Search Model

- Probabilistic Model

Contents

1. Definition
2. Similarity
3. Initial Probability
4. Features

Definitions

- 주어진 Query로 각 Document가 해당 Query에 적합할 확률을 Bayes' Rule을 활용하여 계산

- 독립 가정을 전제로 Bayes' Rule을 이용하여, 비연관문서 집합에서 질의가 포함될 확률에 대한 연관 집합에 포함될 확률을 계산하여 문서를 찾는 모델링

- 특정 질문에 대한 각 문서의 관련 확률과 관련이 없을 확률을 산출

- "관련 확률 > 관련 없을 확률"인 문서를 검색하는 방법

- 각 문서 X가 $X = (x_1, x_2, x_3, \ldots, x_n)$ 형태의 벡터로 표현되며, x_i 는 0또는 1의 값을 가지게 된다.

Similarity

-

Bayes' rule

P(R), P(R)은 모든 문헌에 동일

$w_{ij} \in \{0,1\}, w_{iq} \in \{0,1\}$: index term weight variables are all binary

$$sim(d_j, q) = \frac{P(R \mid \vec{d}_j)}{P(\overline{R} \mid \vec{d}_j)} = \frac{P(\vec{d}_j \mid R) \times P(R)}{P(\vec{d}_j \mid \overline{R}) \times P(\overline{R})} \sim \frac{P(\vec{d}_j \mid R)}{P(\vec{d}_j \mid \overline{R})}$$

유사도:
질의 q에 대하여, 연관문서 집합에서 문서 j가 나올 확률

R : Set of documents known to be relevant

$\overline{R}$: Set of documents known to be non-relevant

$P(R \mid \vec{d}_j)$: Probability that the document d_j is relevant to the query q

$$sim(d_j, q) \sim \frac{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid R)\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(\overline{k}_i \mid R)\right)}{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid \overline{R})\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(\overline{k}_i \mid \overline{R})\right)}$$

색인어 독립성 가정

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{iq} \times w_{ij} \times \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid R)}{P(k_i \mid \overline{R})} \right)$$

연관문서 집합에서 가중치가 1인 키워드가 나올 확률

연관문서 집합에서 가중치가 0인 키워드가 나오지 않을 확률

Log를 취하고, 상수 무시

$$P(k_i \mid R) + P(\overline{k}_i \mid R) = 1$$

Initial Probability

$$P(k_i \mid R) = 0.5$$

$$P(k_i \mid \overline{R}) = \frac{n_i}{N} \qquad n_i : \text{number of documents which contain the index term } k_i$$

Improving Probability

$$P(k_i \mid R) = \frac{V_i}{V} = \frac{V_i + 0.5}{V + 1} = \frac{V_i + \frac{n_i}{N}}{V + 1}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i}{N - V} = \frac{n_i - V_i + 0.5}{N - V + 1} = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$

너무 작은 V와 V_i의 경우, 조정 요소를 더함

V : subset of documents initially retrieved

V_i : subset of V which contain the index term k_i

자동화된 적합성 피드백 :
이용자의 적합성 피드백에 의한 가중치 적용이 아니라 자동으로 학습을 통한 적합성 피드백을 수행하는 경우이다.

Features

- Strength
    - 연관 확률에 따라 문헌 순위화
- Weak
    - 초기 문헌을 연관 / 비연관으로 분리 가정
    - 이진 가중치
        - 색인어의 문헌 내 빈도수 비고려
    - 색인어 독립성 가정