Phản hồi thông tin

Lê Thanh Hương
Bộ môn Hệ thống thông tin
Viện CNTT&TT

Phản hồi thông tin

- Phản hồi thông tin (Information Retrieval - IR) là việc tìm các tài liệu phi cấu trúc (thường là văn bản) thỏa điều kiện tìm kiếm từ một kho dữ liệu lớn (thường được lưu trong máy tính)

Các hệ thống dựa trên từ khóa

- tập các từ khóa có khả năng xuất hiện trong tài liệu (vd., JFK, assasination)
- Các phép toán AND OR:

AND(Kennedy, conspiracy, OR(assasination, murder))
or
AND(OR(Kennedy,JFK), OR(conspiracy, plot), OR(assasination,assasinated,assasinate,murder, murdered,kill,killed)

Các vấn đề

- Đa nghĩa: 1 từ - n nghĩa
- Đồng nghĩa: n từ - 1 nghĩa
- Kích thước: các hệ thống IR phải có khả năng xử lý tập ngữ liệu cỡ ~Gb
- Độ phủ: Các hệ thống IR phải có khả năng xử lý câu truy vấn thuộc bất kỳ lĩnh vực nào

Lấy từ gốc

- Gắn các thuật ngữ trong câu truy vấn với các biến thể của từ (cùng gốc từ) trong các tài liệu
- VD: assassination → assassinat
  - Assassination Assassinations
  - Assassinate Assassinated
  - Assassinating
- Vấn đề:
  - Lỗi: organization - organ past - paste
  - Bỏ qua: analysis - analyzes matrices - matrix

Từ dừng

- Là các từ thường xuất hiện ở hầu hết các tài liệu. Các từ này không chứa nhiều thông tin
- Không đưa vào file nghịch đảo → giảm kích thước của file này
- Các từ dừng: a, an, the, he, she, of, to, by, should, can,…

Nhược điểm của việc bỏ từ dừng

- Có thể bỏ tên người như "The"
- Các từ dừng có thể là thành phần quan trọng của đoạn. Ví dụ, 1 câu nói của Shakepeare: "to be or not to be"
- Một số từ dừng (vd., giới từ) cung cấp các thông tin quan trọng về mối quan hệ
- Bộ nhớ ngày nay đã rẻ hơn → tiết kiệm bộ nhớ không còn là vấn đề quan trọng như trước nữa

Từ chức năng và từ nội dung

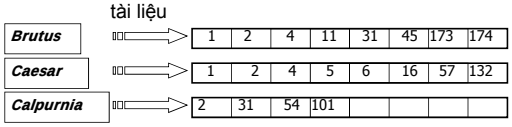
- Muốn loại bỏ các từ chức năng hoặc giảm ảnh hưởng của nó
- Xác định từ nội dung:
  - Nó có xuất hiện thường xuyên không?
  - Nó có xuất hiện trong số ít các tài liệu không?
  - Tần suất của nó có thay đổi trong các tài liệu không?

File nghịch đảo (Inverted Files)

- Để biểu diễn tài liệu trong kho ngữ liệu
- Là 1 bảng từ với 1 danh sách các tài liệu chứa 1 từ
  - Assassination: (doc1, doc4, doc35,…)
  - Murder: (doc3, doc7, doc36,…)
  - Kennedy: (doc24, doc27, doc29,…)
  - Conspiracy: (doc3, doc55, doc90,…)
- Thông tin bổ sung:
  - vị trí của từ trong tài liệu
  - thông tin xấp xỉ: để so khớp hoặc so gần đúng các đoạn

Chỉ số nghịch đảo

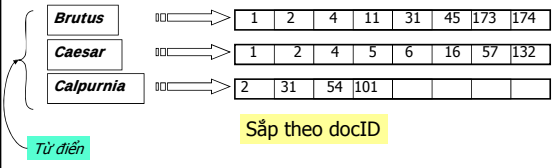
- Với mỗi thuật ngữ t, lưu danh sách các tài liệu chứa t.
  - Định nghĩa mỗi tài liệu bởi **docID**, là số thứ tự của tài liệu



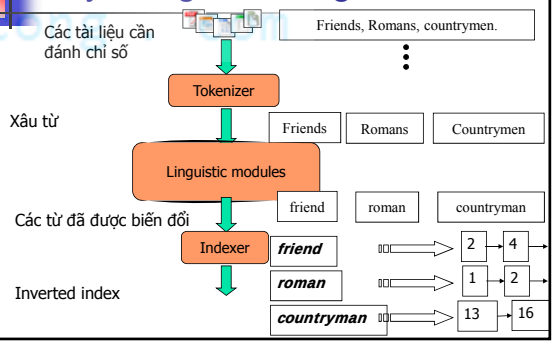
Vấn đề gì xảy ra nếu từ ***Caesar*** được thêm vào tài liệu 14?

Chỉ số nghịch đảo

- Ta cần các danh sách với độ dài thay đổi
  - Có thể sử dụng linked list hoặc mảng có độ dài thay đổi



Xây dựng chỉ số nghịch đảo



Bước đánh chỉ số: Chuỗi từ

- Chuỗi các cặp
(từ đã biến đổi, Document ID)

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Bước đánh chỉ số: Sắp xếp

- Sắp theo từ, rồi theo docID

Bước đánh chỉ số cốt lõi

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

Bước đánh chỉ số: Từ điển và danh sách

- Nhiều chỉ mục từ trong 1 tài liệu được trộn lẫn
- Đưa vào trong từ điển và danh sách
- Thêm số lần xuất hiện của tài liệu

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
I	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Lưu trữ

Danh sách docIDs

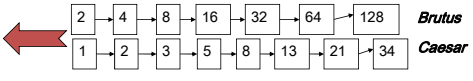
Thuật ngữ và số lần xuất hiện

Con trỏ

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
I	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Xử lý truy vấn: AND

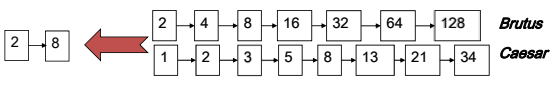
- Xét câu truy vấn:
Brutus *AND* ***Caesar***
  - Định vị ***Brutus*** trong từ điển;
    - Lấy danh sách của nó.
  - Định vị *Caesar* trong từ điển;
    - Lấy danh sách của nó.
  - Trộn 2 danh sách



17

Phép trộn

- Duyệt qua 2 danh sách, thời gian tỉ lệ với số nút



Nếu 2 danh sách có độ dài là x và y, phép trộn có độ phức tạp O(*x+y*) .
Vấn đề cốt yếu: các danh sách sắp theo docID

18

Trộn 2 danh sách

```
INTERSECT(p1, p2)
1  answer ← ⟨ ⟩
2  while p1 ≠ NIL and p2 ≠ NIL
3  do if docID(p1) = docID(p2)
4        then ADD(answer, docID(p1))
5             p1 ← next(p1)
6             p2 ← next(p2)
7        else if docID(p1) < docID(p2)
8               then p1 ← next(p1)
9               else p2 ← next(p2)
10 return answer
```

19

Câu truy vấn logic: so khớp

- Mô hình phản hồi Boolean có thể trả lời câu truy vấn ở dạng biểu thức Boolean
  - Câu truy vấn sử dụng *AND, OR* và *NOT* để kết nối các thuật ngữ
    - Coi mỗi tài liệu là 1 tập các từ
    - Chính xác: tài liệu thỏa điều kiện hoặc không
  - Đây là mô hình IR đơn giản nhất

20

Câu truy vấn logic: phép trộn tổng quát hơn

- Bài tập: Thực hiện phép trộn cho các câu truy vấn:

  Brutus *AND NOT* ***Caesar***

  Brutus *OR NOT* ***Caesar***

Thời gian thực hiện còn là O(*x*+*y*)?

21

Phép trộn

Thực hiện phép trộn cho các câu truy vấn:

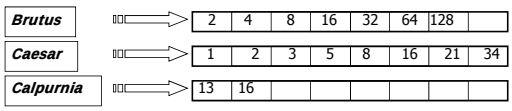
(Brutus *OR* ***Caesar)*** *AND NOT* ***(Antony*** *OR* ***Cleopatra)***

- Có thể luôn thực hiện trong thời gian tuyến tính?
- Có thể làm tốt hơn không?

22

Tối ưu hóa truy vấn

- Đâu là trật tự tốt nhất để xử lý truy vấn?
- Xét 1 câu truy vấn là phép AND của n thuật ngữ
- Với mỗi thuật ngữ, lấy danh sách của nó , sau đó làm phép AND.

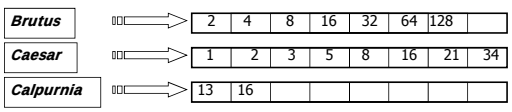


Query: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

23

Tối ưu hóa truy vấn – Ví dụ

- Xử lý theo trật tự tăng của tần suất:
  - *khởi đầu với tập nhỏ, sau đó tiếp tục loại bỏ*



Thực hiện câu truy vấn (***Calpurnia*** *AND* ***Brutus)*** *AND* ***Caesar***.

24

Tối ưu hóa truy vấn

- vd., *(**madding** OR **crowd**) AND (**ignoble** OR **strife**)*
- Lấy tần suất xuất hiện cho mọi thuật ngữ
- Đánh giá kích thước của mỗi câu lệnh OR bằng cách tính tổng các tần suất của nó
- Xử lý theo trật tự tăng của kích thước các danh sách trong phép *OR*

25

Bài tập

- Đưa ra trình tự xử lý truy vấn cho

(tangerine** OR **trees)** AND **(marmalade** OR **skies)** AND **(kaleidoscope** OR **eyes)

Term	Freq
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

26

Bài tập

- Cho câu truy vấn ***friends** AND **romans** AND (NOT **countrymen**)*, ta sử dụng tần suất của ***countrymen*** như thế nào?
- Mở rộng phép trộn cho câu truy vấn ngẫu nhiên. Có thể đảm bảo thực hiện trong thời gian tuyến tính với tổng kích thước các danh sách không

27

Các kỹ thuật nâng cao

- Cụm từ: ***Stanford University***
- Xấp xỉ: Tìm ***Gates** NEAR **Microsoft***.
  - Cần đánh chỉ số để lấy thông tin về vị trí trong các tài liệu
- Vị trí trong tài liệu: Tìm các tài liệu có (*author* = ***Ullman***) *AND* (text contains ***automata***).
- Từ khóa tìm kiếm xuất hiện trong 1 tài liệu nhiều hơn thì tốt hơn
  - Cần thông tin về tần suất của thuật ngữ trong các tài liệu
- Cần độ đo xấp xỉ câu truy vấn với tài liệu
- Cần quyết định trả về 1 tài liệu thỏa câu truy vấn hay một nhóm tài liệu phủ các khía cạnh khác nhau của câu truy vấn

28

Từ và thuật ngữ

- IR quan tâm đến thuật ngữ
- VD: câu truy vấn
  - `What kind of monkeys live in Costa Rica?`

29

Từ và thuật ngữ

- `What kind of monkeys live in Costa Rica?`
  - từ?
  - từ nội dung?
  - gốc từ?
  - các nhóm từ?
  - các đoạn?

30

Cụm từ (các từ thường đi liền nhau)

- kick the bucket
- directed graph
- iambic pentameter
- Osama bin Laden
- United Nations
- real estate
- quality control
- international best practice
- … có ý nghĩa riêng, cách dịch riêng.

31

Tìm cụm từ

- Sử dụng bigrams?
- Không tốt:
  - 80871 of the
  - 58841 in the
  - 26430 to the
  - …
  - 15494 to be
  - …
  - 12622 from the
  - 11428 New York
  - 10007 he said
- Giải quyết: bỏ các từ dừng

32

Tìm cụm từ

- Sử dụng bigrams?
- Tốt hơn: lọc theo thẻ : A N, N N, N P N …
  - 11487 New York
  - 7261 United States
  - 5412 Los Angeles
  - 3301 last year
  - …
  - 1074 chief executive
  - 1073 real estate
  - …

33

Tìm cụm từ

- Vẫn muốn bỏ "new companies"
- Các từ này thường xuất hiện nhưng chỉ vì cả 2 từ đều thường xuất hiện
- Quan sát xác suất của từng từ và xác suất của cụm từ
  - p(new) p(companies)
  - p(new companies)
  - thông tin tương hỗ = p(new) p(companies | new)

34

data from Manning & Schütze textbook (14 million words of NY Times)

Thông tin tương hỗ

	new ___	¬new ___	TOTAL
___ companies	8	4,667 ("old companies")	4,675
___ ¬companies	15,820	14,287,181 ("old machines")	14,303,001
TOTAL	15,828	14,291,848	14,307,676

N

- p(new companies) = p(new) p(companies) ?
- MI = $\log_2$ p(new companies) / p(new)p(companies)
  = $\log_2$ (8/N) /((15828/N)(4675/N)) = $\log_2$ 1.55 = 0.63
- MI > 0 nhưng nhỏ. Với các cụm từ thường xuất hiện, giá trị này lớn hơn

35

data from Manning & Schütze textbook (14 million words of NY Times)

Phép thử mức độ quan trọng

	new ___	¬new ___	TOTAL
___ companies	1	583 ("old companies")	584
___ ¬companies	1978	1,785,898 ("old machines")	1,787,876
TOTAL	1979	1,786,481	1,788,460

- Dữ liệu thưa. Giả sử chia tất cả các giá trị cho 8.
  - Giá trị MI có thay đổi không?
  - Không. Nhưng khả năng là cụm từ của nó ít hơn.
  - Điều gì xảy ra nếu 2 từ mới xuất hiện cạnh nhau?
- Cần thử mức độ quan trọng. Kích thước dữ liệu cũng là 1 yếu tố quan trọng

36

data from Manning & Schütze textbook (14 million words of NY Times)

Mức độ quan trọng nhị thức

	new ___	¬new ___	TOTAL
___ companies	8	4,667	4,675
___ ¬companies	15,820	14,287,181	14,303,001
TOTAL	15,828	14,291,848	14,307,676

- Giả sử có 2 đồng xu dùng để sinh văn bản.
- Tiếp theo new, ta dùng xu A để quyết định xem có từ companies tiếp theo không
- Tiếp theo ¬ new, ta dùng xu B để quyết định xem có từ companies tiếp theo không
- Ta thấy A được tung 15828 lần và 8 lần có mặt ngửa
- B được tung 14291848 lần và 4667 lần có mặt ngửa
- Câu hỏi: 2 đồng xu có trọng số khác nhau không? Nói cách khác, cùng 1 đồng xu hay 2 đồng xu

data from Manning & Schütze textbook (14 million words of NY Times)

Mức độ quan trọng nhị thức

	new ___	¬new ___	TOTAL
___ companies	8	4,667	4,675
___ ¬companies	15,820	14,287,181	14,303,001
TOTAL	15,828	14,291,848	14,307,676

- **Giả thiết Null:** cùng 1 đồng xu
  - giả sử p_{null}(co's | new) = p_{null}(co's | ¬new) = p_{null}(co's) = 4675/14307676
  - p_{null}(data) = p_{null}(8 out of 15828)*p_{null}(4667 out of 14291848) = .00042
- **Giả thiết đồng xuất hiện:** 2 đồng xu khác nhau
  - giả sử p_{coll}(co's | new) = 8/15828, p_{coll}(co's | ¬new) = 4667/14291848
  - p_{coll}(data) = p_{coll}(8 out of 15828)*p_{coll}(4667 out of 14291848) = .00081
- Do đó giả thiết đồng xuất hiện gấp đôi dữ liệu p(data).
  - Ta có thể sắp xếp bigrams theo giá trị log p_{coll}(data)/p_{null}(data)
  - nghĩa là, mức độ chắc chắn "companies" đi sau "new" như thế nào

data from Manning & Schütze textbook (14 million words of NY Times)

Mức độ quan trọng nhị thức

	new ___	¬new ___	TOTAL
___ companies	1	583	584
___ ¬companies	1978	1,785,898	1,787,876
TOTAL	1979	1,786,481	1,788,460

- **Giả thiết Null:** cùng 1 đồng xu
  - giả sử p_{null}(co's | new) = p_{null}(co's | ¬new) = p_{null}(co's) = 584/1788460
  - p_{null}(data) = p_{null}(1 out of 1979)*p_{null}(583 out of 1786481) = .0056
- **Giả thiết đồng xuất hiện:** 2 đồng xu khác nhau
  - giả sử p_{coll}(co's | new) = 1/1979, p_{coll}(co's | ¬new) = 583/1786481
  - p_{coll}(data) = p_{coll}(1 out of 1979)*p_{coll}(583 out of 1786418) = .0061
- Giả thiết đồng xuất hiện vẫn tăng p(data), nhưng khá nhỏ.
  - Nếu không có nhiều dữ liệu, mô hình 2 đồng xu không thuyết phục.
  - Thông tin tương hỗ vẫn có giá trị, nhưng dựa trên ít dữ liệu hơn. Do vậy có thể tin rằng giả thiết Null chỉ là sự trùng hợp ngẫu nhiên.

Phân tích ngữ nghĩa tiềm ẩn

- Mỗi tài liệu được coi là 1 vector có độ dài k

aardvark abacus abandoned abbot abduct above . . . zygote zymurgy

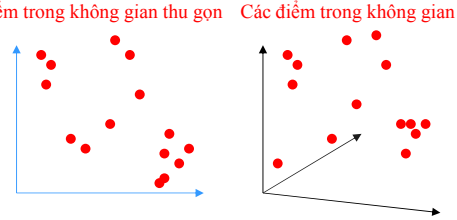
(0, 3, 3, 1, 0, 7, . . . 1, 0)

1 tài liệu

Phân tích ngữ nghĩa tiềm ẩn

- Mỗi tài liệu được biểu diễn thành 1 điểm trong không gian vector

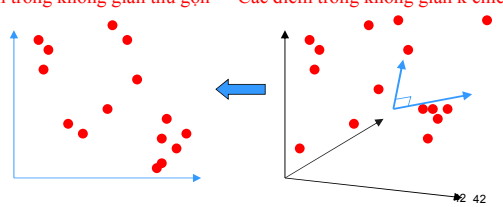
Các điểm trong không gian thu gọn Các điểm trong không gian k chiều



Phân tích ngữ nghĩa tiềm ẩn

- Giảm điểm: các điểm thực được chuyển về không gian ít chiều hơn
- ∃ một lựa chọn tốt nhất cho các chiều - có thể biểu diễn một cách tốt nhất các đặc tính của dữ liệu
  - Tìm được nhờ sử dụng đại số tuyến tính "Singular Value Decomposition" (SVD)

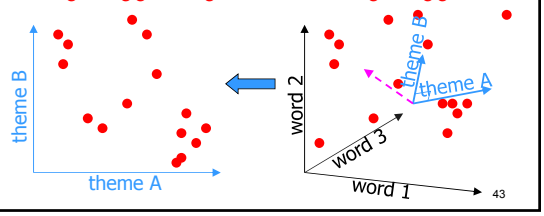
Các điểm trong không gian thu gọn Các điểm trong không gian k chiều



Phân tích ngữ nghĩa tiềm ẩn

- Các điểm SVD cho phép phục hồi các điểm thực (có thể phục hồi không gian 3 chiều với méo ít nhất)
  - Bỏ qua các sai khác trên các cạnh mà nó không chọn
  - Hy vọng các sai khác đó chỉ là nhiễu và chúng ta muốn bỏ qua nó

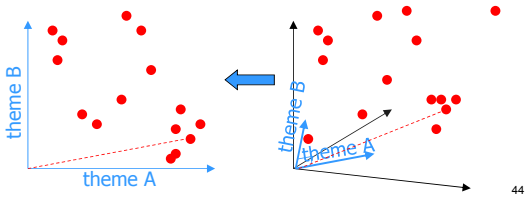
Các điểm trong không gian thu gọn Các điểm trong không gian k chiều



Phân tích ngữ nghĩa tiềm ẩn

- SVD tìm một vài vector chủ đề
- Mỗi tài liệu được xấp xỉ một sự kết hợp tuyến tính các chủ đề
- Liên kết trong không gian thu gọn = hệ số tuyến tính
  - Có bao nhiêu chủ đề A trong tài liệu? Có bao nhiêu chủ đề B trong tài liệu?
  - Có bao nhiêu chủ đề là 1 tập các từ thường xuất hiện cùng nhau

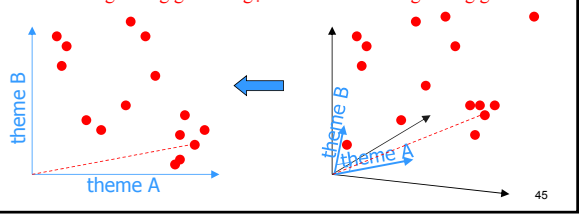
Các điểm trong không gian thu gọn Các điểm trong không gian k chiều



Phân tích ngữ nghĩa tiềm ẩn

- Các tọa độ mới có thể hữu ích trong IR
- Để so sánh 2 tài liệu, hoặc 1 câu hỏi và 1 tài liệu:
  - Chiếu cả 2 vào không gian thu gọn: chúng có cùng chủ đề không?
  - Thậm chí cả khi chúng không có từ nào chung

Các điểm trong không gian thu gọn Các điểm trong không gian k chiều



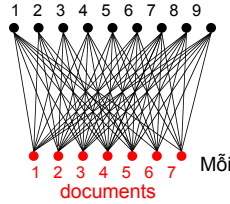
Phân tích ngữ nghĩa tiềm ẩn

- Các chủ đề trong IR có thể dùng trong phân giải nhập nhằng
- Mỗi từ là 1 tài liệu: (0,0,0,1,0,0,…)
- Biểu diễn từ như 1 kết hợp tuyến tính các chủ đề
- Mỗi chủ đề tương ứng với 1 nghĩa?
  - Vd., "Jordan" có các chủ đề Mideast và Sports
  - Nghĩa của từ trong tài liệu: chủ đề nào mạnh nhất trong tài liệu?
- Nhóm và tách các nghĩa
  - Một từ có nhiều nghĩa; nhiều từ có cùng nghĩa

Phân tích ngữ nghĩa tiềm ẩn

- Cách nhìn khác (tương tự mạng nơron):

terms



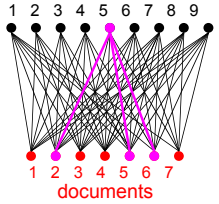
ma trận trọng số
(mỗi thuật ngữ trong tài liệu
có tác dụng như thế nào)

Mỗi cạnh có 1 trọng số cho bởi ma trận

Phân tích ngữ nghĩa tiềm ẩn

- Thuật ngữ 5 đóng vai trò quan trọng trong tài liệu nào

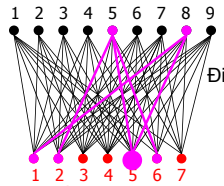
terms



trong các tài liệu 2,5,6

Phân tích ngữ nghĩa tiềm ẩn

- Thuật ngữ 5 và 8 đóng vai trò quan trọng trong tài liệu nào

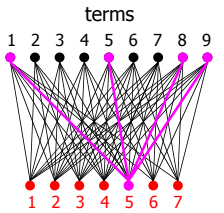


Điều này trả lời cho câu truy vấn chứa thuật ngữ 5 và 8

đó chỉ là phép nhân ma trận:
vector thuật ngữ(query) x trọng số của ma trận = vector tài liệu .

Phân tích ngữ nghĩa tiềm ẩn

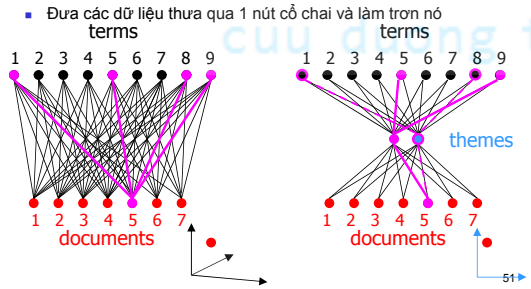
- Ngược lại, các thuật ngữ nào mạnh trong tài liệu 5?



cho các tọa độ của tài liệu 5

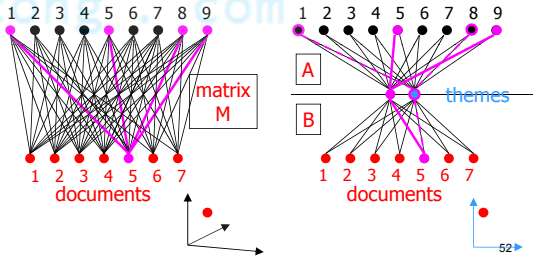
Phân tích ngữ nghĩa tiềm ẩn

- SVD xấp xỉ bằng mạng nơron 3 tầng
- Đưa các dữ liệu thưa qua 1 nút cổ chai và làm trơn nó



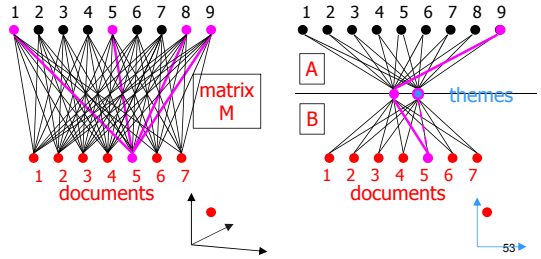
Phân tích ngữ nghĩa tiềm ẩn

- Nghĩa là, làm trơn dữ liệu thưa bằng ma trận xấp xỉ: **M ≈ A B**
  - A được mã hóa qua các chủ đề, B – mỗi tài liệu sẽ có tập thuật ngữ mới



Phân tích ngữ nghĩa tiềm ẩn

Coi A và B là các thuật ngữ và các tài liệu được chuyển về không gian chủ đề ít chiều, tại đó có thể xác định độ tương tự giữa chúng



Phân tích ngữ nghĩa tiềm ẩn

- Phân nhóm tài liệu (có thể giải quyết được dữ liệu thưa)
- Phân nhóm từ
- So sánh 1 từ với 1 tài liệu
- Xác định các chủ đề của 1 từ với các nghĩa của nó
  - Phân giải nhập nhằng bằng cách nhìn vào nghĩa của tài liệu
- Xác định các chủ đề con của tài liệu với chủ đề của nó
  - phân loại chủ đề

IR vs. CSDL: cấu trúc và phi cấu trúc

- Dữ liệu có cấu trúc: thông tin lưu trong bảng

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Cho phép tìm kiếm trong khoảng và tìm kiếm so khớp, ví dụ *Salary < 60000 AND Manager = Smith*.

55

Dữ liệu phi cấu trúc

- Thường đề cập đến dữ liệu văn bản dạng tự do
- Cho phép
  - Các truy vấn sử dụng từ khóa kết hợp các phép toán
  - các truy vấn ngữ nghĩa tinh vi, như
    - tìm tất cả các trang web có liên quan đến *drug abuse*

56

Dữ liệu bán cấu trúc

- Trên thực tế hầu hết dữ liệu đều không ở dạng phi cấu trúc
- Hỗ trợ các tìm kiếm bán cấu trúc như
  - *Title* contains data AND *Bullets* contain search

57

Dữ liệu bán cấu trúc

- *Title* is about Object Oriented Programming AND *Author* something like stro*rup
- Vấn đề:
  - làm cách nào xử lý “about”?
  - xếp hạng kết quả?
- Đây là trọng tâm của tìm kiếm XML

58

Các hệ thống IR phức tạp hơn

- IR đa ngôn ngữ
- Hỏi đáp
- Tóm tắt văn bản
- Khai phá văn bản
- …

59