

Đổi sánh chuỗi

Nguyễn Hồng Phương
 Email: phuong.nguyenhong@hust.edu.vn
 Site: <http://is.hut.edu.vn/~phuongqh>
 Bộ môn Hệ thống thông tin
 Viện Công nghệ thông tin và Truyền thông
 Đại học Bách Khoa Hà Nội

1

Nội dung

- 1. Giới thiệu
- 2. Phát biểu bài toán
- 3. Phương pháp tính độ tương đồng
 - 3.1. Dựa trên chuỗi
 - 3.2. Dựa trên tập hợp

2

1. Giới thiệu

- Là bài toán tìm các chuỗi trở tới cùng một thực thể trong thế giới thực.
- Ví dụ
 - Chuỗi David Smith trong 1 CSDL có thể chỉ tới cùng một người David R. Smith trong CSDL khác.
 - Chuỗi 1210 W. Dayton St, Madison WI và 1210 West Dayton, Madison WI 53706 cùng chỉ tới một địa chỉ vật lý
- Đối sánh chuỗi đóng vai trò then chốt trong bài toán tích hợp dữ liệu, trích rút thông tin,...

3

2. Phát biểu bài toán

- Cho hai tập chuỗi X và Y; tìm tất cả các cặp chuỗi (x, y) , $x \in X$ và $y \in Y$, sao cho x và y trở tới cùng thực thể thế giới thực.
- Gọi các cặp như vậy là các sánh đôi (match)
- Ví dụ:

Tập X	Tập Y	Sánh đôi
$x_1 = \text{Dave Smith}$	$y_1 = \text{Dave D. Smith}$	(x_1, y_1)
$x_2 = \text{Joe Wilson}$	$y_2 = \text{Daniel W. Smith}$	(x_2, y_2)
$x_3 = \text{Dan Smith}$		

4

Thách thức

- Tính chính xác
 - Lỗi chính tả
 - Định dạng khác nhau
 - Tên khác nhau
 - => thước đo độ tương đồng $s(x,y) \in [0,1]$
- Tính mở rộng
 - Độ tương đồng s mở rộng cho nhiều cặp của 2 tập X và Y => bùng nổ tích Đề-các
 - => chỉ áp dụng $s(x,y)$ với các bộ đôi triển vọng

5

3. Phương pháp tính độ tương đồng

- Độ tương đồng s ánh xạ cặp (x,y) vào 1 giá trị $\in [0,1]$
 - s càng lớn thì x,y càng tương đồng
- Thuật ngữ khác: khoảng cách, chi phí
 - giá trị càng nhỏ thì độ tương đồng càng cao

6

3.1. Độ tương đồng dựa trên chuỗi

- Coi các chuỗi là một dãy tuần tự các kí tự
- Tính toán chi phí biến đổi một chuỗi thành chuỗi kia
- Một số phương pháp
 - Edit Distance
 - Needleman-Wunch
 - Affine Gap
 - Smith-Waterman
 - Jaro
 - Jaro-Winkler

7

Phương pháp Edit Distance

- Còn gọi là khoảng cách Levenshtein
- $d(x,y)$ chi phí tối thiểu biến đổi chuỗi x thành chuỗi y
- Việc biến đổi chuỗi sử dụng các thao tác sau: xóa một kí tự, chèn một kí tự, thay thế một kí tự
- Ví dụ: chi phí biến đổi chuỗi $x=David$ Smiths thành chuỗi $y=Davidd$ Simth là 4
 - Thêm d sau David; thay thế m bởi i ; thay thế i bởi m ; xóa kí tự s cuối cùng
- $d(x,y)=d(y,x)$

8

Phương pháp Edit Distance (tiếp)

- Mọi quan hệ giữa hàm khoảng cách $d(x,y)$ và hàm tương đồng $s(x,y)$

$$s(x, y) = 1 - \frac{d(x, y)}{\max(\text{length}(x), \text{length}(y))}$$

- Ví dụ:

$$s(\text{David Smiths}, \text{Davidd Simth}) = 1 - \frac{4}{\max(12,12)} = 0.67$$

9

Phương pháp Edit Distance (tiếp)

- Giá trị của $d(x,y)$ có thể được tính toán dựa trên quy hoạch động
- Cho
 - $x = x_1x_2...x_n, y = y_1y_2...y_m$
 - x_i và y_j là các kí tự
 - $d(i,j)$: khoảng cách soạn thảo giữa $x_1x_2...x_i$ (tiền tố thứ i của x) và $y_1y_2...y_j$ (tiền tố thứ j của y)
- Ý tưởng: sử dụng biểu thức quay lui, tính $d(i,j)$ từ các giá trị đã tính trước đó của d

10

Phương pháp Edit Distance (tiếp)

- Biến đổi chuỗi $x_1x_2...x_i$ thành chuỗi $y_1y_2...y_j$
 - (a) Biến đổi $x_1x_2...x_{i-1}$ thành $y_1y_2...y_{j-1}$, sau đó copy x_i vào y_j nếu $x_i = y_j$
 - (b) Biến đổi $x_1x_2...x_{i-1}$ thành $y_1y_2...y_{j-1}$, sau đó thay thế x_i bởi y_j nếu $x_i \neq y_j$
 - (c) Xóa x_i , sau đó biến đổi $x_1x_2...x_{i-1}$ thành $y_1y_2...y_j$
 - (d) Biến đổi $x_1x_2...x_i$ thành $y_1y_2...y_{j-1}$, sau đó chèn thêm y_j
- Giá trị $d(i,j)$ là tối thiểu của các chi phí biến đổi ở trên

11

Phương pháp Edit Distance (tiếp)

- Biểu thức quay lui:

$$d(i, j) = \min \begin{cases} d(i-1, j-1) & \text{if } x_i = y_j \\ d(i-1, j-1) + 1 & \text{if } x_i \neq y_j \\ d(i-1, j) + 1 & // \text{delete } x_i \\ d(i, j-1) + 1 & // \text{insert } y_j \end{cases}$$

- Hoặc viết gọn:

$$d(i, j) = \min \begin{cases} d(i-1, j-1) + c(x_i, y_j) & // \text{copy or substitute} \\ d(i-1, j) + 1 & // \text{delete } x_i \\ d(i, j-1) + 1 & // \text{insert } y_j \end{cases}$$

$$c(x_i, y_j) = \begin{cases} 0 & \text{if } x_i = y_j \\ 1 & \text{if } x_i \neq y_j \end{cases}$$

12

Phương pháp Edit Distance (tiếp)

- Chú ý: $d(i,0) = i$ và $d(0,j) = j$
- Ví dụ: tính khoảng cách soạn thảo $d(x,y)$ với $x = "dva"$, $y = "dave"$

		Y ₀	Y ₁	Y ₂	Y ₃	Y ₄
			d	a	v	e
X ₀		0	1	2	3	4
X ₁	d	1	0	1		
X ₂	v	2				
X ₃	a	3				

13

Phương pháp Edit Distance (tiếp)

		Y ₀	Y ₁	Y ₂	Y ₃	Y ₄
			d	a	v	e
X ₀		0	1	2	3	4
X ₁	d	1	0	1	2	3
X ₂	v	2	1	1	1	2
X ₃	a	3	2	1	2	2

x = d-va
y = dave

Thay thế a với e,
chèn a sau d

14

Phương pháp Edit Distance (tiếp)

- Mũi tên đi chéo từ ô (3,4) tới ô (2,3): kí tự x₃ (kí tự a) được copy vào hoặc thay thế bởi kí tự y₄ (kí tự e).
- Mũi tên đi chéo từ ô (2,3) tới ô (1,2): x₂ (kí tự v) được copy vào hoặc thay thế bởi y₃ (kí tự v).
- Mũi tên đi ngang từ ô (1,2) sang ô (1,1): một kí tự cách - được chèn vào x và bắt cặp với kí tự a trong y.
- Quá trình dừng lại khi tới ô (0,0)
- Độ phức tạp tính toán là $O(|x||y|)$

15

- Các phương pháp khác: sinh viên tự tìm hiểu, coi đó là bài tập ở nhà.

16

3.2. Tính độ tương đồng dựa trên tập hợp

- Coi xâu kí tự là tập các đa tập token
- Sử dụng tính chất tập hợp để tính toán điểm tương đồng
- Sinh token từ xâu đầu vào:
 - Cách phổ dụng:
 - Xem xét các từ (phân cách nhau bởi kí tự cách)
 - Loại bỏ từ dừng
 - Ví dụ: xâu "david smith" => tập token {david, smith}

17

- Cách khác: q-grams các xâu con độ dài q có mặt trong xâu ban đầu
 - ví dụ: xâu "david smith" có tập tất cả các 3-grams là {##d, #da, dav, avi, ..., ith, th#, h##}
- Một số phương pháp
 - Overlap
 - Jaccard
 - TF/IDF

18

Phương pháp Overlap

- Cho B_x và B_y là các tập các token sinh ra từ xâu x và xâu y
- Độ overlap trả về số token chung
 $O(x,y) = |B_x \cap B_y|$
- Ví dụ:
 - $x = \text{dave}; y = \text{dav}$
 - 2-grams của x: $B_x = \{\#d, da, av, ve, e\# \}$
 - 2-grams của y: $B_y = \{\#d, da, av, v\# \}$
 - $O(x,y) = 3$

19

Phương pháp Jaccard

- Độ tương đồng Jaccard giữa 2 xâu x và y là $J(x,y) = |B_x \cap B_y| / |B_x \cup B_y|$
- Ví dụ:
 - $x = \text{dave}; y = \text{dav}$
 - $B_x = \{\#d, da, av, ve, e\# \}$
 - $B_y = \{\#d, da, av, v\# \}$
 - $J(x,y) = 3/6$

20

Phương pháp TF/IDF

- TF/IDF liên quan đến lĩnh vực tìm kiếm thông tin: tìm các tài liệu phù hợp với các từ khóa truy vấn.
- Hai xâu là tương đồng nếu chúng chia sẻ các term đặc biệt.
- Ví dụ:
 - $x = \text{Apple Corporation, CA}$
 - $y = \text{IBM Corporation, CA}$
 - $z = \text{Apple Corp}$
 - Phương pháp edit distance và Jaccard sẽ cho $s(x,y)$ cao hơn $s(x,z)$

21

Phương pháp TF/IDF (tiếp)

- Phương pháp TF/IDF có thể nhận ra Apple là term đặc biệt, trong khi CA và Corporation là những cái chung nhiều hơn.
- Các cặp xâu đem đối sánh được lấy ra từ một tập chuỗi
- Biến đổi từng chuỗi thành một túi từ, gọi là tài liệu.
- Ví dụ:
 - $x = \text{aab} \Rightarrow B_x = \{a, a, b\}$
 - $y = \text{ac} \Rightarrow B_y = \{a, c\}$
 - $z = \text{a} \Rightarrow B_z = \{a\}$

22

Phương pháp TF/IDF (tiếp)

- Tính term frequency (TF) và inverse document frequency (IDF):
 - Với mỗi từ t và tài liệu d, $tf(t,d) =$ số lần t xuất hiện trong d
 - Với mỗi từ t, tính $idf(t) =$ tổng số tài liệu trong bộ sưu tập chia cho số tài liệu chứa t
 - IDF càng cao nghĩa là sự xuất hiện của từ càng đặc biệt/khác biệt
- | | |
|---------------|--------------------|
| $tf(a,x) = 2$ | $idf(a) = 3/3 = 1$ |
| $tf(b,x) = 1$ | $idf(b) = 3/1 = 3$ |
| ... | $idf(c) = 3/1 = 3$ |
| $tf(c,z) = 0$ | |

23

Phương pháp TF/IDF (tiếp)

- Tiếp theo, biến đổi từng tài liệu d thành vector đặc trưng v_d
- Hai tài liệu càng tương đồng nếu vector tương ứng của chúng gần nhau
- Vector của d có đặc trưng $v_d(t)$ với mỗi từ t. Giá trị của $v_d(t)$ là hàm của TF và IDF
- v_d có nhiều đặc trưng bằng số term trong bộ sưu tập.

24

Phương pháp TF/IDF (tiếp)

	a	b	c
v_x	2	3	0
v_y	3	0	3
v_z	3	0	0

$x=aab \rightarrow B_x=\{a, a, b\}$
 $y=ac \rightarrow B_y=\{a, c\}$
 $z=a \rightarrow B_z=\{a\}$

$tf(a,x)=2$ $idf(a)=3/3=1$
 $tf(b,x)=1$ $idf(b)=3/1=3$
 \dots $idf(c)=3/1=3$
 $tf(c,z)=0$

- Ba vector v_x, v_y, v_z của ba tài liệu B_x, B_y, B_z
- $v_d(t) = tf(t,d).idf(t)$
- $v_x(a) = 2.1 = 2$

25

Phương pháp TF/IDF (tiếp)

- Tính điểm tương đồng TF/IDF giữa 2 xâu bất kỳ p và q
 - T là tập tất cả các từ trong bộ sưu tập
 - Vector v_p và v_q (của xâu p và q): không gian |T| chiều, mỗi chiều tương ứng với 1 từ
 - Điểm TF/IDF giữa p và q được tính như là cosine góc giữa 2 vector này:

$$s(p, q) = \frac{\sum_{t \in T} v_p(t) \cdot v_q(t)}{\sqrt{\sum_{t \in T} v_p(t)^2} \cdot \sqrt{\sum_{t \in T} v_q(t)^2}}$$

26

Phương pháp TF/IDF (tiếp)

- Ví dụ trên:

- điểm TF/IDF giữa 2 xâu x và y là:

$$s(x, y) = \frac{2.3}{\sqrt{2^2 + 3^2} \cdot \sqrt{3^2 + 3^2}} = 0.39$$

27



28

Lời hay ý đẹp

Đường tuy gần, không đi không bao giờ đến.
 Việc tuy nhỏ, không làm chẳng bao giờ nên
Tuân Tử

29