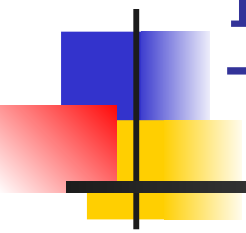
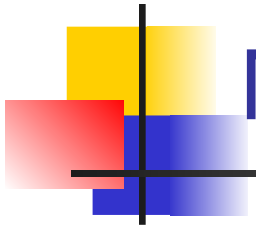


IT4853
Tìm kiếm và trình diễn thông tin

Bài 16. Phát hiện trùng lặp gần

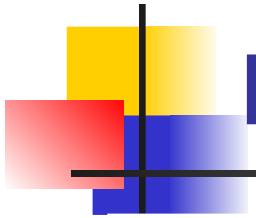
IIR.C19. Web search basics

Bộ môn Hệ thống thông tin
Viện CNTT & TT



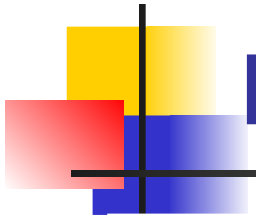
Nội dung chính

- Phát hiện trùng lặp gần
- Tính độ tương đồng bằng hệ số Jaccard
- Ước lượng hệ số Jaccard sử dụng phép trộn



Phân loại trùng lặp

- Trùng lặp tuyệt đối
  - Dễ dàng loại bỏ, v.d., bằng tổng đại diện.
- Trùng lặp gần
  - Khó phát hiện.

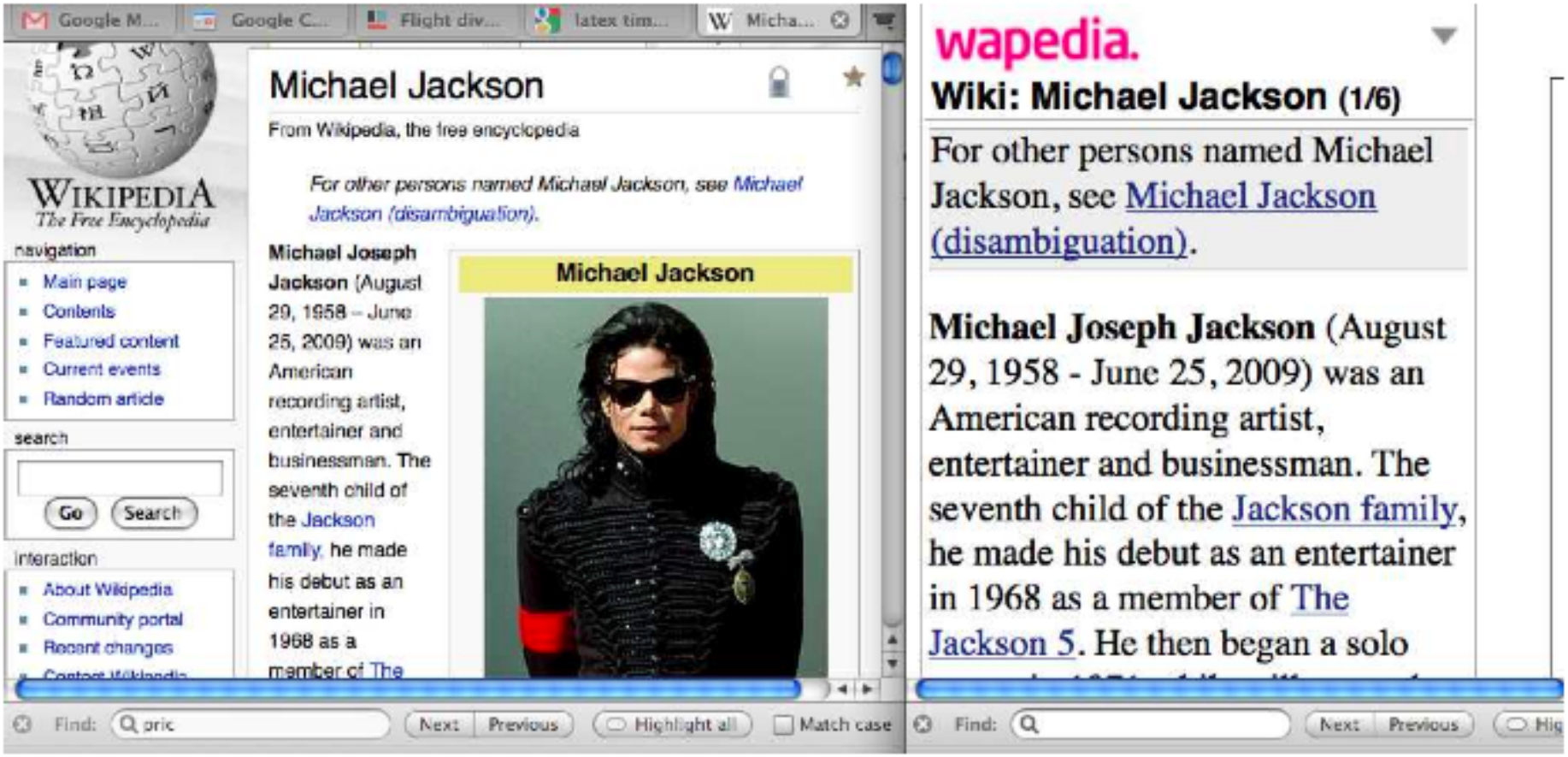


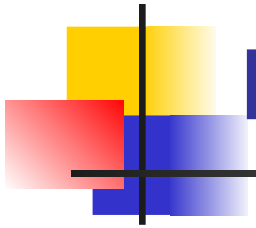
Phát hiện trùng lặp

- Người dùng không mong muốn nhận những kết quả trùng lặp
  - Một tài liệu dù phù hợp có thể bị coi là không phù hợp nếu lặp lại trong danh sách kết quả.

Cần loại bỏ những tài liệu trùng lặp. Chỉ giữ lại một tài liệu nếu có nhiều tài liệu trùng lặp!

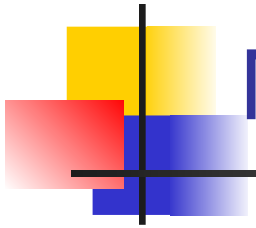
Trùng lặp gần





Phát hiện trùng lặp gần

- Tính độ tương đồng dựa trên "ký tự"
  - Rất khó tính độ tương đồng ngữ nghĩa;
  - Những văn bản cùng nội dung nhưng được diễn đạt khác nhau không phải trùng lặp;
- Sử dụng ngưỡng θ để kết luận "trùng lặp".
  - Coi hai tài liệu là trùng lặp gần nếu độ tương đồng > θ



Nội dung chính

- Phát hiện trùng lặp gần
- Tính độ tương đồng bằng hệ số Jaccard
- Ước lượng hệ số Jaccard sử dụng phép trộn

Biểu diễn văn bản: Mô hình tập shingles

- Shingle là một n-gram trên từ (bộ n-từ).
- Ví dụ, với *n* = 3, "a rose is a rose" có mô hình tập shingles như sau:
  - { a-rose-is, rose-is-a, is-a-rose }

Xác định độ tương đồng của hai tài liệu bằng hệ số Jaccard

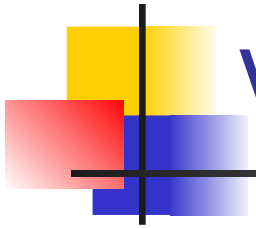
Hệ số Jaccard

- Cho hai tập đặc trưng A và B

$$\text{Với } A \neq \emptyset \text{ hoặc } B \neq \emptyset,$$

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Jaccard(A, A) = 1
- Jaccard(A, B) = 0 nếu $A \cap B = \emptyset$
- Miền giá trị là khoảng [0, 1]



Ví dụ tính hệ số Jaccard

- Cho ba tài liệu:

  d_1: "Jack London traveled to Oakland"

  d_2: "Jack London traveled to the city of Oakland"

  d_3: "Jack traveled from Oakland to London"

- Hãy tính hệ số Jaccard $J(d_1, d_2)$ và $J(d_1, d_3)$ sử dụng các bộ 2-từ?

Ví dụ tính hệ số Jaccard (2)

- Cho ba tài liệu:

  d_1: "Jack London traveled to Oakland"

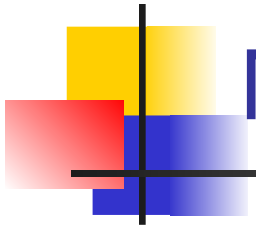
  d_2: "Jack London traveled to the city of Oakland"

  d_3: "Jack traveled from Oakland to London"

- Hãy tính hệ số Jaccard $J(d_1, d_2)$ và $J(d_1, d_3)$ sử dụng các bộ 2-từ?

$J(d_1, d_2) = 3/8 = 0.375$; $J(d_1, d_3) = 0$

Hệ số Jaccard trên tập shingle rất nhạy với trật tự từ



Nội dung chính

- Phát hiện trùng lặp gần
- Tính độ tương đồng bằng hệ số Jaccard
- Ước lượng hệ số Jaccard sử dụng phép trộn

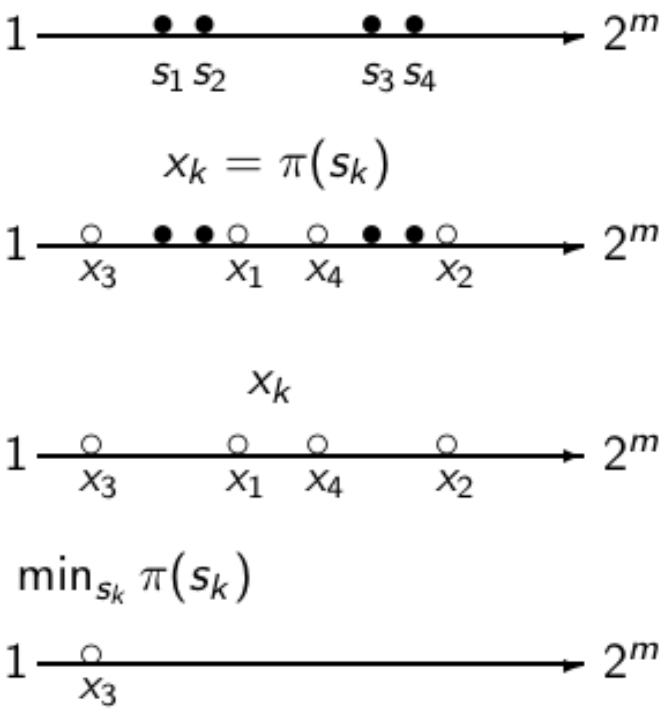
Biểu diễn khung của văn bản

- Biểu diễn khung (sketch) là biểu diễn giản lược của mô hình tập shingles.
  - Tính tổng đại diện cho các shingles
    - Ký hiệu s_k là tổng đại diện của shingle$_k$, $1 <= s_k <= 2^m$.
  - Sử dụng các thao tác trộn $\pi_1 \ldots \pi_K$ trên tổng đại diện.
- Sketch của d được định nghĩa là:

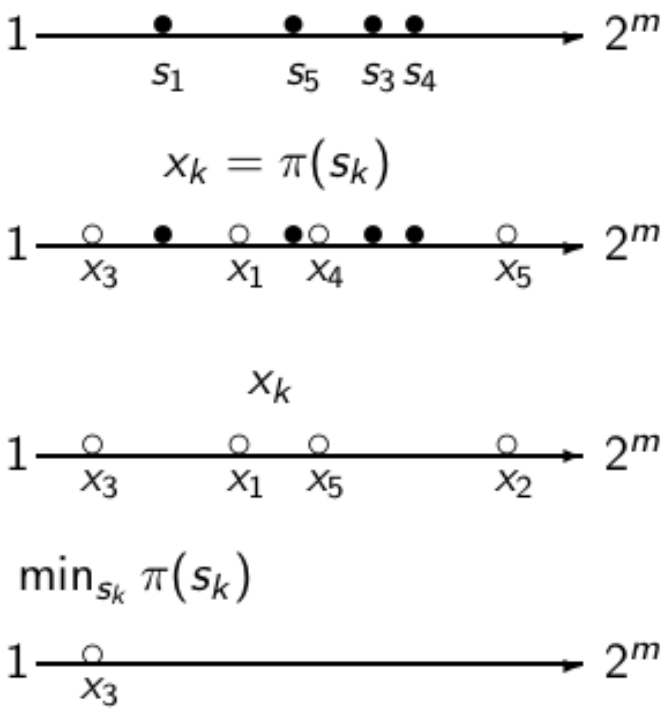
$$< \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \ldots, \min_{s \in d} \pi_K(s) >$$

Phép trộn thành công

tài liệu 1: $\{s_k\}$



tài liệu 2: $\{s_k\}$



Phép trộn π với $\min_{s\in d1} \pi(s) = \min_{s\in d2} \pi(s)$ được gọi là phép trộn thành công.

Trong trường hợp này phép trộn π khẳng định: $d_1 \approx d_2$

Ước lượng hệ số Jaccard

- Đặt $U = d_1 \cup d_2$; $I = d_1 \cap d_2$
- Có $|U|!$ phép trộn trên U
- Với $s' \in I$, có bao nhiêu phép trộn π để

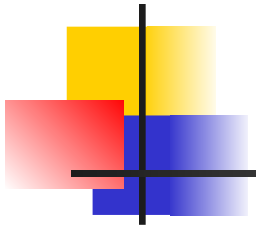
  $\text{argmin}_{s \in d1}\, \pi(s) = s' = \text{argmin}_{s \in d2}\, \pi(s)$?
  - Trả lời: $(|U| - 1)!$
- $\Rightarrow$ có $|I|(|U| - 1)!$ phép trộn thỏa mãn:

  $\text{argmin}_{s \in d1}\, \pi(s) = \text{argmin}_{s \in d2}\, \pi(s)$
- Như vậy, tỉ lệ phép trộn thỏa mãn

  $\min_{s \in d1}\, \pi(s) = \min_{s \in d2}\, \pi(s)$ là: $\dfrac{|I|(|U| - 1)!}{|U|!} = \dfrac{|I|}{|U|} = J(d_1, d_2)$

Ước lượng hệ số Jaccard (2)

- Hệ số Jaccard bằng tỉ lệ phép trộn thành công.
- Ước lượng xác suất trộn thành công
  - 1. Sử dụng K phép trộn, v.d., $K = 200$
  - 2. Đếm số lượng k phép trộn thành công
  - 3. k/K là giá trị gần đúng của $J(d_1, d_2)$.



Cài đặt

- Sử dụng hàm băm với vai trò là phép trộn:

  $h_i : \{1..2^m\} \rightarrow \{1..2^m\}$

Ví dụ

	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1

$h(x) = x \bmod 5$

$g(x) = (2x + 1) \bmod 5$

$\min(h(d_1)) = 1 \neq 0 = \min(h(d_2))$ $\min(g(d_1)) =$

$2 \neq 0 = \min(g(d_2))$

$\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$

	d_1 slot		d_2 slot	
h		∞		∞
g		∞		∞
$h(1) = 1$	1	1	–	∞
$g(1) = 3$	3	3	–	∞
$h(2) = 2$	–	1	2	2
$g(2) = 0$	–	3	0	0
$h(3) = 3$	3	1	3	2
$g(3) = 2$	2	2	2	0
$h(4) = 4$	4	1	–	2
$g(4) = 4$	4	2	–	0
$h(5) = 0$	–	1	0	0
$g(5) = 1$	–	2	1	0

Kết quả trộn

Cực tiểu

Tổng kết: Loại bỏ trùng lặp gần

- Đầu vào:
  - N tài liệu
  - Kích thước shingle, ví dụ, $n = 5$
  - K phép trộn ngẫu nhiên (K hàm băm)
- Tính $\frac{N(N-1)}{2}$ độ phù hợp theo cặp
- Coi hai tài liệu có độ tương đồng > θ là trùng lặp
  - Bỏ qua các trùng lặp khi đánh chỉ mục.

Các giải pháp phát hiện trùng lặp gần khác

- Vấn đề: Cần phải tính N*(N-1)/2 giá trị tương đồng.
  - Khối lượng tính toán lớn.
- Các giải pháp cải thiện hiệu năng:
  - Giải pháp 1: hàm băm cục bộ (LSH)

Andoni, Alexandr, Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. 2006. Locality-sensitive hashing using stable distributions. In Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press. 314, 519, 522, 524,527

  - Giải pháp khác: sắp xếp (Henzinger 2006)

Henzinger, Monika R., Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In Proc. WWW, pp. 295–308. North-Holland. DOI: dx.doi.org/10.1016/S1389-1286(00)00055-4. 442, 524, 527, 528

Bài tập 21.1

Cho mô hình tập shingle của văn bản

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

Hãy ước lượng hệ số Jaccard:
$J(d_1, d_2)$, $J(d_1, d_3)$, $J(d_2, d_3)$

$h(x) = (5x + 5) \bmod 4$
$g(x) = (3x + 1) \bmod 4$

Đáp án

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$h(x) = (5x + 5) \bmod 4$

$g(x) = (3x + 1) \bmod 4$

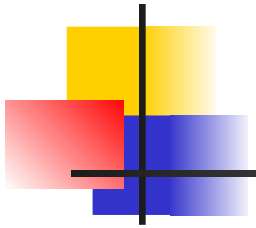
$$\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$$

$$\hat{J}(d_1, d_3) = \frac{0+0}{2} = 0$$

$$\hat{J}(d_2, d_3) = \frac{0+1}{2} = 1/2$$

	d_1 slot	d_2 slot	d_3 slot
	∞	∞	∞
	∞	∞	∞
$h(1) = 2$	– ∞	2 2	2 2
$g(1) = 0$	– ∞	0 0	0 0
$h(2) = 3$	3 3	– 2	3 2
$g(2) = 3$	3 3	– 0	3 0
$h(3) = 0$	– 3	0 0	– 2
$g(3) = 2$	– 3	2 0	– 0
$h(4) = 1$	1 1	– 0	– 2
$g(4) = 1$	1 1	– 0	– 0

Các biểu diễn khung



Bài tập 21.2

Xét một cách ước lượng khác: Lấy ngẫu nhiên một tập con các giá trị từ tập các giá trị của cả S_1 và S_2. Sau đó tính hệ số Jaccard giữa các tập con như vậy.

a) Hãy chứng minh đây là một cách ước lượng không lệch (unbiased estimator);

b) Hãy giải thích vì sao khó áp dụng cách ước lượng này trong thực tế?