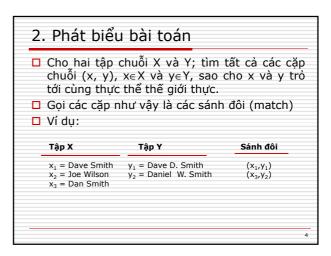
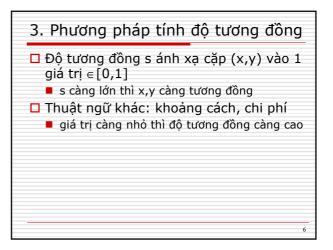
# Đối sánh chuỗi Nguyễn Hồng Phương Email: phuong.nguyenhong@hust.edu.vn Site: http://is.hut.edu.vn/~phuongnh Bộ môn Hệ thống thông tin Viện Công nghệ thông tin và Truyền thông Đại học Bách Khoa Hà Nội

# 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

# 1. Giới thiệu Là bài toán tìm các chuỗi trỏ 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,...



	hách thức
_	Tính chính xác
	□ Lỗi chính tả
	☐ Định dạng khác nhau
	☐ Tên khác nhau
	□ => thước đo độ tương đồng s(x,y)∈[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
	<ul> <li>=&gt; chỉ áp dụng s(x,y) với các bộ đôi triển vọng</li> </ul>



### 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
- Môt số phương pháp
  - Edit Distance
  - Needleman-Wunch
  - Affine Gap
  - Smith-Waterman
  - Jaro
  - Jaro-Winkler

## 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 tư
- ☐ 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
- $\Box$  d(x,y)=d(y,x)

# Phương pháp Edit Distance (tiếp)

☐ Mối quan hệ giữa hàm khoảng cách d(x,y) và hàm tương đồng s(x,y)

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

□ Ví du:

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

Phương pháp Edit Distance (tiếp)

- ☐ Giá trị của d(x,y) có thể được tính toán dựa trên quy hoạch động
- ☐ Cho
  - $\mathbf{x} = \mathbf{x}_1 \mathbf{x}_2 ... \mathbf{x}_n, \ \mathbf{y} = \mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_m$
  - x<sub>i</sub> và y<sub>i</sub> là các kí tự
  - d(i,j): khoảng cách soạn thảo giữa x<sub>1</sub>x<sub>2</sub>...x<sub>i</sub> (tiền tố thứ i của x) và y<sub>1</sub>y<sub>2</sub>...y<sub>j</sub> (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)

- $\square$  Biến đổi chuỗi  $x_1x_2...x_i$  thành chuỗi  $y_1y_2...y_j$ 
  - (a) Biến đổi x<sub>1</sub>x<sub>2</sub>...x<sub>i-1</sub> thành y<sub>1</sub>y<sub>2</sub>...y<sub>j-1</sub>, sau đó copy x<sub>i</sub> vào y<sub>j</sub> nếu x<sub>i</sub> = y<sub>j</sub>
  - (b) Biến đổi x₁x₂...x₁.1 thành y₁y₂...y₁.1, sau đó thay thế x₁ bởi y₁ nếu x₁ ≠ y₁
  - (c) Xóa x<sub>i</sub>, sau đó biến đổi x<sub>1</sub>x<sub>2</sub>...x<sub>i-1</sub> thành y<sub>1</sub>y<sub>2</sub>...y<sub>j</sub>
  - (d) Biến đổi x<sub>1</sub>x<sub>2</sub>...x<sub>i</sub> thành y<sub>1</sub>y<sub>2</sub>...y<sub>j-1</sub>, sau đó chèn thêm y<sub>i</sub>
- ☐ Giá trị d(i,j) là tối thiểu của các chi phí biến đổi ở trên

Phương pháp Edit Distance (tiếp)

☐ Biểu thức quay lui:

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

☐ Hoặc viết gon:

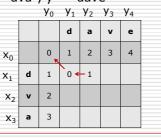
$$d(i, j) = min \begin{cases} d(i-1, j-1) + c(x_i, y_j) & \text{// copy or substitute} \\ d(i-1, j) + 1 & \text{// delete } x_i \\ d(i, j-1) + 1 & \text{// insert } y_j \end{cases}$$

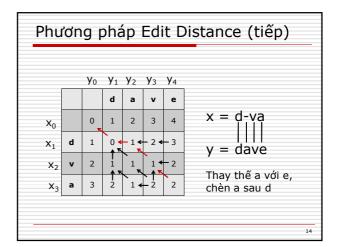
$$c(x_i, y_j) = \begin{cases} 0 & \text{if } x_i = y_j \\ 1 & \text{if } x_i \neq y_j \end{cases}$$

2

# Phương pháp Edit Distance (tiếp)

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





# Phương pháp Edit Distance (tiếp)

- Mũi tên đi chéo từ ô (3,4) tới ô (2,3): kí tự x<sub>3</sub> (kí tự a) được copy vào hoặc thay thế bởi kí tự y<sub>4</sub> (kí tự e).
- Mũi tên đi chéo từ ô (2,3) tới ô (1,2): x<sub>2</sub> (kí tự v) được copy vào hoặc thay thế bởi y<sub>3</sub> (kí tự v).
- ☐ Mũi tên đi ngàng 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

13

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

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 dung:
    - Xem xét các từ (phân cách nhau bởi kí tự cách)
    - ☐ Loại bỏ từ dừng
    - Ví dụ: xâu "david smith" => tập token {david, smith}

- 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 3grams là {##d, #da, dav, avi, ..., ith, th#, h##}
- Một số phương pháp
  - Overlap
  - Jaccard
  - TF/IDF

18

# Phương pháp Overlap

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

19

# Phương pháp Jaccard

- Dộ tương đồng Jaccard giữa 2 xâu x và y là J(x,y)=|Bx∩By|/|Bx∪By|
- □ Ví du:
  - x = dave; y = dav
  - B<sub>x</sub>={#d, da, av, ve, e#}
  - B<sub>v</sub>={#d, da, av, v#}
  - I(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 = Apple Corporation, CA
  - y = IBM Corporation, CA
  - z = 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 du:

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

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

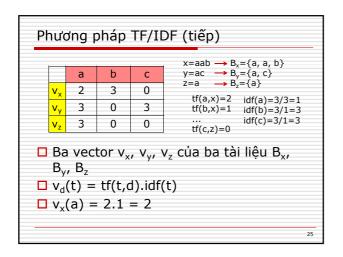
tf(a,x)=2 idf(a)=3/3=1 tf(b,x)=1 idf(b)=3/1=3... idf(c)=3/1=3

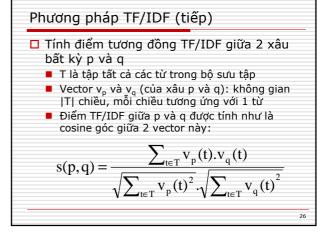
tf(c,z)=0

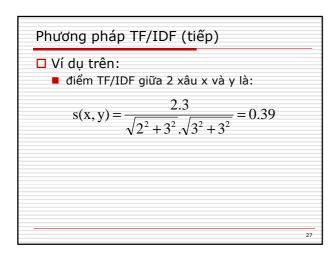
Phương pháp TF/IDF (tiếp)

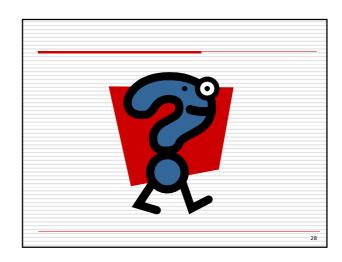
- ☐ Tiếp theo, biến đổi từng tài liệu d thành vector đặc trưng v<sub>d</sub>
- ☐ Hai tài liệu càng tương đồng nếu vector tương ứng của chúng gần nhau
- □ Vector của d có đặc trưng v<sub>d</sub>(t) với mỗi từ t. Giá trị của v<sub>d</sub>(t) là hàm của TF và IDF
- □ v<sub>d</sub> có nhiều đặc trưng bằng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

24









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