Tạp chí Tin học và Điều khiển học, T. 13, S. 4 (1997) (91–97)

VỀ GIẢI THUẬT SO SÁNH HAI TỪ TRONG THUẬT TOÁN TRUY NGUYÊN ĐỒNG NHẤT

PHƯƠNG MINH NAM[1], LÊ TIẾN VƯƠNG[2]

Abstract. The problems for automatical indentification algorithm in databases with Latinized names has discussed overview in [2], [3], this paper is issuing detail algorithms that compares words of Latinized name following occurrences probability.

Trong các hệ thống thông tin không đầy đủ, việc tìm kiếm, đồng nhất một đối tượng dựa trên một số đặc trưng nào đó luôn là một vấn đề được nhiều nhà nghiên cứu quan tâm và có nhiều hướng giải quyết khác nhau. Các phần mềm cơ sở dữ liệu (CSDL) thương phẩm như SQL Server, Oracle, Sysbase... đều cài đặt các toán tử tìm kiếm thông minh như LIKE, MATCH, SOUND... Một trong các cách tiếp cận được tập trung nghiên cứu là sử dụng lý thuyết tập mờ để biểu diễn và tìm kiếm thông tin khá thành công.

Trong [3] đã đưa ra một cách tiếp cận tìm kiếm mờ dựa trên cơ sở lý thuyết thống kê chọn mẫu và lập ma trận sai sót (ma trận độ đo của sai lệch). Việc đồng nhất hai đối tượng bất kỳ trong một CSDL có thể phải sử dụng tới một số thuộc tính nào đó. Không làm mất tính tổng quát, ở đây chỉ đưa ra cách thức đồng nhất dựa trên việc so sánh hai xâu kí tự biểu diễn họ và tên của đối tượng.

1. ĐẶT BÀI TOÁN

Giả sử rằng thuộc tính họ và tên có độ dài trung bình không quá 24 kí tự.

X, Y là hai giá trị của thuộc tính này với $X = (X_1X_2...X_n)$ và $Y = (Y_1Y_2...Y_n)$, trong đó X_i, Y_j, $i, j < 24$ là kí hiệu cho từng kí tự thuộc xâu X và Y. Gọi $\rho(X, Y)$ là hàm độ đo giữa hai xâu X và Y; δ là một số cho trước. Khi đó nếu $\rho(X, Y) < \delta$ nói rằng X và Y là cùng một lớp theo ngưỡng δ.

Vấn đề cần giải quyết là :

- Xác định ngưỡng độ đo δ.

- Xác định thuật toán chấp nhận được để so sánh hai sâu (hai giá trị họ và tên) theo ngưỡng độ đo đã cho.

Trong bài này chúng tôi quan sát và giải quyết trong trường hợp các giá trị thuộc tính họ và tên thuộc họ La tinh hoặc đã phiên âm ra La tinh.

2. MA TRẬN SAI SÓT VÀ PHÂN LỚP KÝ TỰ

Trong [3], [4] đã trình bày một thuật toán để truy nguyên đồng nhất tên họ, gọi là *thuật toán truy nguyên codfonetics* (truy nguyên theo mã phát âm). Thuật toán này còn khá nhiều khiếm khuyết khi sử dụng ở Việt Nam. Để khắc phục những hạn chế của thuật toán, đã tiến hành các quan sát phân tích trên khoảng 25 ngàn cặp mẫu tên họ, đã lập được ma trận độ đo sai sót (gọi tắt là ma trận sai sót) biểu thị tần suất sai lệch giữa các kí tự trong bảng chữ cái $A \to Z$.

Có thể liệt kê một số nguyên nhân chính dẫn đến sai lệch giữa các kí tự đã nêu như sau:

- Viết tên họ theo phát âm.
- Ngôn ngữ khác nhau dẫn đến cách viết khác nhau.
- Yếu tố tâm lý, khi mệt mỏi thường viết lẫn đặc biệt ở cuối từ.
- Gõ nhầm trên bàn phím.
- Sai số trên đường truyền.

Gọi $P(X_i, X_j)$, $1 \le i, j \le 26$ là độ đo sai sót của kí tự ở vị trí thứ i sang kí tự ở vị trí thứ j trong bảng chữ cái Latin từ A đến Z, $P(X_i, X_j)$ chính là một phần tứ trong ma trận độ đo sai sót. Xác định hàm $R(X_i, X_j)$ như sau:

$$R(X_i, X_j = 1 - \frac{P(X_i, X_j) + P(X_j, X_i)}{2} \tag{2.1}$$

Hàm $K(X_i, X_j)$ đo độ "gần" giữa X_i và X_j xác định như sau:

$$K(X_i, X_j) = \frac{1}{1 + a^* R(X_i, X_j)^2} = \frac{1}{1 + \alpha\left(1 - \frac{P(X_i, X_j) + P(X_j, X_i)}{2}\right)^2} \tag{2.2}$$

Ở đây $\alpha = \dfrac{1}{2^5}$.

Trong [3] đã sử dụng thuật toán phân loại tự động với các công thức (2.1), (2.2) để phân loại bảng chữ cái trong vùng tên họ thành các nhóm, mỗi kí tự trong nhóm có độ sai lệch trung bình sang các kí tự cùng nhóm là lớn nhất so với các kí tự của nhóm khác.

Thực tế kiểm định, phân bảng chữ cái $A \to Z$ thành 8 nhóm với các đại diện sau đây là tối ưu:

Nhóm 1 là A	Nhóm 5 là V
Nhóm 2 là M	Nhóm 6 là I
Nhóm 3 là N	Nhóm 7 là Z
Nhóm 4 là W	Nhóm 8 là B

Các nhóm nói trên được sử dụng để tạo codfonetics cho các từ trong thuộc tính tên họ.

Sai sót theo vị trí và việc chọn ngưỡng

Các kí tự ở từng vị trí khác nhau trong từ, tần suất sai sót cũng khác nhau. Thống kê trên 2 ngàn cặp tên có sai sót, lập bảng tần suất sai sót theo vị trí; ở các vị trí 1 - 3, độ sai lệch dao động ở khoảng 5,1%, sai nhiều nhất ở các vị trí thứ 5 và 6 sai lệch tới 55%.

3. XÁC ĐỊNH NGƯỠNG ĐỘ ĐO

X được coi là trùng với Y nếu:

- 3 kí tự đầu (của X và Y) sai khác nhau không quá 1 kí tự (3.1)
- Sau kí tự thứ 3, sai khác nhau không vượt quá 2 trong mỗi 5 kí tự (3.2)

Kí hiệu s_1 là tổng số 3 kí tự (đầu khác trắng), δ_1 là số kí tự sai sót trong 3 kí tự đầu, s_2 là tổng số kí tự sau kí tự thứ 3 (khác trắng), δ_2 là số kí tự sai trong số xác kí tự thứ 3. Khi đó có thể viết:

$$\delta_1 = \sum_{i=1}^{3} \rho(X_i, Y_i), \qquad \delta_2 = \sum_{i=4}^{n} \rho(X_i, Y_i)$$

và do đó

$$\rho(X, Y) = \sum_{i=1}^{3} \rho(X_i, Y_i) + \sum_{i=4}^{n} \rho(X_i, Y_i) = \delta_1 + \delta_2$$

Hiển nhiên rằng từ 3.1 và 3.2 suy ra $X \equiv Y$ nếu $\delta_1 \leq \frac{1}{3} s_1$ và $\delta_2 \leq \frac{2}{5} s_2$.

Vấn đề còn lại là tìm giải thuật so sánh hai từ X và Y

4. THUẬT TOÁN

Gọi *tam giác lân cận* của kí tự bất kỳ trong hai từ

$$X = \{X_1 X_2 X_3 \ldots X_n\}, \quad Y = \{Y_1 Y_2 Y_3 \cdots Y_m\}, \quad n, m \leq 24$$

là *kí tự của từ này, được xem xét với 3 kí tự trước một, cùng và sau một kí tự của kí tự trong từ kia.*

Ví dụ, *tam giác lân cận của* X_i là (Y_{i-1}, Y_i, Y_{i+1}).

Quy tắc so sánh hai từ được xác lập như sau:

Tiến hành so sánh từng kí tự giữa hai từ.

- Kí tự thuộc từ này, so sánh với kí tự ở từ kia trong phạm vi *tam giác lân cận*, gọi là *giải thuật so sánh chéo nhau*, trước hết là so sánh kí tự ở vị trí thẳng hàng.

- Trường hợp các từ độ dài khác nhau thì so sánh *lân cận* bắt đầu từ vị trí đuôi các từ ngược trở lại đầu từ, gọi là *giải thuật so sánh giật lùi.*

Gọi e_1 là số kí tự sai trong 3 kí tự đầu, s_1 là tổng số số kí tự (khác trắng) trong 3 kí tự đầu, e_2 là số kí tự từ kí tự thứ 4, s_2 là tổng số số kí tự (khác trắng) từ kí tự thứ 4.

Nếu $X \equiv Y$ thì có các biểu diễn sau:

- Với 3 kí tự đầu $\delta_1 \leq \frac{1}{3}s_1$, khi đó $e_1 \leq 1$ nếu $s_1 = 3$

$$e_1 = 0 \text{ nếu } s_1 \leq 2 \tag{4.1}$$

- Các kí tự sau $\delta_2 \leq \frac{2}{5}s_2$, khi đó $e_2 \leq 1$ nếu $3 \leq s_2 \leq 4$

$$e_2 \leq 2 \text{ nếu } 5 \leq s_2 \leq 7 \tag{4.2}$$

(trường hợp tham khảo: $e_2 \leq 1$ nếu $e_1 = 0$ và $0 < s_2 \leq 2$)

Thuật toán so sánh hai từ như sau:

Trường hợp 1: X thuộc trọn trong Y $(X \subset Y)$

Không giảm tổng quát coi như X trùng với Y ở đoạn đầu, từ trái sang sang phải.

$e_1 + e_2 = (m - n)$ và sử dụng quy tắc (4.1) và (4.2) để quyết định xem $X = Y$ hay không?

Trường hợp thứ hai: X và Y bất kỳ

- Nếu $n = m$ (hai từ có độ bằng nhau), thực hiện *giải thuật tìm cặp chéo nhau.*
- Nếu $n \neq m$ (hai từ có độ dài bằng nhau), thực hiện *giải thuật so sánh giật lùi.*

GIẢI THUẬT *tìm cặp chéo nhau*

Bước 1. Ba kí tự đầu tính $\sum_{i=1}^{3} \rho(Xi, Yi)$

So tương ứng X_i *với* Y_i,

nếu $X_i \neq Y_i$, tăng e_1 lên 1 $(e_1 = e_1 + 1)$, tăng i lên 1 $(i = i + 1)$

Với mỗi i, kiểm tra e_1

Nếu $e_1 \geq 2$ thì hai từ khác nhau, kết thúc so sánh

Nếu $i > 3$, *chuyển tới bước 2*

Bước 2. Các kí tự sau kí tự thế 3, tính $\sum_{i=4}^{n} \rho(Xi, Yi)$

Thực hiện giải thuật *tìm cặp kí tự chéo nhau*

Bước 3. So sánh từng kí tự X_i với Y_i

Nếu $X \equiv Y$, tăng i lên 1 $(i = i + 1)$, lặp lại bước 3

Ngược lại $(X_i \neq Y_i)$, tăng e_2 lên 1 $(e_2 = e_2 + 1)$ chuyển tới bước 4

Bước 4. So sánh X_{i+1} ví Y_i

Nếu $X_{i+1} \equiv Y_i$, *có nửa đường chéo trên* (X_{i+1}, Y_i)

Ghi nhận vị trí Y_i

So sánh X_i với Y_{i+1}

Nếu $X_i \equiv Y_{i+1}$, *có nửa đường chéo dưới* (X_i, Y_{i+1})

Ghi nhận vị trí X_i

Ghi nhận có cắp chéo cặp, không tính sai $(e_2 = e_2 - 1)$

Tăng i lên 2 $(i = i + 2)$

Lặp lại bước 3

Nếu $X_i \neq Y_{i+1}$, X_i là kí tự thừa (chèn), coi như sai

Tăng i lên 1 $(i = i + 1)$

Lặp lại bước 3

Nếu $X_{i+1} \neq Y_i$, so sánh X_i với Y_{i+1}

Nếu $X_i \equiv Y_{i+1}$, *có nửa đường chéo trên* (X_i, Y_{i+1})

Y_i là kí tự thứa (chèn)

Ghi nhận vị trí X_i

Tăng i lên 1 $(i = i + 1)$

Lặp lại bước 2

Nếu $X_i \neq Y_{i+1}$, có hai cặp đều sai là (X_i, Y_i) và (X_{i+1}, Y_{i+1})

tăng e_2 lên 2 $(e_2 = e_2 + 2)$

tăng i lên 2 $(i = i + 2)$

Lặp lại bước 3

Trong quá trình tính toán, e_2 luôn được kiểm tra với ngưỡng, nếu vượt quá ngưỡng ở bất cứ bước nào, đều dừng, ngược lại, kết hợp với e_1 để quyết định xem $X \equiv Y$ hay không.

GIẢI THUẬT *so sánh giật lùi*

Nếu $n \neq m$ (hai từ độ dài không bằng nhau), ta chỉ xét các từ hơn kém nhau không vượt quá 2 kí tự.

Không mất tính tổng quát, giả sử $m = n - 1$, tức là X dài hơn Y một kí tự (chính là X_n).

Bước 1. Ba kí tự đầu, tính $\sum_{i=1}^{3} \rho(X_i, Y_i)$

So tương ứng X_i với Y_i như bước 1 thuật toán tìm cặp chéo

Bước 2. Các kí tự sau kí tự thứ 3, tính $\sum_{i=4}^{n} \rho(X_i, Y_i)$

Thực hiện *giải thuật toán so sánh "giật lùi"*,

So sánh X_n với Y_m

Nếu $X_n = Y_m$ thì X_n khôg phải là kí tự thừa

Bỏ cặp kí tự (X_n, Y_m), thay Y_m bằng dấu $*$

Quay lại trường hợp so sánh 2 chuỗi bằng nhau

$(X_1, X_2, \ldots X_m)$ với $(Y_1, Y_2, \ldots Y_m)$, với $Y_m = *$

Nếu $X_n \neq Y_m$ thì X_n là kí tự thừa, bỏ X_n, gán $e_2 = 1$

Quay lại trường hợp so sánh 2 chuỗi bằng nhau

$(X_1, X_2, \ldots X_m)$ với $(Y_1, Y_2, \ldots Y_m)$

5. ĐÁNH GIÁ THUẬT TOÁN

- Khắc phục về cơ bản các hạn chế của thuật toán No. 1 trong [3], nhất là đã tạo codfonetics đủ mịn, phản ánh đúng thực tế quy luật sao sót ngẫu nhiên; mặt khác tận dụng các trường hợp không đủ thông tin nhưng vẫn tìm kiếm để giúp việc xác minh được thực hiện trên các tập hạn chế về số lượng.
- Tốc độ tìm kiếm rất nhanh đối với trường hợp tên họ có độ dài bằng nhau.
- Đã đồng nhất hàng triệu trường hợp, độ chính xác cao.
- Tập hợp được hàng chục ngàn trượng hợp tên họ, năm sinh có sai sót và với số lượng mẫu đồng nhất này, ma trận sai sót không ngừng được hoàn chỉnh và thuật toán liên tục được bổ sung (tự học).
- Thuật toán trên được triển khai cài đặt trong hệ thống thông tin Bộ Nội vụ đạt hiệu quả tốt. Hoàn toàn có giá trị ứng dụng trong các hệ thống thông tin tương tự.

TÀI LIỆU THAM KHẢO

1. Phương Minh Nam, *Phân tích hệ thống và phần mềm hệ thống thông tin XNC*, Báo cáo tổng kết đề tài NCKH cấp nhà nước, mã số 90-68-051.

2. Phương Minh Nam, *Đồng nhất trong cơ sở dữ liệu xuất nhập cảnh*, Tạp chí NCKHCA, T.3 (1997) 38-41.
3. Phương Minh Nam, *Đồng nhất trong cơ sở dữ liệu tên họ Latin*, Tạp chí Tin học và Điều khiển học, T.12, S.2 (1997) 31-41.
4. Nguyễn Doãn Tiến, Trương Đức Thái, Phương Minh Nam, *Về thuật toán đồng nhất theo trọng số*, Kỉ yếu Tuần lễ tin học lần thứ 2 - 1992, tại Tp. Hồ Chí Minh.

(1) Cục Quản lý XNC Bộ Nội vụ.
(2) Viện Điều tra quy hoạch đất đai.

Nhận bài ngày 12-7-1997