IT4853
Tìm kiếm và trình diễn thông tin

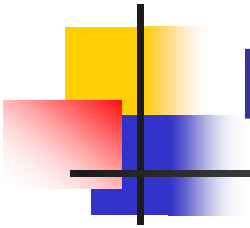
Bài 2. Thực hiện truy vấn trên chỉ mục ngược

IIR.C1. Boolean retrieval

IIR.C2. The term vocabulary and postings lists

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

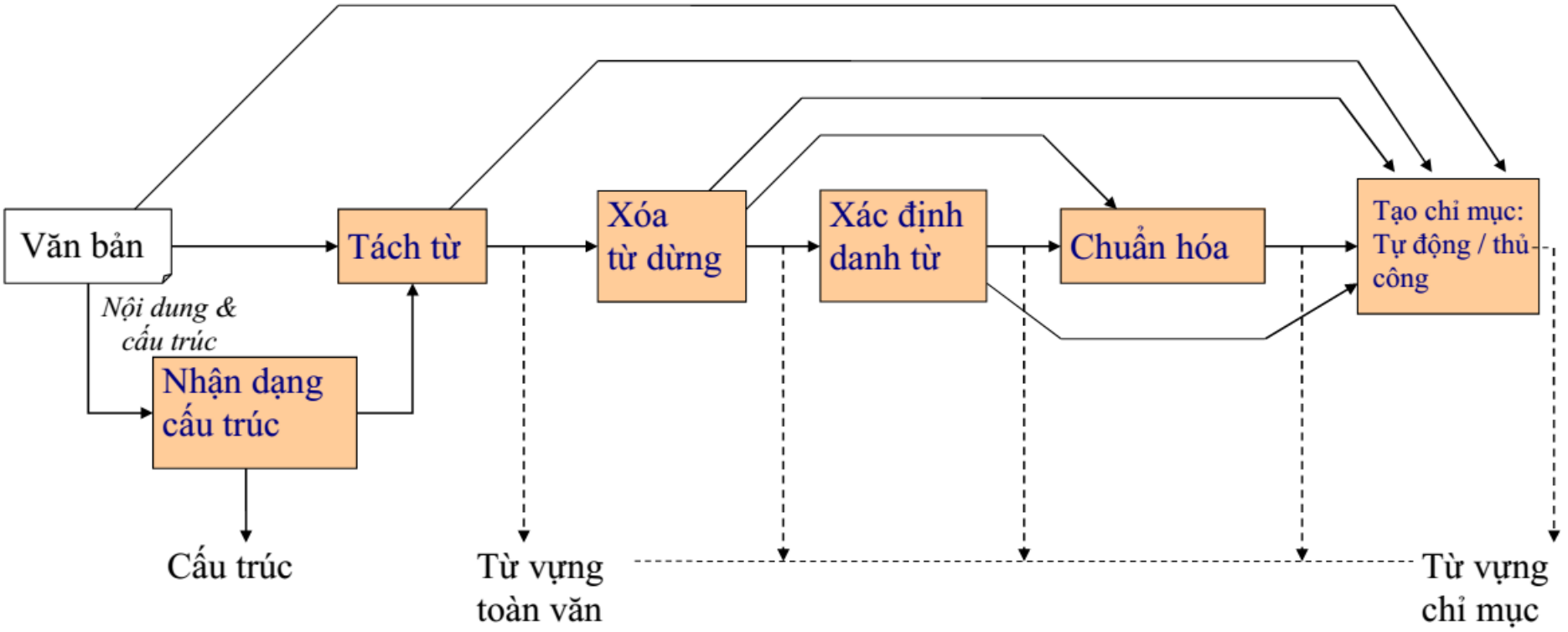
Hà Nội, 2016

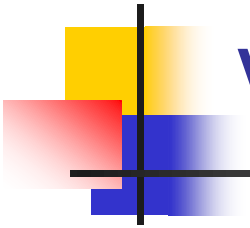


Nội dung chính

- 1. Bộ từ vựng
- 2. Thực hiện truy vấn Boolean
  - Thực hiện truy vấn AND trên chỉ mục ngược
  - Cải tiến giải thuật lấy giao hai danh sách
  - Trình tự tối ưu thực hiện truy vấn Boolean

Quy trình đọc văn bản





Vấn đề chuẩn hóa từ tiếng Việt

- Chuẩn hóa bảng mã
- Chuẩn hóa dấu ngữ âm
- V.v.

Mã hóa tiếng việt Unicode

- Tổ hợp (composite)

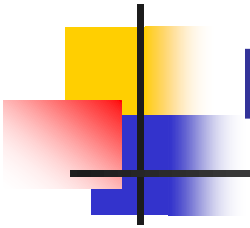
a ă â e ê i o ô ơ u ư y

◌̀ ◌̉ ◌̃ ◌́ ◌̣

- Dựng sẵn (precomposed)

1EA0	Ạ	Chữ A hoa với dấu nặng

- TCVN 6909:2001



Nội dung chính

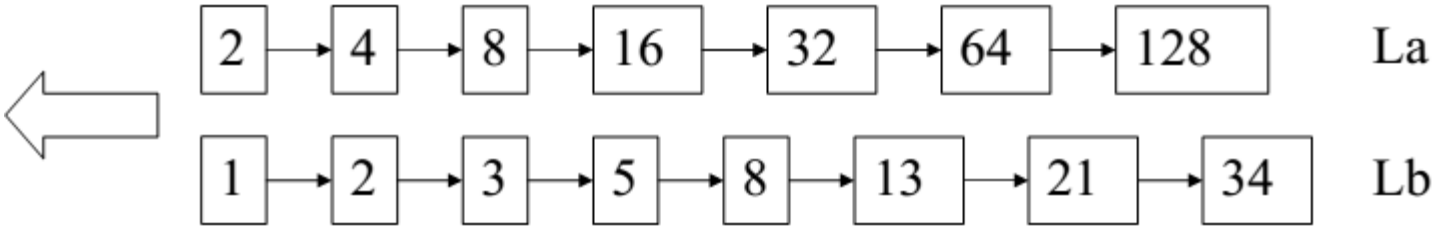
- 1. Bộ từ vựng
- 2. Thực hiện truy vấn Boolean
  - Thực hiện truy vấn AND trên chỉ mục ngược
  - Cải tiến giải thuật lấy giao hai danh sách
  - Trình tự tối ưu thực hiện truy vấn Boolean

Truy vấn AND

- Các bước thực hiện truy vấn kiểu:

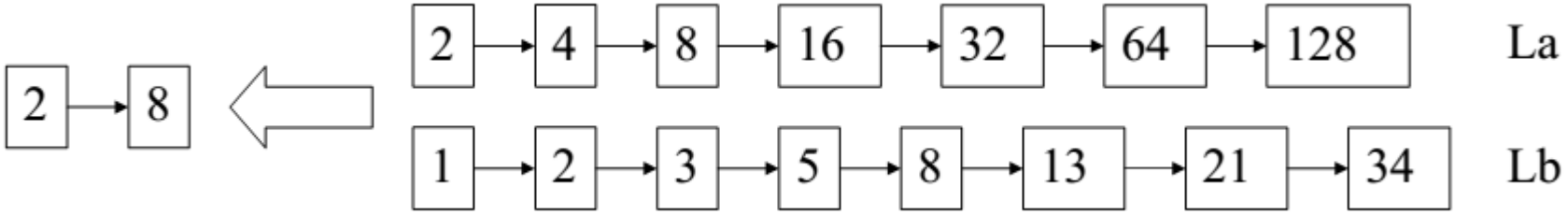
a AND *b*

1. Tìm *a* trong từ điển và lấy danh sách thẻ định vị *La*
2. Tìm *b* trong từ điển và lấy danh sách thẻ định vị *Lb*
3. Lấy các phần tử chung (giao) của *La* và *Lb*



Lấy giao của hai danh sách

- Duyệt đồng thời cả hai danh sách



Thuật toán 2.1.

Nếu các danh sách được sắp xếp theo mã văn bản, thì số lượng so sánh không vượt quá *La* + *Lb*.

Thuật toán 2.1

```
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
```

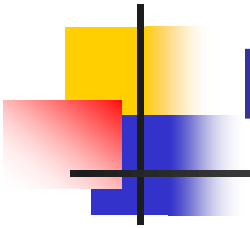
Minh họa thuật toán

2, 4, 8, 16, 32, 64, 128 La

1, 2, 3, 5, 8, 13, 21, 34 Lb

answer = {2, 8}

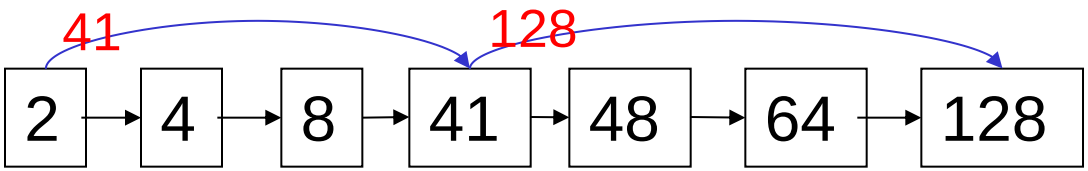
La	*Lb*	*answer*
2	1	
2	2	2
4	3	
4	5	
8	5	
8	8	2, 8
16	13	
16	21	
32	21	
32	34	
64	34	
64	NIL	

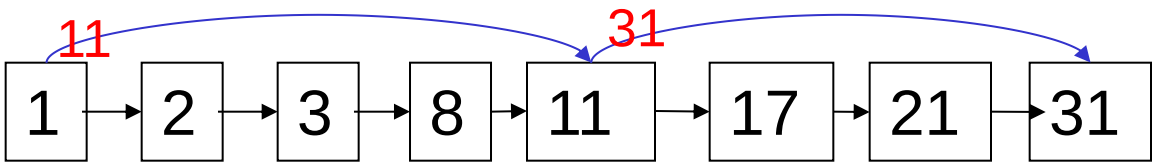


Nội dung chính

- 1. Bộ từ vựng
- 2. Thực hiện truy vấn Boolean
  - Thực hiện truy vấn AND trên chỉ mục ngược
  - Cải tiến giải thuật lấy giao hai danh sách
  - Trình tự tối ưu thực hiện truy vấn Boolean

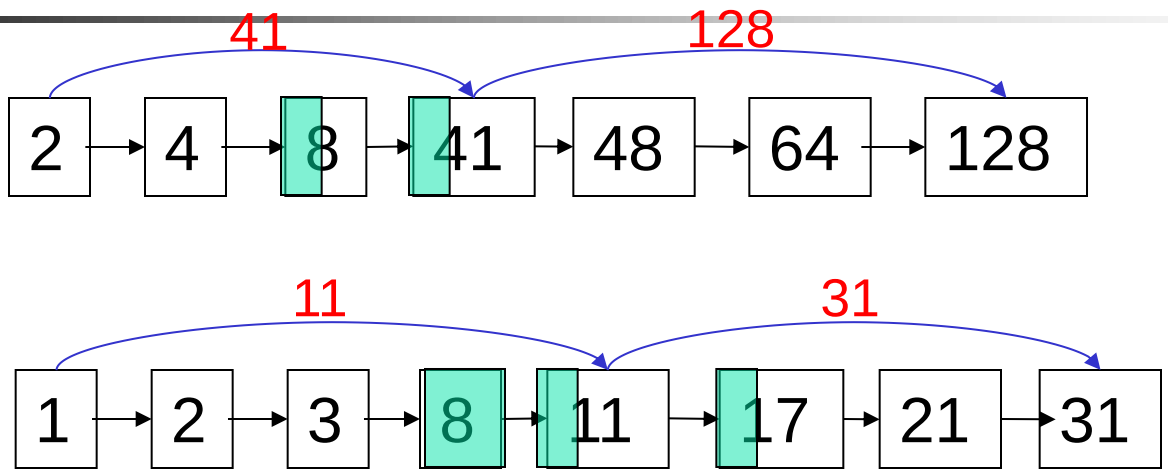
Sử dụng bước nhảy





- Bổ xung bước nhảy vào danh sách thẻ định vị;
- Sử dụng bước nhảy để bỏ qua những thẻ định vị không thỏa mãn điều kiện.

Lấy giao hai danh sách có bước nhảy



Giả sử trong quá trình duyệt danh sách, các con trỏ đang ở vị trí số 8 ở cả hai danh sách, các thao tác là:

Lưu giá trị 8 và,

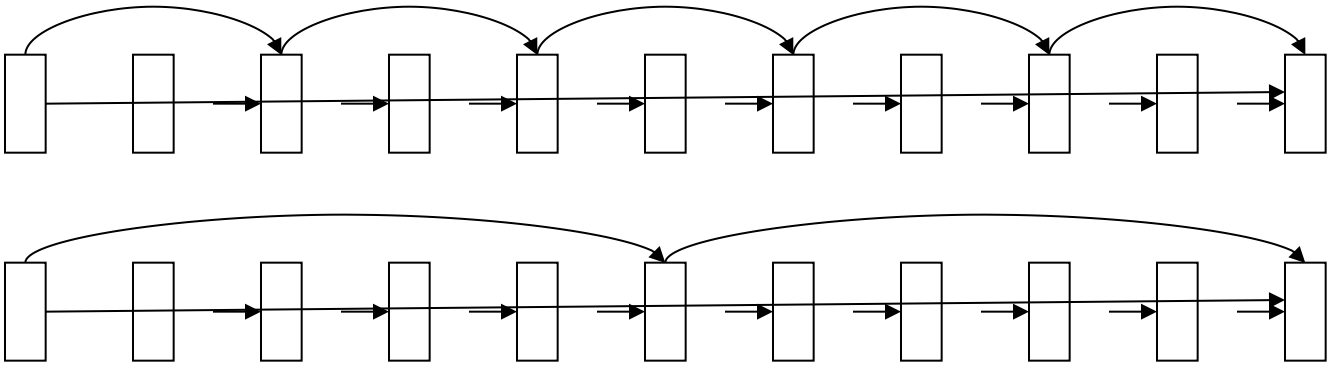
Dịch chuyển con trỏ sang phải ở cả hai danh sách, vị trí mới là (41, 11),

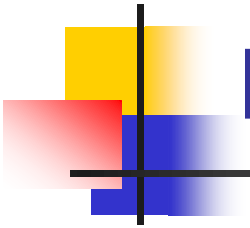
Thực hiện bước nhảy (vì 31 < 41), và kết thúc giải thuật.

Trong trường hợp này chúng ta đã bỏ qua một phần danh sách

Độ dài của bước nhảy

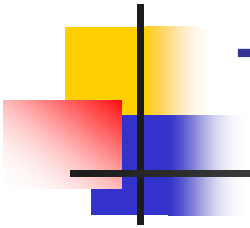
- Nếu nhiều bước nhảy$\rightarrow$ khoảng cách nhỏ $\Rightarrow$ xác suất di chuyển theo bước nhảy cao. Nhưng phải so sánh bước nhảy nhiều lần.
- Ít bước nhảy $\rightarrow$ ít so sánh hơn, nhưng khoảng cách lớn hơn $\Rightarrow$ xác suất di chuyển theo bước nhảy thấp hơn.





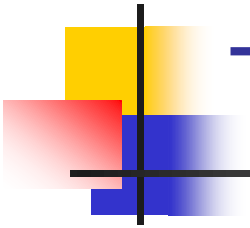
Nội dung chính

- 1. Bộ từ vựng
- 2. Thực hiện truy vấn Boolean
  - Thực hiện truy vấn AND trên chỉ mục ngược
  - Cải tiến giải thuật lấy giao hai danh sách
  - Trình tự tối ưu thực hiện truy vấn Boolean



Tối ưu hóa truy vấn AND

- Số kết quả không lớn hơn độ dài danh sách thẻ định vị ngắn nhất



Tối ưu hóa truy vấn AND

1. Với mỗi thuật ngữ truy vấn *t*
   - Tìm *t* trong bộ từ vựng
2. Sắp xếp thuật ngữ tăng dần theo *df*(*t*)
3. Khởi tạo tập kết quả *answer* là danh sách ngắn nhất
4. Tiếp tục thực hiện truy vấn theo thứ tự đã sắp xếp

Ví dụ

- Cho truy vấn a AND b AND c với các danh sách thẻ định vị như trong hình vẽ

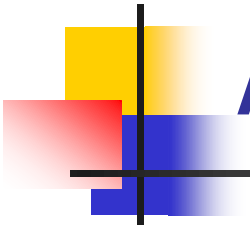
2	4	8	16	32	64	128		La

1	2	3	5	8	16	21	34	Lb

13	16							Lc

Thứ tự tối ưu với truy vấn a AND b AND c là

(c AND a) AND b



AND of OR's

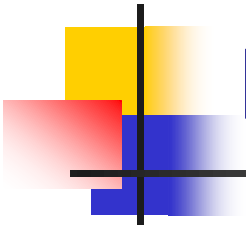
- Ví dụ truy vấn dạng AND of OR's:

  (văn bản OR dữ liệu OR hình ảnh) AND

  (nén OR gom nhóm) AND

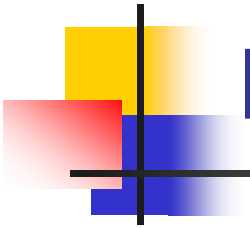
  (tìm kiếm OR đánh chỉ mục OR lưu trữ)

- Tối ưu hóa truy vấn
  - Lấy độ dài danh sách thẻ vị trí cho mỗi từ
  - Ước lượng số kết quả cho mỗi truy vấn OR
  - Sắp xếp các truy vấn OR theo thứ tự tăng dần số lượng kết quả



Bài tập 2.1

- Đối với truy vấn AND, thứ tự tăng dần theo độ dài danh sách thẻ định vị có luôn là thứ tự tối ưu hay không? Chứng minh?



Bài tập 2.2

- Tương tự thuật toán 2.1, hãy viết thuật toán thực hiện các truy vấn dạng a OR b và a AND NOT b với độ phức tạp tuyến tính.

Bài tập 2.3

- Những phát biểu sau đây đúng hay sai?

  a. Trong mô hình tìm kiếm Boolean, loại bỏ dấu không bao giờ làm giảm tính chính xác.

  b. Trong mô hình tìm kiếm Boolean, loại bỏ dấu không bao giờ làm giảm tính đầy đủ.

  c. Loại bỏ dấu làm tăng kích thước bộ từ vựng.

  d. Nên thực hiện các thao tác chuẩn hóa trong quá trình xây dựng chỉ mục thay vì khi thực hiện truy vấn.

Bài tập 2.4

- Đề xuất một trình tự thực hiện cho truy vấn

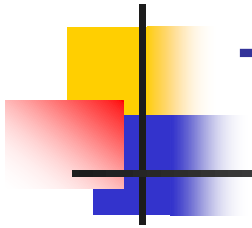
(tangerine** OR **trees)** AND **(marmalade** OR **skies)** AND **(kaleidoscope** OR **eyes)

- Chúng ta nên xử lý biểu thức nào trước tiên?

Từ khóa	Tần suất
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Bài tập 2.5

- Cho truy vấn:
- (Brutus OR Caesar) AND NOT (Antony OR Cleopatra)
- a) Hãy sử dụng luật phân tích và viết lại truy vấn đã cho dưới dạng OR of ANDs (disjuntive normal form).
- b) Truy vấn thu được ở mục a hiệu quả hơn hay kém truy vấn ban đầu?
- c) Kết luận này đúng trong trường hợp tổng quát? Hay là còn phụ thuộc vào từ khóa và nội dung văn bản?



Tham khảo

- IIR 1.5:
  - Memex
  - Information retrieval term