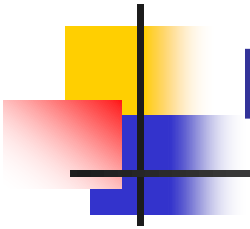
IT4853
Tìm kiếm và trình diễn thông tin

Bài 1. Phương pháp tìm kiếm Boolean

IIR.C1. Boolean retrieval

TS. Nguyễn Bá Ngọc, *Bộ môn Hệ thống thông tin, Viện CNTT & TT*
ngocnb@soict.hust.edu.vn

Hà Nội, 2016



Nội dung chính

- 1. Khái niệm tìm kiếm thông tin
- 2. Khái niệm mô hình
- 3. Mô hình Boolean và chỉ mục ngược

Tìm kiếm thông tin là gì?

Tìm kiếm thông tin là tìm kiếm các tài nguyên thông tin phi cấu trúc (thường là văn bản) từ một nguồn thông tin lớn (thường được lưu trên máy tính), đáp ứng được nhu cầu thông tin.

Thuật ngữ tiếng Anh là Information Retrieval (IR).

TKTT vs. CSDL: Dữ liệu có cấu trúc vs phi cấu trúc

- Dữ liệu có cấu trúc thường thể hiện được dưới dạng bảng

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

Cho phép truy xuất dạng so khớp và giới hạn miền giá trị, vd, *Salary < 60000 AND Manager = Smith*.

http://nlp.stanford.edu/IR-book/newslides.html

TKTT vs. CSDL: Dữ liệu có cấu trúc vs phi cấu trúc (2)

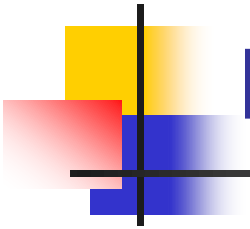
- Dữ liệu phi cấu trúc: Điển hình là những văn bản tự do.
- Cho phép:
  - Truy xuất bằng từ khóa
    - có thể kết hợp với ràng buộc logic
  - Sử dụng quan hệ ngữ nghĩa giữa các khái niệm, v.d,
    - tìm tất cả những trang web liên quan tới *công nghệ*

http://nlp.stanford.edu/IR-book/newslides.html

Dữ liệu bán cấu trúc

- Trong thực tế, hầu như rất hiếm dữ liệu văn bản tuyệt đối phi cấu trúc.
  - Nếu tính đến cả khả năng suy diễn cấu trúc yếu từ dữ liệu phi cấu trúc:
    - vd., có thể chia slide này thành hai phần là tiêu đề và nội dung
- Khái niệm bán cấu trúc nằm giữa khái niệm phi cấu trúc và khái niệm có cấu trúc theo mức độ chặt chẽ,
  - Có thể kết hợp phong cách tìm kiếm trên dữ liệu phi cấu trúc và phong cách tìm kiếm trên dữ liệu có cấu trúc cho dữ liệu bán cấu trúc,
    - vd., Tiêu đề có từ *thông tin* và Nội dung có từ *tìm kiếm*
    - *Tiêu đề* nói về lập trình C++ và *Tác giả* có tên như là stro*rup

http://nlp.stanford.edu/IR-book/newslides.html



Nội dung chính

- 1. Khái niệm tìm kiếm thông tin
- 2. Khái niệm mô hình
- 3. Mô hình Boolean và chỉ mục ngược

Mô hình tìm kiếm thông tin (1)

"Mô hình tìm kiếm là nền tảng lý thuyết để xây dựng công cụ tìm kiếm."

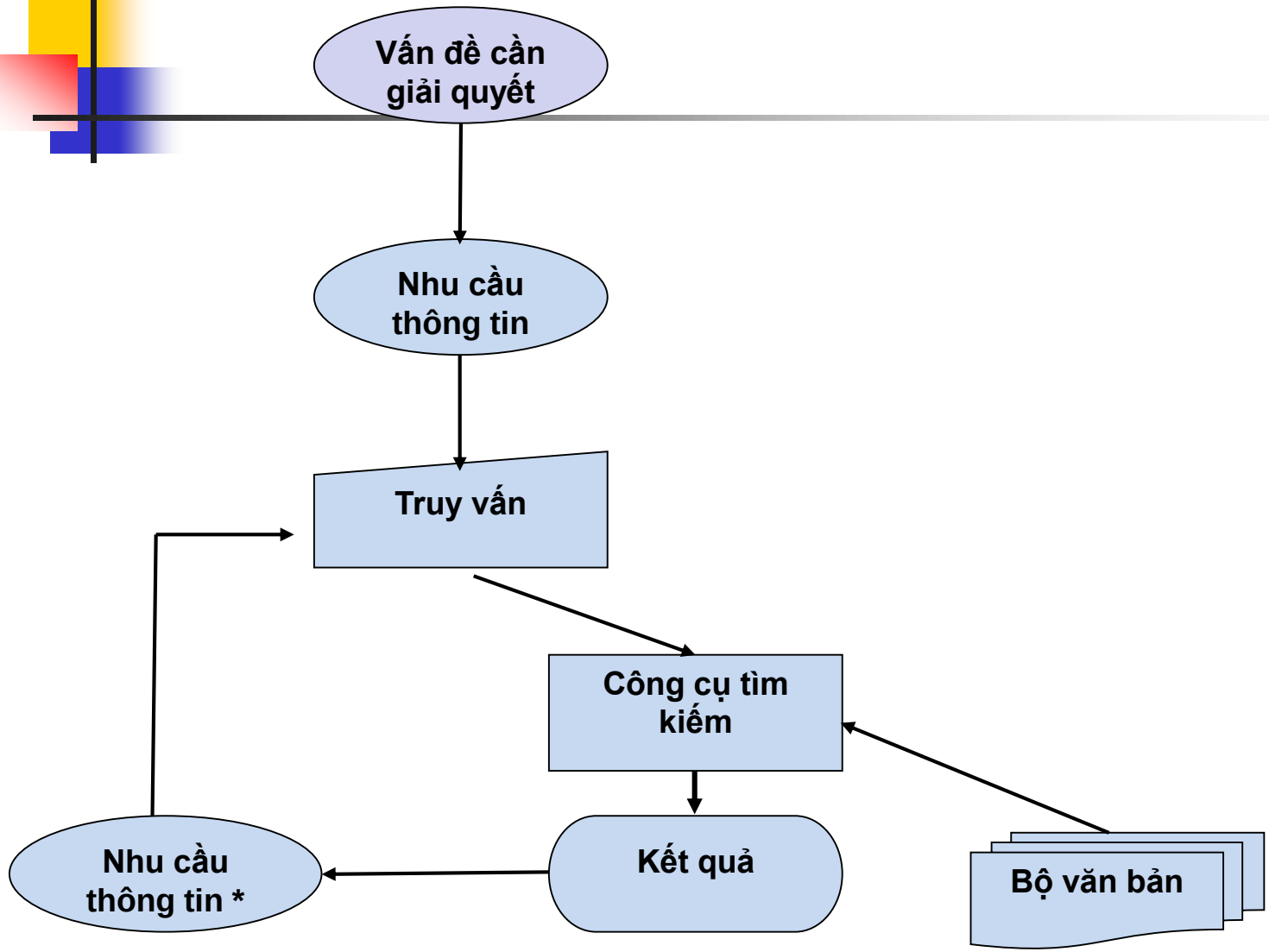
Nếu biết mô hình được sử dụng để xây dựng công cụ tìm kiếm thì có thể giải thích và dự đoán được hành vi của hệ thống tìm kiếm, v.d., vì sao văn bản A được trả về trước văn bản B? vì sao văn bản C không được trả về? làm thế nào để chiếm thứ hạng cao trong xếp hạng? V.v.

Mô hình tìm kiếm thông tin (2)

- Mô hình tìm kiếm quyết định các yếu tố sau:
  - **D:** Cách biểu diễn văn bản;
  - **Q:** Cách biểu diễn truy vấn;
  - **F:** Nền tảng lý thuyết (toán học) tương thích với D và Q, giữ vai trò cơ sở để thực hiện các suy diễn xếp hạng;
  - **R(d, q):** Hàm xếp hạng, là hàm định lượng mức độ phù hợp giữa văn bản và truy vấn.

Biểu diễn văn bản còn được gọi là mô hình văn bản;
Truy vấn về bản chất là biểu diễn của nhu cầu thông tin bằng ngôn ngữ của hệ thống tìm kiếm;
Một vài nền tảng lý thuyết quan trọng: tập hợp, đại số, xác suất,...

Mô hình tìm kiếm thông tin (3)



*Sau khi nhận kết quả tìm kiếm, người dùng chịu tác động của kết quả tìm kiếm và có thể dẫn đến thay đổi nhu cầu thông tin sau đó thiết lập lại truy vấn.

Nội dung chính

- 1. Khái niệm tìm kiếm thông tin
- 2. Khái niệm mô hình
- 3. Mô hình Boolean và chỉ mục ngược

Mô hình Boolean

- Ra đời từ khoảng 3 thập kỷ trước đây và là mô hình được sử dụng rộng rãi nhất trong thời gian đó.
- Hiện nay vẫn đang được sử dụng trong nhiều hệ thống,
  - Vd, thư viện số : http://www.westlaw.com
    - nhiều TB dữ liệu, > 700K người dùng

Mô hình Boolean (2)

D: Văn bản được biểu diễn dưới dạng tập từ;

Q: Biểu thức Boolean trên từ, ràng buộc sự xuất hiện của từ trong văn bản;

F: Lý thuyết tập hợp, đại số Boolean;

R: Một văn bản phù hợp nếu nó thỏa mãn biểu thức truy vấn. R(d, q) chỉ trả về hai giá trị 0: không phù hợp, 1: phù hợp.

Ví dụ phù hợp Boolean

Truy vấn: ((*văn bản* ∨ *thông tin*) ∧ *tìm kiếm* ∧ ¬*lý thuyết*)

Văn bản:

1. *"Tìm kiếm thông tin*
2. *"Lý thuyết thông tin"*
3. *"Tìm kiếm thông tin hiện đại: lý thuyết và thực hành"*
4. *"Phương pháp nén văn bản"*

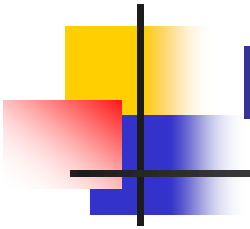
Ví dụ phù hợp Boolean

Truy vấn: ((*văn bản* ∨ *thông tin*) ∧ *tìm kiếm* ∧ ¬*lý thuyết*)

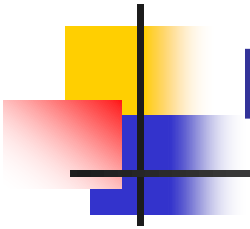
Văn bản:

1. *"Tìm kiếm thông tin*
2. *"Lý thuyết thông tin"*
3. *"Tìm kiếm thông tin hiện đại: lý thuyết và thực hành"*
4. *"Phương pháp nén văn bản"*

Thực hiện truy vấn Boolean trên dữ liệu nhỏ



- Kiểm tra tuần tự tất cả văn bản:
  - Đơn giản, nhưng…
  - .. Sẽ rất chậm khi chạy trên bộ dữ liệu lớn



Khái niệm chỉ mục

"Chỉ mục là cấu trúc dữ liệu chuyên biệt để tối ưu hóa tốc độ thực hiện truy vấn."

Thuật ngữ tiếng anh là Index

Ý tưởng sử dụng chỉ mục

Từ\Văn bản	d1	d2	d3	d4	d5	d6	d7
a	1	1	0	1	0	0	1
b	1	0	0	1	1	0	1
c	0	0	1	1	0	1	0
d	0	1	0	0	1	0	1
e	1	0	0	1	0	1	0

1: từ xuất hiện trong văn bản; 0: từ không xuất hiện.

Xử lý truy vấn trên ma trận đánh dấu

- Xử lý các truy vấn Boolean có thể quy về thực hiện phép toán logic theo bit:
  - Ví dụ, truy vấn a AND b AND NOT d được thực hiện như sau:
  - 1101001 AND
  - 1001101 AND
  - 1011010 =
  - 1001000

Từ\Văn bản	d1	d2	d3	d4	d5	d6	d7
a	1	1	0	1	0	0	1
b	1	0	0	1	1	0	1
c	0	0	1	1	0	1	0
d	0	1	0	0	1	0	1
e	1	0	0	1	0	1	0

Ưu điểm: Nhanh hơn kiểm tra tuần tự;

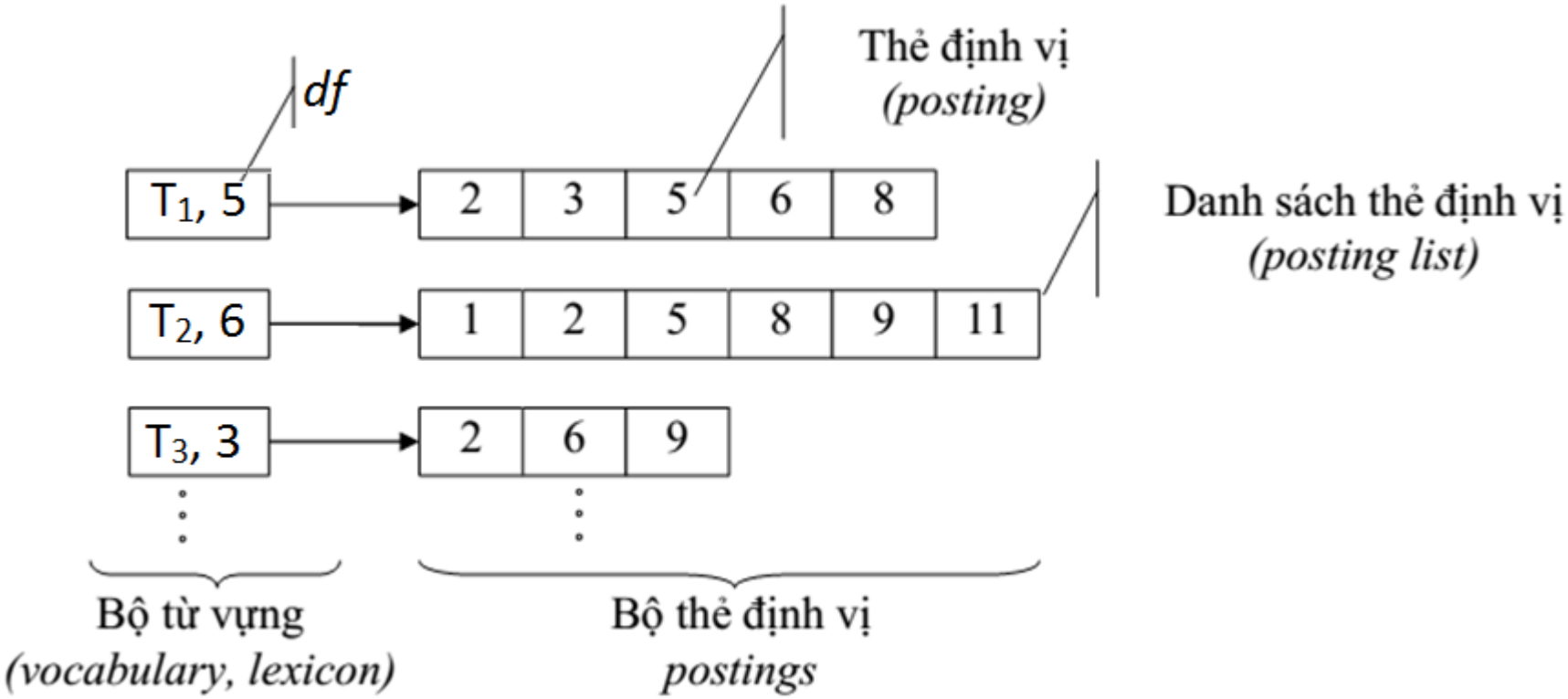
Nhược điểm: nhưng sẽ cần rất nhiều bộ nhớ;

Phương hướng giải quyết nhược điểm: chỉ mục ngược.

Chỉ mục ngược (1)

- Ý tưởng: Gần giống với ma trận đánh dấu, chỉ lưu các giá trị 1.
  - Tối ưu hơn ma trận đánh dấu về mặt lưu trữ;
  - Thực hiện truy vấn trên các danh sách:
    - Không thực hiện phép toán logic trên bit như đối với ma trận đánh dấu;
    - Thực hiện các phép toán tập hợp trên danh sách: lấy phần tử chung của hai danh sách (∩), kết hợp hai danh sách (∪);
    - Nếu sắp xếp văn bản theo trật tự tăng dần mã văn bản, thì có thể thực hiện truy vấn với độ phức tạp tuyến tính.

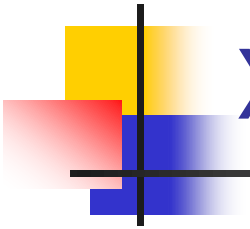
Chỉ mục ngược (2)



Từ ngược trong chỉ mục ngược có nghĩa gì?

Chỉ mục ngược (3)

- Các thuật ngữ:
  - Mỗi mục từ là một bộ ba gồm một từ duy nhất trong bộ từ vựng, *df* và con trỏ tới danh sách thẻ định vị của từ đó;
  - Thẻ định vị, là một cấu trúc lưu thông tin tương ứng với một văn bản (mã văn bản, các vị trí xuất hiện từ, v.v.). Từ định vị mang ý nghĩa xác định vị trí xuất hiện của từ;
  - Tất cả các danh sách thẻ định vị gộp lại được gọi chung là bộ thẻ định vị.



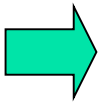
Xây dựng chỉ mục ngược

Các bước cơ bản xây dựng chỉ mục ngược trong bộ nhớ:

Tách từ→ Sinh thẻ định vị → Sắp xếp thẻ định vị → Tổng hợp danh sách thẻ định vị → Lưu bộ từ vựng và bộ thẻ định vị

Tách từ

- D1. DMPLK là tác phẩm văn xuôi đặc sắc và nổi tiếng nhất của Tô Hoài viết về loài vật, dành cho lứa tuổi thiếu nhi
- D2. **Tô Hoài** (sinh ngày 27-9-1920) là một nhà văn Việt Nam nổi tiếng. Một số tác phẩm đề tài thiếu nhi của ông được dịch ra ngoại ngữ.



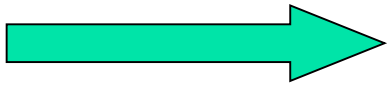
- D1. DMPKL| là | tác phẩm | văn xuôi | đặc sắc | và | nổi tiếng nhất | của | Tô Hoài | viết về | loài vật | dành cho | lứa tuổi thiếu nhi
- D2. Tô Hoài | sinh ngày | 27-9-1920 | là một | nhà văn | Việt Nam | nổi tiếng | Một số | tác phẩm | đề tài | thiếu nhi | của ông | được | dịch ra | ngoại ngữ

*Ký hiệu viết tắt trong slide: DMPLK: Dế mèn phiêu lưu kí

Sinh thẻ định vị

- D1. DMPKL| là | tác phẩm | văn xuôi | đặc sắc | và | nổi tiếng nhất | của | Tô Hoài | viết về | loài vật | dành cho | lứa tuổi thiếu nhi
- D2. Tô Hoài | sinh ngày | 27-9-1920 | là một | nhà văn | Việt Nam | nổi tiếng | Một số | tác phẩm | đề tài | thiếu nhi | của ông | được | dịch ra | ngoại ngữ

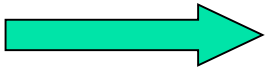
Sinh thẻ định vị



Từ	Mã văn bản
<DMPLK,	1>
<là,	1>
<tác phẩm,	1>
<văn xuôi,	1>
<đặc sắc,	1>
<và ,	1>
<nổi tiếng nhất,	1>
<của,	1>
<Tô Hoài,	1>
<viết về,	1>
<loài vật,	1>
<dành cho,	1>
<lứa tuổi thiếu nhi,	1>
<Tô Hoài,	2>
<sinh ngày,	2>
<27-9-1920,	2>
<là một,	2>
<nhà văn,	2>
<Việt Nam,	2>
<nổi tiếng,	2>
<Một số,	2>
<tác phẩm,	2>
<đề tài,	2>
<thiếu nhi,	2>
<của ông,	2>
<được,	2>
<dịch ra,	2>
<ngoại ngữ,	2>

Từ	Mã văn bản
<DMPLK,	1>
<là,	1>
<tác phẩm,	1>
<văn xuôi,	1>
<đặc sắc,	1>
<và ,	1>
<nổi tiếng nhất,	1>
<của,	1>
<Tô Hoài,	1>
<viết về,	1>
<loài vật,	1>
<dành cho,	1>
<lứa tuổi thiếu nhi,	1>
<Tô Hoài,	2>
<sinh ngày,	2>
<27-9-1920,	2>
<là một,	2>
<nhà văn,	2>
<Việt Nam,	2>
<nổi tiếng,	2>
<Một số,	2>
<tác phẩm,	2>
<đề tài,	2>
<thiếu nhi,	2>
<của ông,	2>
<được,	2>
<dịch ra,	2>
<ngoại ngữ,	2>

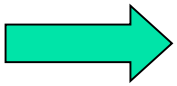
Sắp xếp



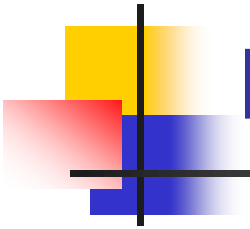
Từ	Mã văn bản
<27-9-1920,	2>
<DMPLK,	1>
<Một số,	2>
<Tô Hoài,	1>
<Tô Hoài,	2>
<Việt Nam,	2>
<của,	1>
<của ông,	2>
<dành cho,	1>
<dịch ra,	2>
<đặc sắc,	1>
<đề tài,	2>
<được,	2>
<là,	1>
<là một,	2>
<loài vật,	1>
<lứa tuổi thiếu nhi,	1>
<ngoại ngữ,	2>
<nhà văn,	2>
<nổi tiếng,	2>
<nổi tiếng nhất,	1>
<sinh ngày,	2>
<tác phẩm,	1>
<tác phẩm,	2>
<thiếu nhi,	2>
<và ,	1>
<văn xuôi,	1>
<viết về,	1>

Từ	Mã văn bản
<27-9-1920,	2>
<DMPLK,	1>
<Một số,	2>
<Tô Hoài,	1>
<Tô Hoài,	2>
<Việt Nam,	2>
<của,	1>
<của ông,	2>
<dành cho,	1>
<dịch ra,	2>
<đặc sắc,	1>
<đề tài,	2>
<được,	2>
<là,	1>
<là một,	2>
<loài vật,	1>
<lứa tuổi thiếu nhi,	1>
<ngoại ngữ,	2>
<nhà văn,	2>
<nổi tiếng,	2>
<nổi tiếng nhất,	1>
<sinh ngày,	2>
<tác phẩm,	1>
<tác phẩm,	2>
<thiếu nhi,	2>
<và ,	1>
<văn xuôi,	1>
<viết về,	1>

Tổng hợp danh sách



Từ , df		danh sách thẻ định vị
27-9-1920, 1	→	2
…		
Tô Hoài, 2	→	1, 2
…		
tác phẩm, 2	→	1, 2
…		
văn xuôi, 1	→	1
viết về, 1	→	1



Lưu bộ từ vựng và bộ thẻ định vị

- Bộ từ vựng và bộ thẻ định vị thường được lưu tách biệt
  - Có thể nén bộ từ vựng và bộ thẻ định vị;
  - Giải thuật nén bộ từ vựng khác giải thuật nén bộ thẻ định vị.

Bài tập 1.1

Cho các văn bản sau:

- **Doc1:** [breakthrough drug for schizophrenia]
- **Doc2:** [new schizophrenia drug]
- **Doc3:** [new approach for treatment of schizophrenia]
- **Doc4:** [new hopes for schizophrenia patients]

a) Vẽ biểu diễn chỉ mục ngược;

b) Các văn bản nào sẽ được trả về cho truy vấn:

- schizophrenia AND drug
- for AND NOT(drug OR approach)