp nhận được từ quan niệm này được trình bày như sau.

Một mã BCH đối với t sửa lỗi được định nghĩa bằng một bộ sinh đa thức g(x) sao cho:

$$G(\alpha^i) = 0 \quad i = 1, 3, \ldots, 2t - 1$$

trong đó α là phần tử gốc của GF(2) đối với một mã có độ dài $2^m - 1$. Syndrome s của bất kỳ lỗi đa thức e(x) được cho bởi $s = [S_1, S_3, \ldots, S_{2t-1}]$, trong đó: $S_i = c(\alpha^i)$ $i = 1, 3, \ldots, 2t - 1$ là tổng luỹ thừa. Tính toán $f(s_{d \oplus q})$ phải xác định trọng số có thể cực tiểu của e(x), mà syndrome của nó là $s_{d \oplus q}$.

Định lí sau đây là thiết yếu đối với thủ tục của chúng tôi.

Định lí Peterson [14]

$$\text{Cho } M_{t_0} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ S_2 & S_1 & 1 & 0 & 0 & \cdots & 0 \\ S_4 & S_3 & S_2 & S_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ S_{2t_0-2} & S_{2t_0-3} & S_{2t_0-4} & S_{2t_0-5} & S_{2t_0-6} & \cdots & S_{2t_0-1} \end{bmatrix}$$

Sau đó, M_{t_0} là không suy biến nếu S_i thuộc M_{t_0} là tổng luỹ thừa của t_0 hoặc t_0 - 1 các phần tử $\neq 0$ riêng biệt của GF(2^m); M_{t_0} là suy biến nếu S_i là tổng luỹ thừa của t_0 - 2 hoặc ít hơn các phần tử $\neq 0$ riêng biệt của GF(2^m).

Theo định lí Peterson : nếu $f(s_{d \oplus q}) \leq t$ thì $f(s_{d \oplus q})$ có thể được xác định bằng cách tính toán $|M_{t_0}|$ đối với t_0 sao cho $t_0 \leq t$. Vì S_i đã biết, $| M_{t_0}|$ có thể được tính dễ dàng đối với $t_0 \leq t$. Tuy nhiên, nếu $f(s_{d \oplus q}) > t$ thì $| M_{t_0}|$ đối với bất kỳ $t_0 \leq t$ không đưa ra một thông tin hữu ích nào.

Cho u là giá trị lớn nhất của t_0 ($t_0 \leq t$) sao cho $| M_{t_0}| \neq 0$. Chúng tôi trình bày một điều kiện cần đối với d để phủ q ở bổ đề sau đây.

Bổ đề 2

Cho $f(s_q) \neq 0$.

Nếu $f(s_{d \oplus q}) = f(s_d) - f(s_q)$ thì $u = f(s_d) - f(s_q) + 1$ (2)

Chứng minh:

Đối với bất kỳ trường hợp không tầm thường $f(s_q) \geq 1$; do đó, giả thiết được thoả mãn. Mặt khác, bằng bổ đề 1, $f(s_{d \oplus q}) = f(s_d) - f(s_q)$ hàm ý $d \supseteq q$, lần lượt hàm ý $f(s_{d \oplus q}) \leq f(s_d) - 1 \leq t - 1$. Đối với bất kỳ e(x) có trọng số $\leq t - 1$, $u - 1 = f(s_{d \oplus q})$ là kết quả của định lí Peterson và bổ đề tiếp theo.

Như vậy, (2) phải được thoả mãn nếu d phải phủ q. Các tài liệu này đối với nó (2) không có hiệu lực không phủ q và có thể không được để ý đến ngay tức thì. Bổ đề 2 có thể được dùng để sắp xếp ra ngoài hầu hết tài liệu không phủ q.

Nhận xét 1: Đối với các hệ thống có thể chịu lỗi một số đáp ứng sai, bổ đề 2 cung cấp một tiêu chuẩn lựa chọn cho tìm kiếm.

Lý thuyết mã cũng cung cấp một câu trả lời hoàn chỉnh đối với hệ thống cho phép không có đáp ứng sai nào. Khi cho rằng điều này kéo theo sự tính toán bổ sung nào đó. Xét phương trình

$$S_i = \sum_{j=1}^{u} X_j^i \qquad i = 1, 3, 5, \ldots, 2u - 1 \qquad (3)$$

Đối với mã sửa t lỗi, (3) có thể được giải đối với X_i trong GF(2) nếu và chỉ nếu S_i là tổng luỹ thừa của t hoặc ít hơn phần tử của GF(2). Kết quả, nếu (3) không thể giải được thì $f(s_{d \oplus q}) > t$. $f(s_{d \oplus q}) > t$ hàm ý sự kiện $q \not\subset d$. Nếu (3) có thể giải được để sinh ra các phần tử riêng biệt u của GF(2) thì nếu

a> $u = t$ và

b> tất cả phần tử $u \neq 0$ thì $f(s_{d \oplus q}) = t$. Đây là trường hợp suy biến, đối với $f(s_{d \oplus q}) = t$ hàm ý $f(s_d) = t$ và $f(s_q) = 0$. Cuối cùng, nếu (3) có thể được giải để sinh ra các phần tử riêng biệt và $u \leq t - 1$ thì $d \supseteq q$.

Chú ý: ở thảo luận trên $s_{d \oplus q}$ được giả thiết là $\neq 0$. Khi $s_{d \oplus q} = 0$ thì $f(s_{d \oplus q}) = 0$ và $d = q$.

Nghiên cứu thủ tục đề xuất kỹ lưỡng, rõ ràng ở đa số trường hợp $u \neq f(s_d) - f(s_q) + 1$ và do đó, thủ tục được kết thúc sau khi tính định thức. Dễ dàng nhận thấy quá trình tính định thức có thể được tổ chức theo cách sao cho ngay khi chúng ta biết định thức lớn nhất bằng 0 hay không, chẳng hạn, phép khử Gauss có thể được dùng, ma trận ở dạng thích hợp để giải nhanh (3). Các kỹ thuật giải (3) được

thảo luận ở các công trình về giải mã BCH [2], [14], [18]. Berlekamp [2] đề xuất một kỹ thuật biến đổi đơn giản làm giảm đáng kể khối lượng tính toán đòi hỏi khi giải (3).

Thủ tục lựa chọn sau đây có thể được. Cho $t_1 = f(s_d) - f(s_q)$. Xét định thức $|M_{t_1}|$. Nếu $|M_{t_1}| = 0$ thì theo bổ đề 1 và định lí Peterson, $q \not\subset d$. Nếu $|M_{t_1}| \neq 0$ thì có thể xác định định thức $f(s_{d \oplus q})$ bằng cách giải (3), với t_1 thay cho u. Nói chung, không rõ ràng là thủ tục lựa chọn này có hiệu quả hơn thủ tục chính mô tả ở trên, nhưng nó dường như khá nhanh khi t_1 nhỏ.

Nhận xét 2: Điều kiện $|M_{t_1}| \neq 0$ là một điều kiện cần đối với $d \supseteq q$: do đó, nó có thể sử dụng là một tiêu chuẩn tìm kiếm ở hệ thống có thể chịu một số đáp ứng sai.

IV - SƠ ĐỒ MÃ CÓ ĐỘ DÀI THAY ĐỔI

Ở đây, chúng tôi mô tả một hệ thống tìm kiếm tài liệu sử dụng ý tưởng về mã có độ dài thay đổi để cực tiểu các yêu cầu hệ thống ở cả lưu trữ lẫn tính toán.

Ở sơ đồ mã có độ dài thay đổi, độ dài syndrome của một vector tài liệu (vector truy vấn) có liên quan tới trọng số của nó w_d (w_q), trong đó

$$s = [S_1, S_2, \ldots, S_{2w_d - 1}]$$

Tương tự đối với w_q . Do đó, độ dài syndrome được sử dụng để tính trực tiếp trọng số của vector.

Khi một truy vấn cho trước, chúng tôi muốn tìm kiếm tất cả tài liệu d sao cho $d \supseteq q$. Rõ ràng, chúng tôi phải có trọng số của vector tài liệu tối thiểu lớn bằng trọng số của truy vấn. Do đó, chỉ cần xét các tài liệu có syndrome dài hơn syndrome của truy vấn. Trường hợp có độ dài bằng nhau có thể chú ý chỉ bằng cách thử $s_{d \oplus q} = 0$.

Khi $w_d > w_q$ chúng tôi có thể tính từ syndrome truy vấn lấy tổng luỹ thừa bổ sung

$$S_{2w_q + 1}, S_{2w_q + 2}, \ldots, S_{2w_d - 1}$$

theo bổ đề sau đây.

Bổ đề 3

Đối với một vector truy vấn có trọng số w_q và bất kỳ $w_d > w_q$, các hàm

$$S_{2w_q + 1}, S_{2w_q + 2}, \ldots, S_{2w_d - 1}$$

có thể được tính từ các hàm $S_1, S_2, \ldots, S_{2w_d - 1}$

Chứng minh:

Các đồng nhất thức Newton đối với GF(2) là

$$S_1 + \sigma_1 = 0$$
$$S_3 + \sigma_1 S_2 + \sigma_2 S_1 + \sigma_3 = 0$$
$$\vdots$$
$$S_{2w_d + 1} + \sigma_1 S_{2w_d - 2} + \ldots + \sigma_{2w_d - 1} = 0$$

Vì trọng số của vector truy vấn là w_q

$$\sigma_{w_q + 1} = \sigma_{w_q + 2} = \ldots = \sigma_{2w_d - 1} = 0$$

và bây giờ, các đồng nhất thức Newton trở thành

$$S_1 + \sigma_1 = 0$$
$$S_3 + \sigma_1 S_2 + \sigma_2 S_1 + \sigma_3 = 0$$
$$\vdots$$

$$S_{2w_d - 1} + \sigma_1 S_{2w_d - 2} + \ldots + \sigma_{w_q} S_{2w_d - w_q - 1} = 0$$

Theo định lí Peterson, phương trình w_q đầu tiên của hệ là độc lập tuyến tính và do đó, nó được sử dụng để giải đối với σ_k ($k = 1, 2, \ldots, w_q$). Phần còn lại của tổng luỹ thừa và σ_k thuộc các phương trình cuối cùng $w_d - w_q$.

Sau khi tổng luỹ thừa bổ sung được tính đối với truy vấn, thủ tục tìm kiếm ở phần III được áp dụng.

Ưu điểm của cách tiếp cận mã có độ dài thay đổi là:

1. Độ dài syndrome của mỗi một tài liệu được cực tiểu hoá, do đó, các yêu cầu lưu trữ được giảm.

2. Lượng tính toán yêu cầu để xác định liệu một tài liệu d phủ truy vấn q cũng được giảm nhiều như bây giờ chúng tôi xử lý với syndrome có độ dài mw_d đúng hơn là mt ($t = \text{Max}\{w_d\}$). Lượng tính toán yêu cầu để nhận được tổng luỹ thừa $S_{2w_q + 1}$, $S_{2w_q + 2}$, ..., $S_{2t - 1}$ là tương đối nhỏ hơn như nó được thực hiện chỉ một lần trong toàn bộ quá trình tìm kiếm về một truy vấn.

3. Hệ thống hoàn toàn linh động vì không một mã riêng biệt nào cần được lựa chọn. Hệ thống chỉ tính toán một số đủ về tổng luỹ thừa để biểu diễn duy nhất một vector có một trọng số nhất định. Biểu diễn không làm thay đổi khi trọng số cực đại của vector tài liệu bị thay đổi.

V- KẾT LUẬN

Phương pháp tìm kiếm dựa vào mã đại số BCH trong thư viện số được trình bày ở bài báo này. Nó có hiệu quả trong tìm kiếm thông tin và không yêu cầu tính toán quá nhiều khi cài đặt.

Nhận xét 3: Theo định lí Peterson, sự thêm các bộ mô tả mới vào hệ thống có thể thực hiện chỉ bởi thêm syndrome gốc vào syndrome tính toán chỉ từ các bộ mô tả mới. Nói riêng, nếu các bộ mô tả mới được thêm một mỗi lần, cập nhật syndrome đặc biệt trở nên đơn giản như $S_i = S_1^i$ *đối với một vector có trọng số 1.*

Cuối cùng, chúng ta nhận xét về ưu điểm tương đối của kỹ thuật này so với các kỹ thuật khác đưa ra đối với bài toán như nhau.

Nhận xét 4: Hai kỹ thuật cạnh tranh chính là

a> mã chồng;

b> biểu diễn tường minh của mỗi một tài liệu bằng một danh sách bộ mô tả.

Ở trường hợp trước, thử nghiệm tìm kiếm là một trong số truy vấn n-bộ phải là tập con của chúng có n-bộ đối với bất kỳ tài liệu tìm kiếm và ở trường hợp sau, thử nghiệm kể cả đối với tập được cài đặt qua so sánh các danh sách bộ mô tả, tương ứng với tài liệu và truy vấn.

Ở trường hợp mã chồng, thực tế hoặc logic trái với hoặc loại trừ (mod 2 cộng) không có một phép đảo duy nhất bắt buộc nó đúng là "tiến đến độ dài lớn" để nhận được tính giải mã được. Kautz và Singleton đề xuất một kỹ thuật đơn giản, chẳng hạn, yêu cầu một độ dài 525 bit để giải một bài toán cùng cỡ với ví dụ mô tả trước đây ($t = 10$, $|V| = 15625$). Như vậy, khi thử nghiệm tìm kiếm tầm thường, các yêu cầu lưu trữ lớn hơn nhiều đối với mã chồng.

Sự so sánh với thao tác danh sách là dễ xử lý hơn nhiều. Độ dài của biểu diễn yêu cầu đối với hai phương pháp có thể so sánh được và ở đây, chúng tôi nhận thấy tính toán định thức cạnh tranh với các thao tác danh sách và so sánh.

TÀI LIỆU THAM KHẢO

1. W.Y. Arms - Digital Libraries, MIT Press, Cambridge, 2003.
2. E.R. Berlekamp - Algebraic Coding Theory, Aegan Park Press, California, 1984.
3. G. Birkhoff, S. MacLane - Tổng quan về đại số hiện đại, 2 tập, Nhà xuất bản Đại học và trung học chuyên nghiệp, Hà Nội, 1979.

4. Csiszar Imre, Korner Janos - Information Theory, Academic Press, Orlando, 1981.
5. E.A. Fox - Advanced Digital Libraries, Virginia Polytechnic Institue and State University, 2000.
6. J. von zur Gathen, J. Gerhard - Modern Computer Algebra, Cambridge University Press, 1999.
7. R.A. Korfhage - Information Storage and Retrieval, John Wiley, New York, 1997.
8. G. Kowalski - Information Retrieval Systems, Kluwer Academic Publishers, Boston, 1997.
9. S. Lawrence, C. Lee Giles - Searching the World Wide Web, Science **280**(3) (1998) 98-100.
10. S. Lawrence, C. Lee Giles - Searching the Web: General and Scientific Information Access, IEEE Communications **37**(1) (1999) 116-122.
11. M. Lesk - Practical Digital Libraries, Morgan Kaufmann, San Francisco, 1997.
12. C.T. Meadow - Text Information Retrieval Systems, Academic Press, San Diego, 1992.
13. P. Perrin, F. Petry - An Information–theoretic based Model for Large-scale Contextual Text Processing, Information Sciences **116** (1999) 229-252.
14. W.W. Peterson - Error-Correcting Codes, MIT Press, Cambridge, 1971.
15. B.R. Schatz - Information Retrieval in Digital Libraries, Science **275** (1997) 327-334.
16. J.C.A. Van der Lubbe - Information Theory, Cambridge University Press, 1997.
17. C.J. Van Rijsbergen - Information Retrieval, 2nd Edition, Butterworths, London, 1979.
18. Nguyễn Thuý Vân - Lý thuyết mã, xuất bản lần 2, Nhà xuất bản Khoa học và kỹ thuật, Hà Nội, 2001.
19. M.J. Wester - Computer Algebra Sytems, John Wiley, New York, 2000.
20. I.H. Witten, D. Bainbridge - How to Build a Digital Library, Morgan Kaufmann, San Francisco, 2003.

SUMMARY

AN METHOD OF INFORMATION RETRIEVAL
BASED ON BCH CODES IN DIGITAL LIBRARIES

Retrieval method based on BCH codes in digital libraries are presented in this paper. It is efficient in information retrieval and they do not require excessive computation in implemention.

In Sections III, the retrieval procedure is discussed on basis of Theorem Peterson and binary BCH codes. We remark that by Lemma 1, addition of new descriptors to the system can be complished simply by addition of the original syndrome to the syndrome computed from the new desciptors only.

For systems that can tolerate some false responses, Lemma 2 provides a selection criterion for retrieval.

We remark that condition $| M_{t_1}| \neq 0$ is a necessary condition for $d \supseteq q$: hence, it can be used as the criterion for retrieval in systems that can tolerate some false responses.

In Section IV, we describe a document retrieval system that utilizes the idea of varable length coding to minimize system requirements in both storage and computation.

Finally, the two principal competing techniques appear to be a> superimposed coding and b> explicit representation of each document by a list of descriptors.