Input:

+ nhu cầu tìm kiếm (query)

+ kho dữ liệu

Ouput: thông tin cần tìm => danh sách hoặc tập hợp các đối tượng tìm kiếm phù hợp với nhu cầu tìm kiếm

Các hệ thống tìm kiếm phải định nghĩa khái niệm thế nào là relevance

* Boolean Retrieval => relevance là có chứa tù khóa và thõa mãn các quan hệ luận lý
* Vector Space => relevance là similarity giống query và phù hợp nhu cầu tìm kiếm

Đo được

* Rút trích thông tin đối tượng
* So sánh vector (đo khoảng cách vector)

Tham số ảnh hưởng (factor) đến vector space model

+ số lượng từ trong từ điển (dictional)

+ Phân phối (distribute) từ trong từ điển

(chọn từ không cover được document)

Hạn chế của Boolean retrieval (không thể ranking)

Vấn đề của Boolean retrieval là scalibitity

Để ranking Boolean retrieval => sử dụng chỉ số phụ (chỉ số phụ tùy mỗi hệ thống phù hợp với từng hệ thống)

Các bước vector space

Step 0: vocabulary

+ size

+ distribution

Step 1: vector => chuyển document thành vector

Step 2: Distance (tính khoáng cách)

**Trường hợp độ đo euclic tốt hơn độ đo cosine (thường thì độ đo cosine tốt hơn)**

Gap

Intension gap : sự liên quan giữa information need và query

Sematic gap

Sensetive gap

Precision (độ chính xác) tỉ lệ % kết quả trả về chính xác

Recall: (độ bao phủ) -> lấy kết quả đúng trong kho dữ liệu

Evaluation

Bài tập tính toán độ đo nào tốt hơn

Các độ đo

Độ đo AUC

AUC càng lớn càng tốt

Tổng của 11 điểm (mức) recall

Mối liện hệ dựa vào AUC -> dựa trên đường cong Pre-Recall

Dấu nhân vì agreement (sự đồng thuận)

Độ đo Precision at K

Độ đo R – Precision

Độ đo F – Measure

Độ đo MAP (Mean avegage Precision) (sensitive to rank)

Trung bình của precision ở nhiều mức recall của 1 câu query

Độ đo R – Precision (Sensetive to rank)

Độ do none-binary trustman

+ CG not sensitive to rank

+ DCG ( sensitive to rank)

+ NDCG

Các yếu tố ảnh hưởng đến hệ thống tìm kiếm

+ Subjective ( chủ quan)

+ Situational ( tình huống cụ thể, nhu cầu hiện tại)

+ Cognitive ( nhận thức)

+ Dynamic (thay đổi theo thời gian)

Trước khi đánh giá hệ thống tìm kiếm chúng ta cần làm gì?

+ Tiêu chí đánh giá (muốn đo năng lực gì của hệ thống) xác định độ đo phù hợp

+ Dữ liệu chuẩn (có đáp án)

Dữ liệu có chuẩn bao gồm:

+ Query

+ Collection

+ label

* Không có đáp án thì không đánh giá được
* Trong các độ đo, độ đo nào có Precision và recall (F-Measure, R-Presion, MAP, AUC)

Releven Feedback

Sử dụng feedback => revice câu query

Sử dụng kết quả để improve hệ thống không thay đổi hệ thống

+ reweighting

+ query expansion (them từ)

Global – nhập nhằng từ cao

Local –

Sử dụng top kết quả trả về => local

Local sẽ tốt hơn global nhưng local tốn chi phí cao hơn global

Global chỉ thống kê 1 lần ( pha offline)

Local thống kê nhiều lần (mỗi lần tìm kiếm)

Image Search

Sensetive gap

Sematic gap (từ bức ảnh rút trích được ý nghĩa cáo)

Extension gap

Nhiều bài toán tiếp cận theo hướng relevance là similar

Large scale (Localize Search Hashing) băm không gian đặc trưng chuyển về bit

Relevance Feedback => revice query

Revice query

+ re - weighting

+ expansion