Search Model

- Boolean Model

Contents

Definitions

- is one of classical Information Retrieval model

- is based on Boolean Logic (AND, OR, NOT)

- is based on classical Sets Theory in that both documents to be searched and the user's query are conceived as sets of terms

- is based on whether or not the documents contain the query terms.

Boolean Query

- 매우 직관적

- 불리언 모델을 사용하여 검색질의어 처리를 하면 포함(TRUE), 포함하지 않는다(FALSE)로 단순화
  → Boolean Query
  (자연어에 좀 더 가까운 사용자 질의어 처리를 위해서 Vector Space Model 또는 Extended Boolean Model을 사용)

- 예)
  1. "A"와 "B"를 모두 포함하는 문서의 검색 → "A" AND "B"
  2. "A"를 포함하지만 "B"를 포함하지 않는 문서의 검색 → "A" AND (NOT "B")

  3. "정보 검색 시스템" → "정보" OR "검색" OR "시스템"
     "정보" AND "검색" AND "시스템"

- Query에 대해서도 형태소 분석을 통해서 term 추출 후 어떻게 Boolean Logic을 만들 것인가가 관건
  1. term들간의 OR Logic → Recall ↑
  2. term들간의 AND Logic → Precision ↑

Extension of Query

- 문서 Collection
  문서 A : 정보 검색 시스템이란 …
  문서 B : 포털에서의 정보 검색은 주로 …
  문서 C : 정보검색시스템은 구글에서 …
  문서 D : 정보 검색은 초창기 문헌 정보 …

Boolean Query	
문제점	직관적이고 단순하나 사용자의 의도를 충분히 파악하기 힘듬
예	User Query : "정보검색시스템" →변환 Query : 정보검색시스템(동일) True인 경우에만 찾는다면 "C" 문서만 검색

Extension of Query	
해결책	질의 명백화와 Level, Priority를 할당 (형태소 분석기와 밀접한 관계)
예	User Query : "정보검색시스템" → 1. ((정보 AND 검색) OR 정보검색) AND ((검색 AND 시스템) OR 검색시스템) → 2. 정보검색시스템 OR 정보검색 OR 정보 OR 검색 OR 검색시스템 OR 검색 OR 시스템

- Extension of Query : 질의 명백화
  - "정보검색시스템"이라는 질의어는 정보, 검색, 시스템이라는 term이 전부 존재하는 문서만 검색하라는 의미로만 해석
  → 변환 Query : ((정보 AND 검색) OR 정보검색) AND ((검색 AND 시스템) OR 검색시스템)
  → 해석 : ("정보"와 "검색") 혹은 "정보검색"를 포함하고 있는 문서 중에
  ("검색"과 "시스템") 혹은 "검색시스템"을 포함하고 있는 문서를 검색
  → 장점 : - 사용자의 의도를 확대 해석하지 않는 범주에서 명확하게 검색
  - Precision ↑
  → 단점 : - "정보검색", "검색시스템"만 존재하는 의미있는 문서는 제외
  - Recall ↓

Extension of Query

- Extension of Query : Level, Priority 할당

"정보검색시스템"이라는 질의어는 정보, 검색, 시스템이라는 term이 하나라도 존재하는 문서를 검색하되
각 term들에 대한 Level과 Priority를 달리 줌으로써 사용자의 의도를 최대한 만족시킨다는 의미

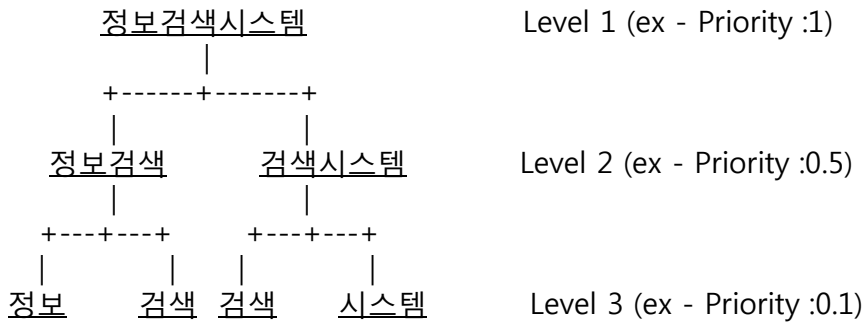
→ 변환 Query : (정보검색시스템:L1) OR (정보검색:L2) OR (정보:L3) OR
(검색:L3) OR (검색시스템:L3) OR (검색:L3) OR (시스템:L3)

→ 해석 : "정보검색시스템", "정보검색", "정보", "검색", "검색시스템", "검색", "시스템"이 하나라도 있는 문서를 검색하되, Level에 따른 Priority를 달리 하여 높은 Priority일수록 상위에 랭킹

→ 장점 : - 사용자의 의도를 확대는 하되 상위 검색 결과는 명확
- Precision ↔

→ 단점 : - 연산 속도가 느려진다.
- Recall ↑

➔ 보완 : 검색 대상 문서가 너무 많아지며, 연산 속도가 현저히 떨어진다.
Level1에서의 결과 획득 후 Recall율을 체크 후 Level2의 문서 추가, Level3의 문서 추가...



※ 검색 서비스와 컬렉션의 성격에 따라서 Extension of Query의 방법 중 하나를 선택할 것

Limitations

- 문서의 우선순위나 사용자 질의에 대한 가중치 등을 부여할 수 없다.
- 사용자의 의도와 다른 결과물을 보여줄 확률이 높다.
- 직관적이고 단순하지만, 사용자의 의도를 충분히 판단하기 어렵다.
- 유사도를 계산할 수 없다.