Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

Nguyễn Đình Hóa
dinhhoa@gmail.com 0942807711

Tóm tắt nội dung bài 7

- Siêu dữ liệu và CSDL ĐPT
  - Sự khác biệt giữa CSDL ĐPT và thế giới thực
    - Cảm nhận
    - Ngữ nghĩa
    - Đa đạng dữ liệu
  - Khái niệm siêu dữ liệu đa phương tiện
    - Miêu tả (description)
    - Nhận diện (identification)
    - Truyền đạt (interpretation)
  - Từ điển CSDP ĐPT

Chỉ số hóa dữ liệu văn bản

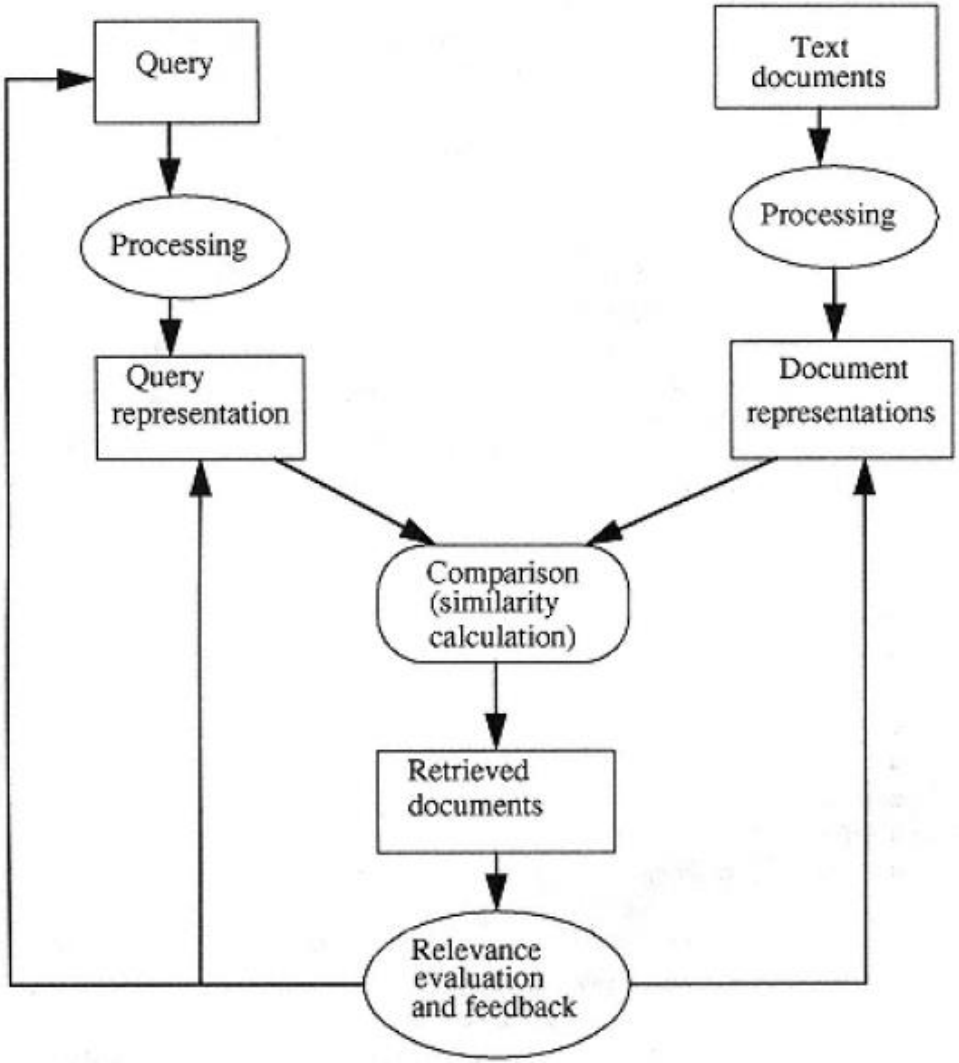
- Chỉ số hóa (indexing): tìm những điểm đặc trưng nhất của văn bản để lưu trữ và tìm kiếm thông tin.
- Sự khác biệt giữa Hệ quản trị CSDL (DBMS) với hệ thống chỉ số hóa và truy vấn dữ liệu (Indexing and Retrieval systems)

DBMS	IR Systems
Các bản ghi có cấu trúc đồng nhất bao gồm bộ các thuộc tính cố định để mô tả rõ ràng và tổng quát thông tin của từng bản ghi.	Các bản ghi không lưu trữ theo cấu trúc, không có các thuộc tính cố định. Văn bản được đánh chỉ số dựa trên các từ khóa, thông tin miêu tả, từ đánh dấu, … dùng để diễn tả nội dung theo một khía cạnh nào đó.

Chỉ số hóa dữ liệu văn bản

DBMS	IR Systems
Tra cứu dữ liệu dựa trên sự so khớp dữ liệu trên từng thuộc tính	Tra cứu dựa trên sự tương đồng về nội dung giữa lệnh truy vấn và nội dung văn bản
Kết quả tra cứu phải chính xác với câu truy vấn	Kết quả tra cứu không nhất thiết đúng với nội dung truy vấn

Chỉ số hóa dữ liệu văn bản



Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Hệ thống truy vấn Boolean
  - Các văn bản được đánh chỉ số bởi các bộ từ khóa
  - Các câu truy vấn cũng bao gồm các từ khóa kèm với các phép toán logic (boolean)
    - AND (term 1 AND term 2)
    - OR (term 1 OR term 2)
    - NOT (term 1 AND NOT term 2) thường dùng để giới hạn kết quả tìm kiếm.

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Cấu trúc tệp dữ liệu
  - Tệp phẳng (Flat files):
    - Chứa một hoặc vài văn bản trong cùng một file
    - Văn bản không được chỉ số hóa
    - Tìm kiếm thông qua so khớp khối văn bản (pattern)
  - Tệp đặc trưng (Signature files):
    - Chứa các đặc trưng (chuỗi bit) về thông tin văn bản
    - Có nhiều cách để trích đặc trưng cho từng văn bản
    - Câu truy vấn cũng bao gồm các đặc trưng cần tìm kiếm

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Tệp ngược (Inverted files): là cấu trúc thông dụng nhất
  - Mỗi từ khóa (thuật ngữ - term) được gán một chỉ số riêng
  - Mỗi văn bản (bản ghi) được gán một số nhận dạng (ID)
  - Mỗi từ khóa được dùng để lưu trữ toàn bộ các bản ghi có chứa nó
  - Các bản ghi có chứa cùng một từ khóa được lưu theo một hàng ứng với từ khóa đó
  - Ví dụ
    - Term 1: Record 1, Record 3
    - Term 2: Record 1, Record 2
    - Term 3: Record 2, Record 3, Record 4
    - Term 4: Record 1, Record 2, Record 3, Record 4

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Tệp ngược: sử dụng mô hình Boolean để truy vấn dữ liệu
  - AND: Term 1 AND Term 3 cho kết quả Record 3
  - OR: Term 1 OR Term 2 cho kết quả: Record 1, Record 2, Record 3
  - NOT: Term 4 AND NOT Term 1 cho kết quả: Record 2, Record 4

- 2 vấn đề cần quan tâm khi tra cứu theo từ khóa (term):
  - vị trí của từ khóa trong văn bản,
  - độ quan trọng của từ khóa đối với nội dung của văn bản

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Các thông tin về vị trí của các từ khóa cần phải thêm vào trong tệp.
  - Term *i:* Record no., Paragraph no., Sentence no., Word no.
  - VD:
    - *world: R99, 10, 8, 3; R155, 15, 3, 6; R166, 2, 3, 1*
    - *battle: R77, 9, 7, 2; R99, 10, 8, 6; R166, 10, 2, 5; R166, 2, 3, 2*
    - Truy vấn: **“world” trong cùng một câu với “battle”**
- Chỉ số về khoảng cách
  - “Trong cùng một câu”
    - Term 1 trong cùng một câu với Term 2
  - “Liền kề”
    - Term 1 liền kề với Term 2

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Đánh chỉ số tự động cho từ khóa văn bản
  - Loại bỏ các stop words: là các từ/mục không quan trọng hoặc hữu ích

Excerpt of Common Stop Words

A	ALTHOUGH	ANYONE
ABOUT	ALWAYS	ANYTHING
ACROSS	AMONG	ANYWHERE
AFTER	AMONGST	ARE
AFTERWARDS	AN	AROUND
AGAIN	AND	AS
AGAINST	ANOTHER	AT
ALL	ANY	BE
ALSO	ANYHOW	BECOME

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Đánh chỉ số tự động
  - Rút gọn: các từ có cùng một nghĩa hoặc chung gốc từ được thay thế bởi một từ duy nhất (từ gốc)
    - VD: retrieval, retrieved, retrieving, retrieve -> retriev
  - Tìm các từ đồng nghĩa, chọn một từ đại diện
    - VD: study, learning, schoolwork, reading,… -> study
  - Đánh trọng số cho các từ khóa
    - VD:
      - Term1: R1, 0.3; R3, 0.5; R6, 0.8; R7, 0.2; R11, 0.1
      - Term2: R2, 0.7; R3, 0.6; R7, 0.5; R9, 0.5
      - Term3: R1, 0.8; R2, 0.4; R9, 0.6

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Truy vấn Boolean với từ khóa có trọng số
  - OR: Nếu một bản ghi có nhiều từ khóa truy vấn thì trọng số cao nhất sẽ được chọn. Các bản ghi được sắp xếp theo thứ tự trọng số từ cao xuống thấp.
    - Term1: R1, 0.3; R3, 0.5; R6, 0.8; R7, 0.2; R11, 0.1
    - Term2: R2, 0.7; R3, 0.6; R7, 0.5; R9, 0.5
    - Term3: R1, 0.8; R2, 0.4; R9, 0.7
    - Truy vấn: **Term 2 OR Term 3** trả về kết quả R1, R2, R9, R3, R7
  - AND: trọng số thấp sẽ được chọn cho bản ghi có nhiều từ khóa truy vấn
    - VD: truy vấn **Term 2 AND Term 3** trả về kết quả R9, R2
  - NOT: hiệu số giữa các trọng số của các bản ghi có cùng từ khóa sẽ được sử dụng
    - VD: truy vấn **Term 2 AND NOT Term 3** trả về kết quả R3, R7, R2.

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Cách đánh trọng số
  - Dựa vào tần suất xuất hiện của từ khóa trong từng văn bản (tf_{ij})
  - Dựa vào số văn bản chứa cùng một từ khóa (df_j)
  - $W_{ij} = tf_{ij} * \log(N / df_j)$
    - W_{ij} là trọng số của từ khóa j trong văn bản i
    - tf_{ij} là tần suất của từ khóa j trong văn bản i
    - df_j là tổng số văn bản chứa từ khóa j
    - N là tổng số văn bản trong kho dữ liệu

Chỉ số hóa dữ liệu văn bản và mô hình truy vấn Boolean

- Tổng hợp các bước đánh chỉ số tự động
  - Tìm toàn bộ từ có trong tiêu đề, tóm tắt, nội dung văn bản
  - Loại bỏ các stop words
  - Tìm các từ đồng nghĩa, thay thế bằng một từ chung
  - Tìm từ gốc của các từ có cùng gốc
  - Tìm tần suất xuất hiện của các từ khóa trong từng văn bản
  - Tìm trọng số của từng từ khóa
  - Tạo tệp ngược.