ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

TRUY HỒI THÔNG TIN TRONG VĂN BẢN Y KHOA

GVHD: GS. TS. Cao Hoàng Trụ
GVPB: TS. Nguyễn An Khương

SVTH: Nguyễn Văn Lành

Nội dung

1. Giới thiệu
2. Phương pháp đề xuất
3. Kết quả thí nghiệm
4. Tổng kết

Nội dung

1 Giới thiệu

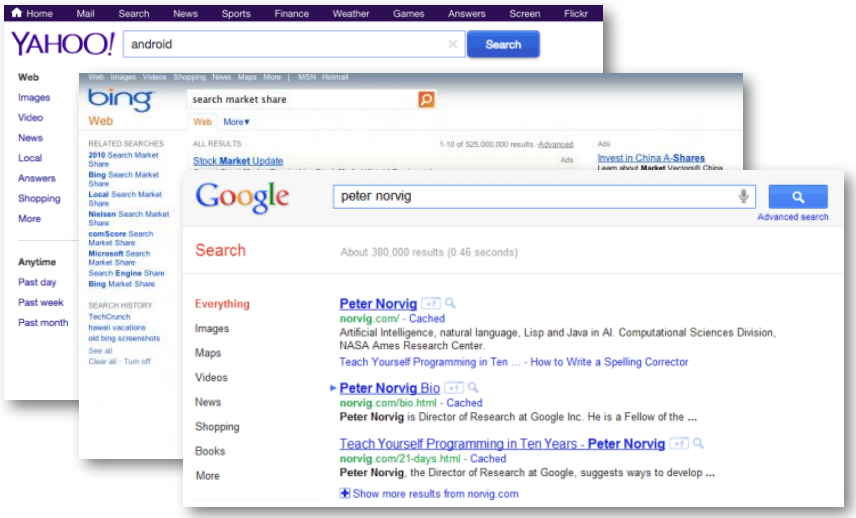
2 Phương pháp đề xuất

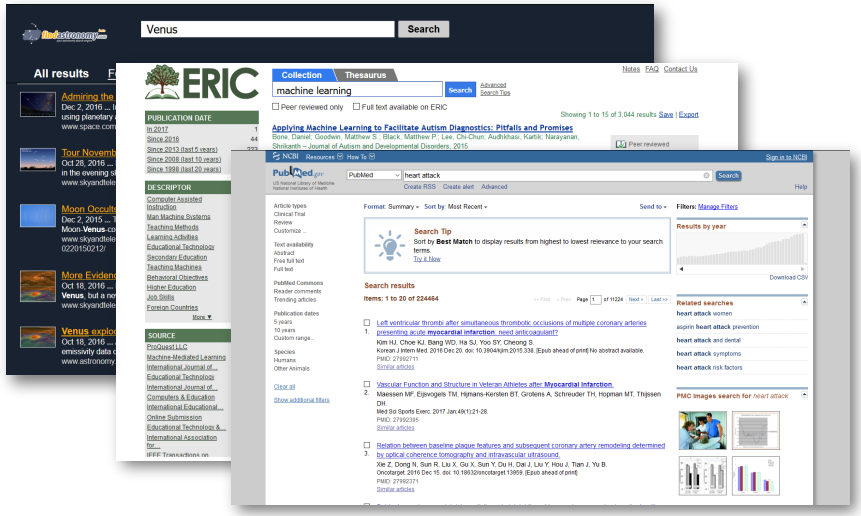
3 Kết quả thí nghiệm

4 Tổng kết

Giới thiệu

- 85% người dùng Internet sử dụng các công cụ tra cứu.
- Các công cụ tìm kiếm phổ biến (*Google, Bing, Yahoo*) hoặc trong từng lĩnh vực chuyên biệt (*FindAstronomy - thiên văn học, Eric - giáo dục, PubMed - y khoa,..*)





Giới thiệu

- Truy hồi thông tin: quá trình tìm kiếm thông tin (chủ yếu dưới dạng phi cấu trúc) được lưu trữ trên các hệ thống máy tính để đáp ứng nhu cầu tìm kiếm của người dùng.

Giới thiệu

- Truy hồi thông tin: quá trình tìm kiếm thông tin (chủ yếu dưới dạng phi cấu trúc) được lưu trữ trên các hệ thống máy tính để đáp ứng nhu cầu tìm kiếm của người dùng.
- Truy hồi thông tin trong lĩnh vực y khoa để trả lời các loại câu hỏi lâm sàng đang được chú trọng.
- Tài liệu trong lĩnh vực y khoa được tổ chức đặc thù và phức tạp $\rightarrow$ vấn đề tìm kiếm đặt ra nhiều thách thức.

Giới thiệu

- Ba loại câu hỏi lâm sàng phổ biến (chiếm 52.72%)*
  - Chẩn đoán (diagnoses)
  - Điều trị (treatments)
  - Xét nghiệm (tests)

Giới thiệu

- Ba loại câu hỏi lâm sàng phổ biến (chiếm 52.72%)*
  - Chẩn đoán (diagnoses)
  - Điều trị (treatments)
  - Xét nghiệm (tests)

- Mẫu câu truy vấn (thông tin truy vấn và loại câu hỏi lâm sàng) và tập danh sách tài liệu theo mức độ phù hợp.

Thông tin bệnh nhân	**Loại câu hỏi**	**Tài liệu**
58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.	Diagnosis	Tài liệu 1 Tài liệu 2 ...

[*] Matthew S. Simpson, Ellen M. Voorhees, and William Hersh. Overview of the TREC 2014 Clinical Decision Support Track. In *Proceedings of the 23rd Text Retrieval Conference (TREC)*, 2014.

Bài toán

- **Nội dung:** Xây dựng hệ thống Truy hồi thông tin trong văn bản y khoa.
- **Đầu vào:**
  - Tập tài liệu (văn bản) y khoa → cơ sở dữ liệu
  - Câu truy vấn gồm 2 phần: thông tin truy vấn (văn bản) và loại câu hỏi lâm sàng.
- **Đầu ra:** Danh sách các tài liệu phù hợp với câu truy vấn được sắp xếp theo mức độ giảm dần.

Nội dung

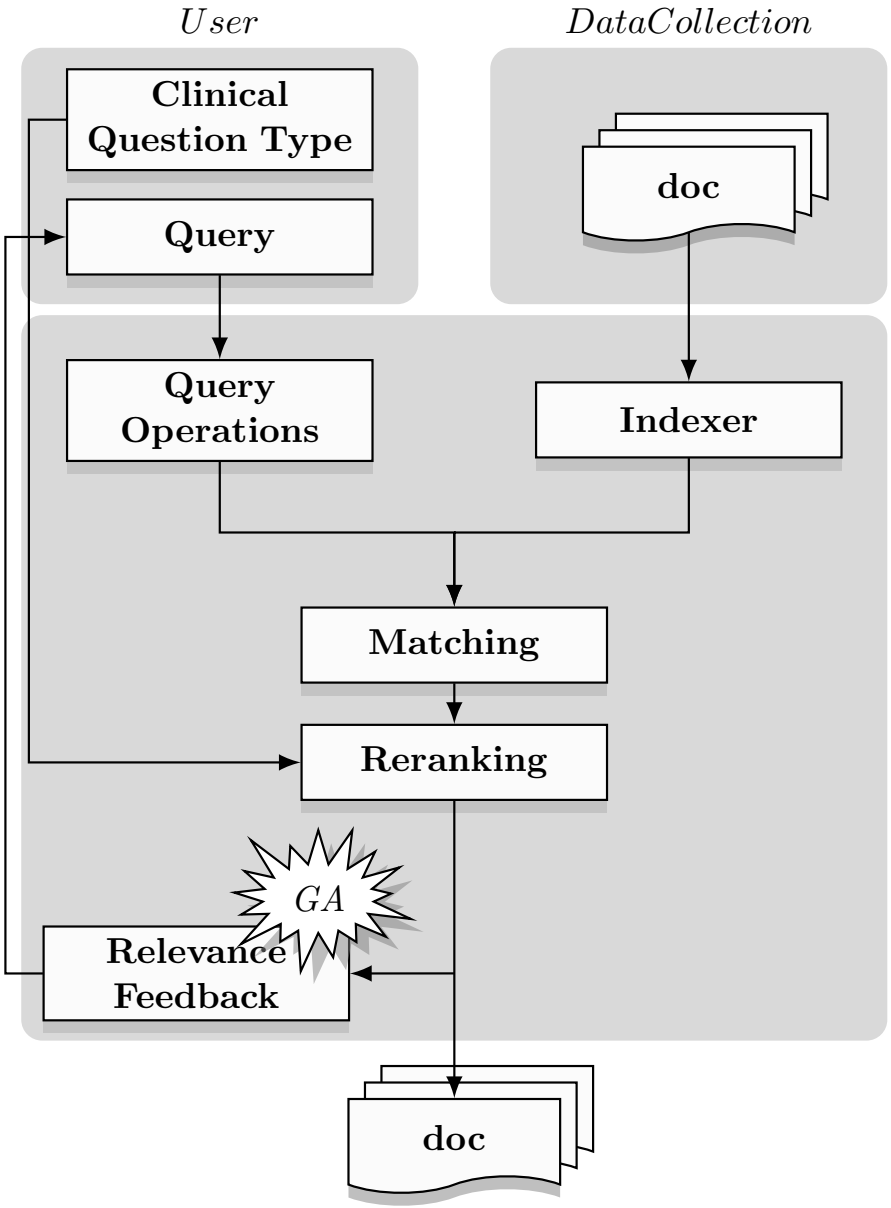
1 Giới thiệu

2 Phương pháp đề xuất

3 Kết quả thí nghiệm

4 Tổng kết

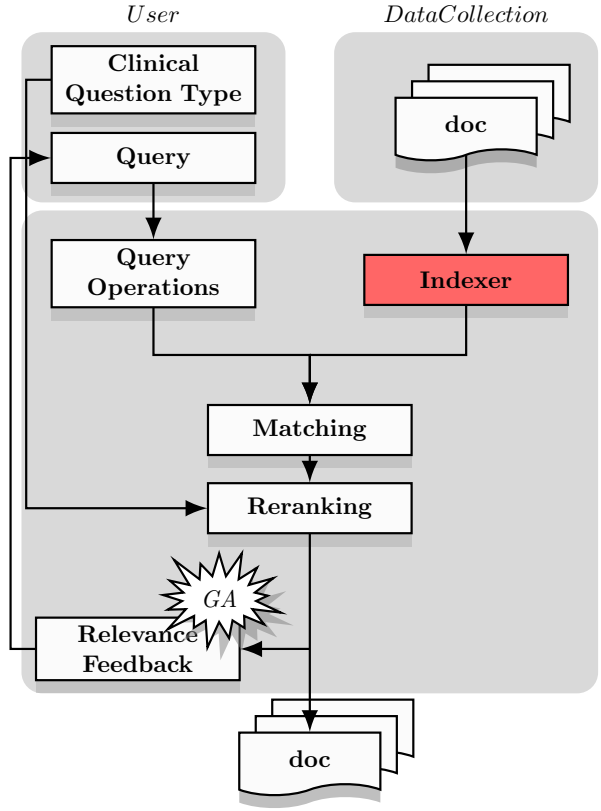
Kiến trúc hệ thống



André Mourão, Flávio Martins and João Magalhães. NovaSearch at TREC 2014 Clinical Decision Support Track. In *Proceedings of the 23rd Text Retrieval Conference (TREC)*, 2014.

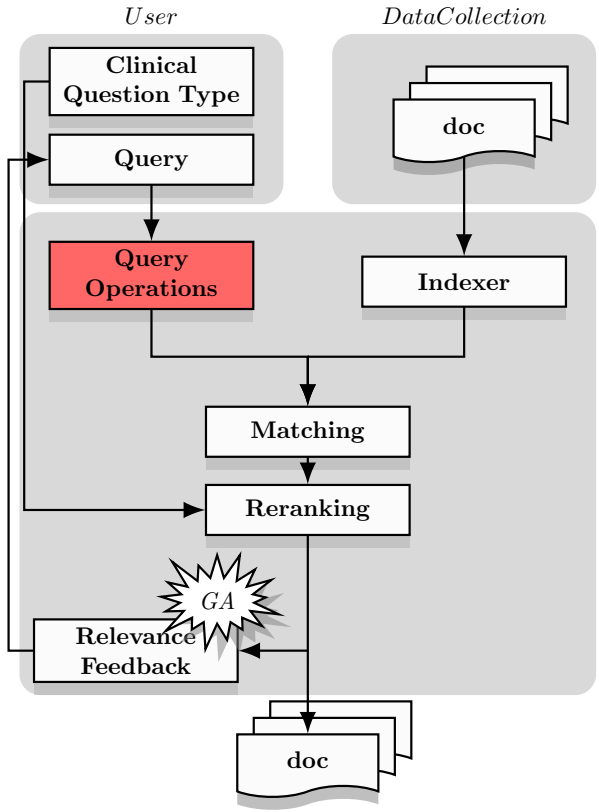
Lập chỉ mục tài liệu

- Quá trình sắp xếp và tổ chức lưu trữ nội dung các tài liệu văn bản trong cơ sở dữ liệu nhằm:
  - Loại bỏ các yếu tố dư thừa
  - Tăng tốc độ tìm kiếm
  - Giảm không gian lưu trữ
- Phương pháp:
  - Loại bỏ các từ dừng (*stop word*), chuẩn hóa từ gốc (*stemming*).
  - Tính trọng số *TF-IDF (Term Frequency-Inverse Document Frequency)* cho các từ.



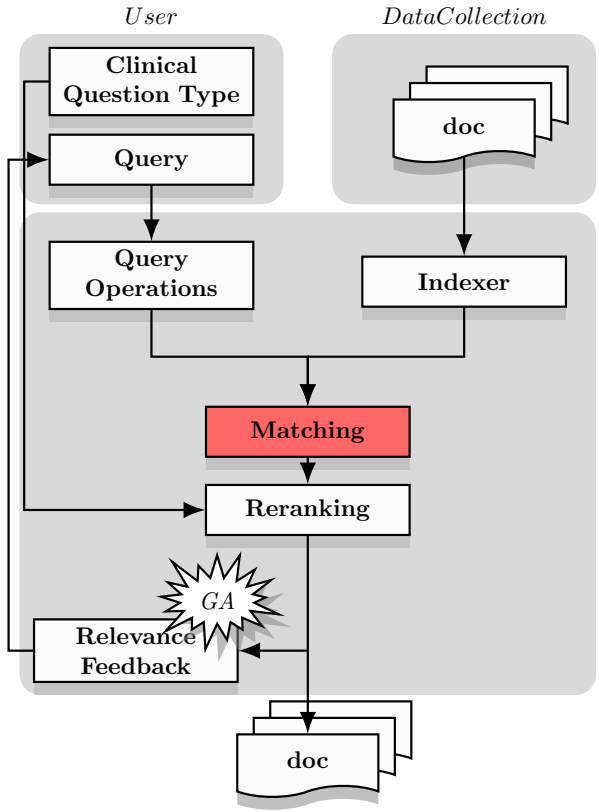
Xử lý câu truy vấn

- Loại bỏ các yếu tố dư thừa và cải thiện chất lượng thông tin truy vấn.
- Phương pháp:
  - Tiền xử lý câu truy vấn
  - Mở rộng câu truy vấn qua từ điển y sinh *MeSH*

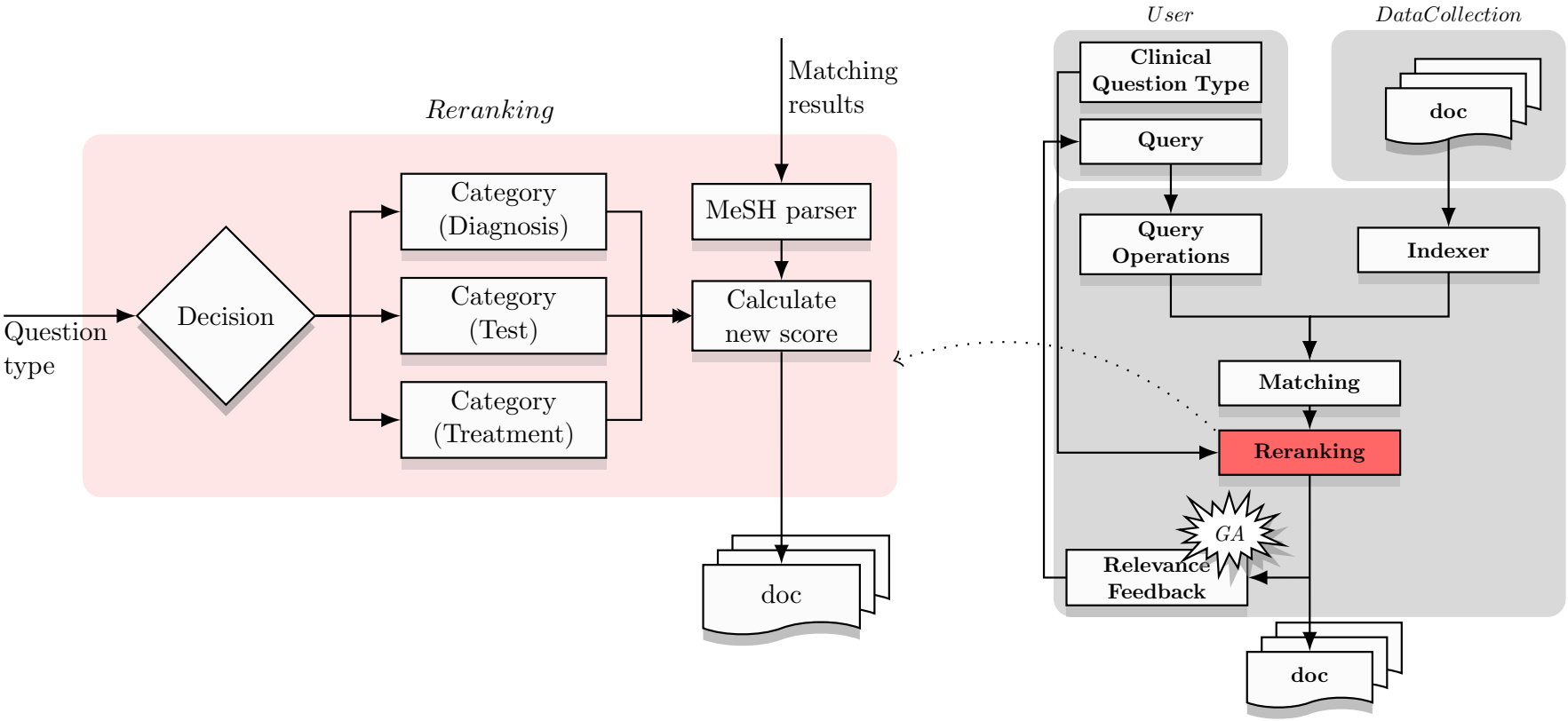


Tìm tài liệu phù hợp

- Các mô hình chấm điểm:
  - Mô hình không gian vector (*Vector Space Model - VSM*)
  - Mô hình tài liệu phù hợp (*Best Matching Model - BM*)
  - Mô hình ngôn ngữ (*Language Model - LM*)



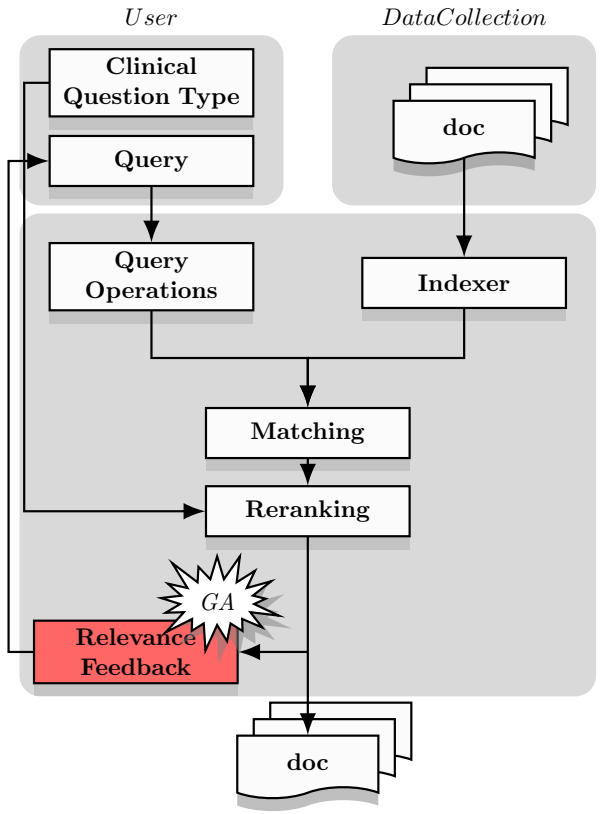
Xếp hạng lại tài liệu



- Sắp xếp các tài liệu phù hợp với loại câu hỏi lâm sàng
  - Tính toán thứ hạng mới dựa trên số lượng từ hạt giống đã phân loại trong phần tóm tắt (*abstract*) tài liệu.
  - Phân loại nhóm từ dựa trên cây phân cấp ▸ MeSH.

Phản hồi sự liên quan

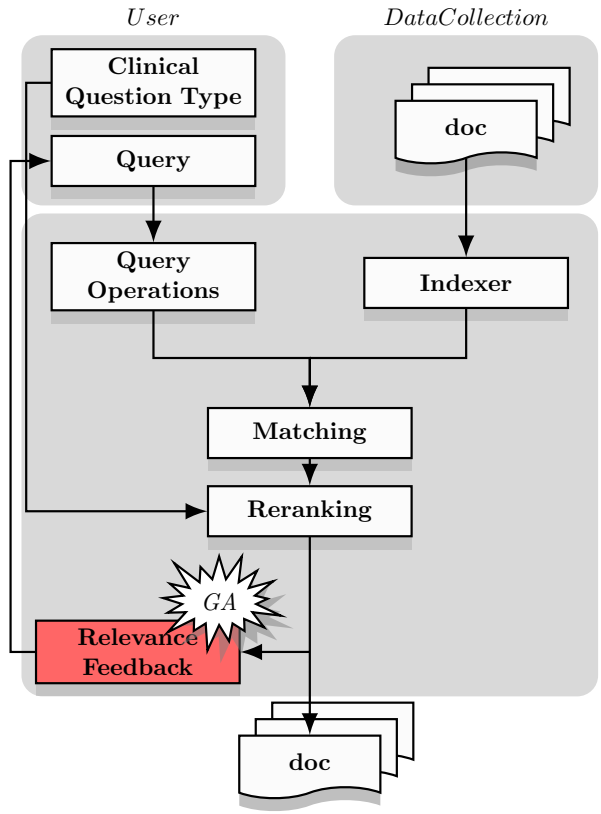
- Cải thiện chất lượng kết quả danh sách tài liệu
  - Trích xuất thông tin từ danh sách các tài liệu phù hợp.
  - Bổ sung từ mới vào câu truy vấn để tiếp tục quá trình tìm kiếm.



Phản hồi sự liên quan

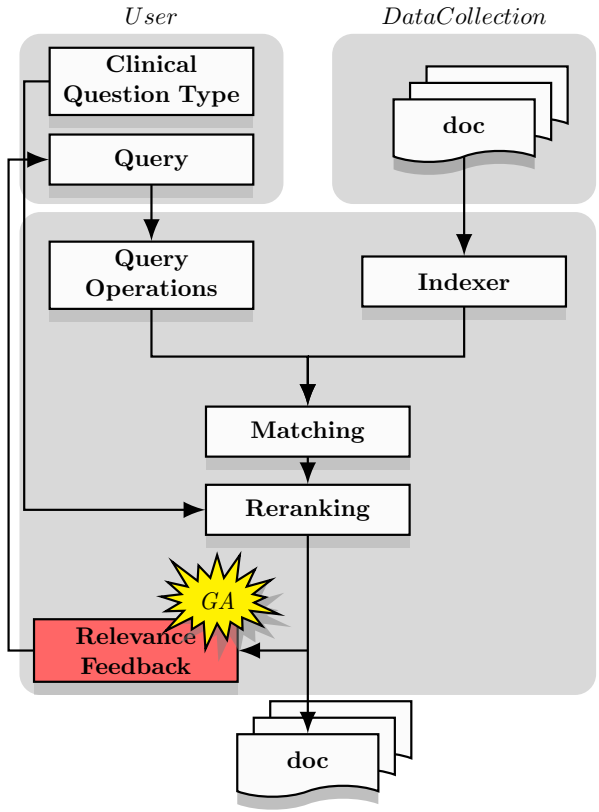
- Cải thiện chất lượng kết quả danh sách tài liệu
  - Trích xuất thông tin từ danh sách các tài liệu phù hợp.
  - Bổ sung từ mới vào câu truy vấn để tiếp tục quá trình tìm kiếm.

→ Giả lập phản hồi sự liên quan (Pseudo Relevance Feedback - PRF)



Giải thuật di truyền trong phản hồi sự liên quan*

- Chọn lọc tập từ vựng phù hợp để bổ sung vào câu truy vấn.
- Phương pháp:
  - Khởi tạo tập từ vựng từ các tài liệu phù hợp trả về (quần thể).
  - Mã hóa các tài liệu được chọn (cá thể)
  - Thao tác trên quần thể để tìm cá thể tốt nhất (bao gồm quá trình chọn lọc, lai hóa và đột biến) qua hệ số tương đồng ▸ Jaccard.
  - Giải mã, bổ sung các từ vựng mới đã được chọn và câu truy vấn.



Nội dung

1 Giới thiệu

2 Phương pháp đề xuất

3 Kết quả thí nghiệm

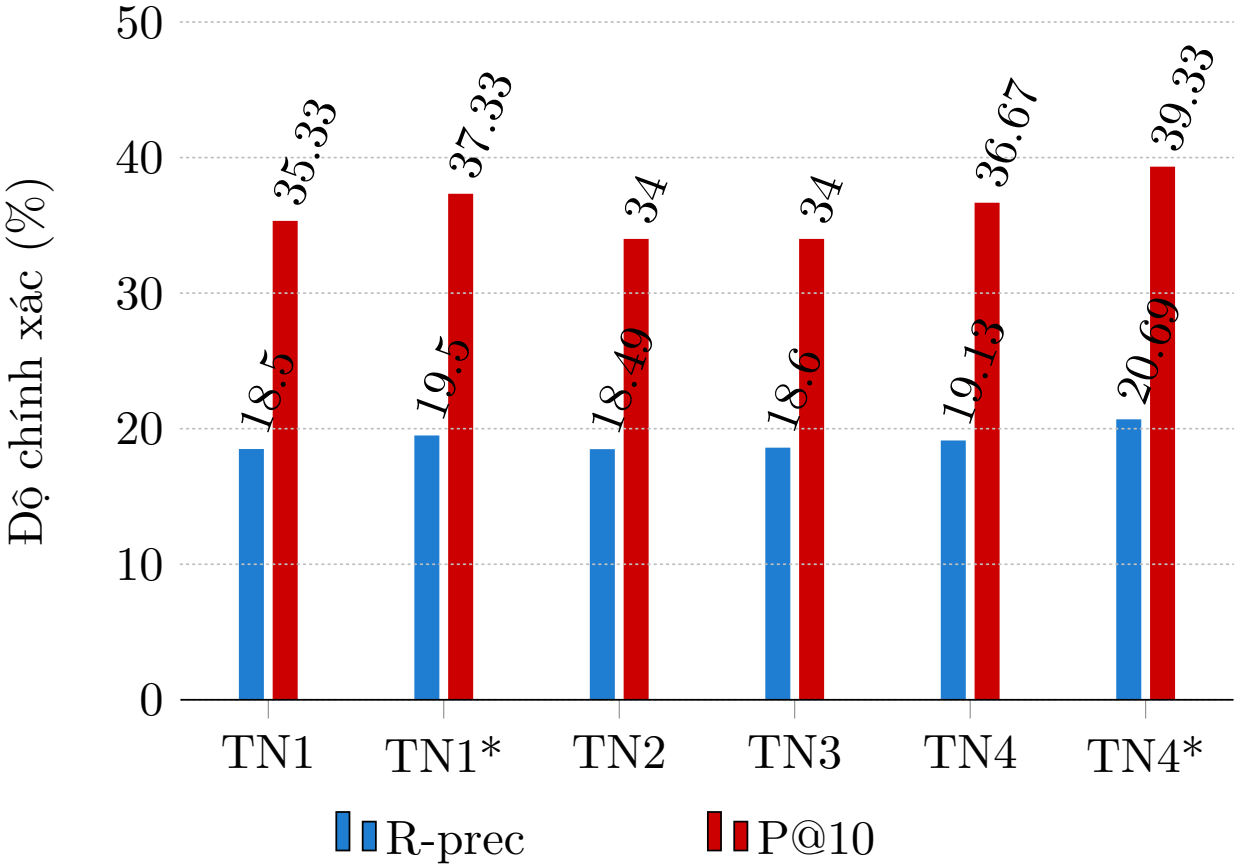
4 Tổng kết

Thí nghiệm

- **Tập dữ liệu:** Được cung cấp bởi *TREC* 2014
  - 733,328 tài liệu văn bản y khoa (cấu trúc *nxml*)
  - 30 câu truy vấn (theo tỷ lệ câu hỏi lâm sàng 1:1:1)
  - Danh sách gồm 1,000 tài liệu phù hợp/câu truy vấn.

- **Phương pháp đánh giá:**
  - Precision
  - Recall
  - P@10
  - R-precision

Kết quả thí nghiệm



TN1: BM25L, MeSH, PRF

TN1*: BM25L, MeSH, PRF, GA

TN2: BM25L, MeSH, PRF, QTW

TN3: TN1, TN2, RRF

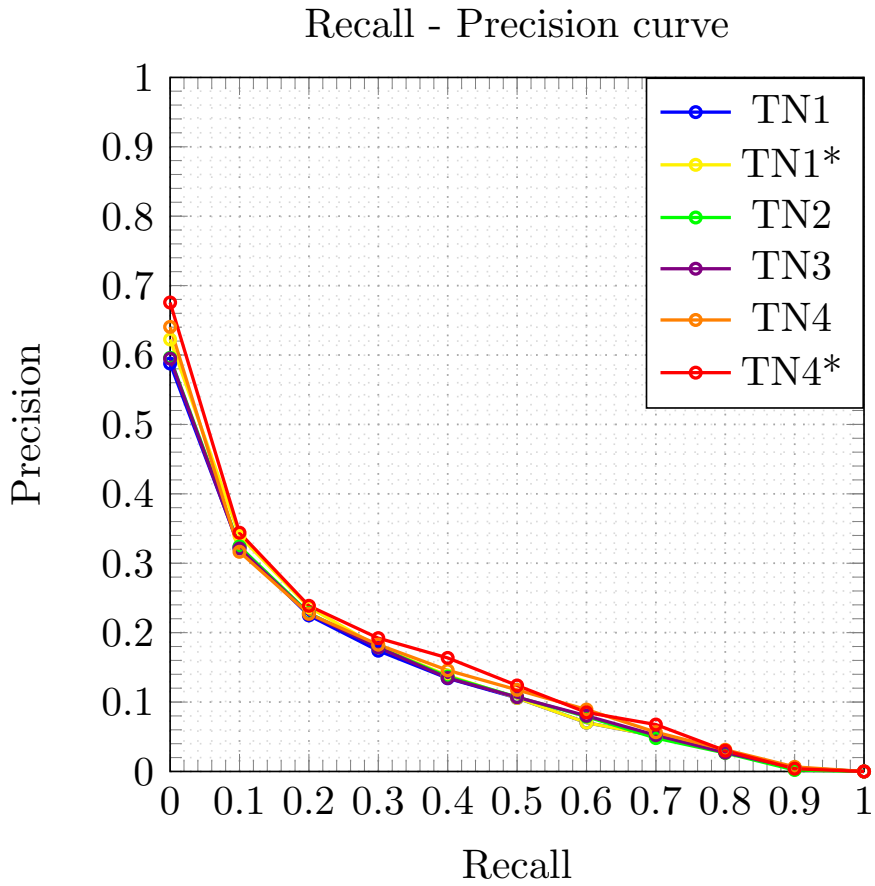
TN4: (BM25L, BM25+, VSM, LM), MeSH, PRF, RRF

TN4*: (BM25L, BM25+, VSM, LM), MeSH, PRF, GA, RRF

(RRF: Reciprocal Rank Fusion)

- Kết hợp nhiều mô hình chấm điểm làm tăng độ chính xác.
- Số lượng tài liệu phù hợp ở những thứ hạng cao tăng lên khi áp dụng giải thuật di truyền.

Kết quả thí nghiệm



TN1: BM25L, MeSH, PRF

TN1*: BM25L, MeSH, PRF, GA

TN2: BM25L, MeSH, PRF, QTW

TN3: TN1, TN2, RRF

TN4: (BM25L, BM25+, VSM, LM), MeSH, PRF, RRF

TN4*: (BM25L, BM25+, VSM, LM), MeSH, PRF, GA, RRF

(RRF: Reciprocal Rank Fusion)

- Kỹ thuật hợp nhất xếp hạng RRF giúp tăng số lượng tài liệu đúng trong tổng số tài liệu trả về.

Nội dung

1 Giới thiệu

2 Phương pháp đề xuất

3 Kết quả thí nghiệm

4 Tổng kết

Tổng kết

- **Kết quả đạt được**
  - Xây dựng thành công hệ thống tìm kiếm trên tập tài liệu văn bản y khoa
  - Khai thác đặc trưng trong y khoa để mở rộng câu truy vấn dựa vào từ điển
- **Hạn chế tồn tại**
  - Chưa khai thác hết cấu trúc tập dữ liệu liệu lớn
  - Các công cụ mở rộng truy vấn còn hạn chế
- **Hướng phát triển**
  - Phát triển hệ thống tìm kiếm tài liệu y khoa trên tập dữ liệu tiếng Việt

Cám ơn hội đồng đã lắng nghe!

Hệ số tương đồng Jaccard

Cho hai tập hợp X, Y, hệ số tương đồng *Jaccard* được tính như sau:

$$J(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}, \qquad 0 \leqslant J(X,Y) \leqslant 1$$
$$= \frac{|X \bigcap Y|}{|X| + |Y| - |X \bigcap Y|}$$

back

Prajakta Mitkal, Prof. (Ms.) Deipali Gore. Improving the Performance of Information Retrieval System using AGA in Distributed Environment. In *International Journal of Innovative Research in Computer and Communication Engineering*, 2016.

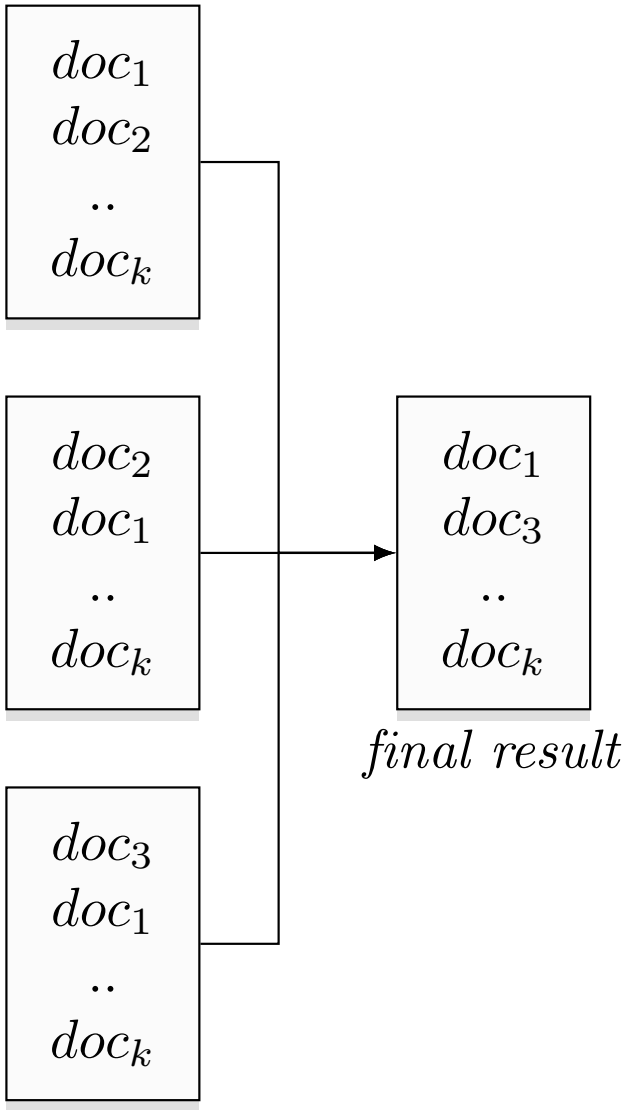
Medical Subject Headings (MeSH)

- **Diagnosis:**
  - B03 - Organisms.Bacteria
  - B04 - Organisms.Viruses
  - C - Diseases
- **Treatment:** D02, D04, D06, D26, D27, E02, E04
- **Test:** E01

back

André Mourão, Flávio Martins and João Magalhães. NovaSearch at TREC 2014 Clinical Decision Support Track. In *Proceedings of the 23rd Text Retrieval Conference (TREC)*, 2014.

Reciprocal Rank Fusion - RRF



- Điểm số mới của từng tài liệu được tính theo công thức sau

$$RRFScore(doc_i) = \frac{1}{\sum_j^n \frac{1}{pos(doc_{ij})}}$$

trong đó:

- n số danh sách tài liệu
- doc_{ij} tài liệu thứ i trong tập kết quả thứ j

back

G. V. Cormack, C. L. A. Clarke, Stefan Buttcher. Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods. In *Special Inspector General for Iraq Reconstruction (SIGIR)*, Boston, Massachusetts, USA, 2009.