

3장

3. Probabilistic Model

* Information Needs, Queries

- Information needs는 query에 모두 담겨있다.
- 단순히 1번으로 끝나는 게 아님. 계속 해야함.

* Retrieval Models

- query, document \rightarrow matching.
- q, d 에 대한 relevance : $P(R=1 | q, d)$
- Relevance 정의
 - 정의가 어렵다.
 - Topical relevance : topic에 따름.
User relevance : user에 따름
 - 모호하기 때문에, 관련성을 uncertainty로 표현하자.
probabilistic

* Document Ranking

- 주어진 query에 대해 각 문서 ranking 기준
 $f(q, d) = P(R=1 | q, d)$ → q는 편의상 고정
기준

- 각 문서에 대해 f 를 계산, 이로 ranking.
- relevant 문서 = $\{d | f(q, d) > 1 - f(q, d)\}$

- 계산

$$P(R=1 | q, d) = \frac{\text{Count}(R \geq 1, d, q)}{\text{Count}(q, d)}$$

- heuristic한 data 필요.

* Relevance feedback

- query 결과 \rightarrow update 반복. User 개입된 retrieval process.
 \nwarrow
feedback

* Probability Ranking Principle (PRP)

- 수학적 Solidar model.

- return top K documents in $p(R=1 | d, q)$

$$p(R=1 | d, q) = \frac{p(d | R=1, q) p(R=1 | q)}{p(d | q)}$$

q는 고정

- $p(R=1 | q)$ 는 prior probability of relevant document

$$p(R=1 | q) = \frac{\text{Count}(R=1, q)}{\text{Count}(q)}$$

관련문서수
전체

사실 $p(R=0 | q)$ 가 훨씬 더 큼.

- $p(d | R=1, q)$ 는 likelihood.

$$p(d | R=1, q) = \frac{\text{Count}(d, R=1, q)}{\text{Count}(R=1, q)}$$

주의) $p(d | R=0, q) \neq 1 - p(d | R=1, q)$

- 관련없는 문서

- $p(R=1 | d, q) > p(R=0 | d, q)$

- $p(R=1 | d, q) > 0.5$

* Binary Independence Model (BIM)

- 문서 query는 binary vector로 표현

- $d = (x_1, \dots, x_m)$. $x_i = 1$ if x_i is in d .

- 문서 query 내 각 term의 발생 확률은 서로 independent함.

$$p(d | R) = p(x_1 | R) p(x_2 | R) \cdots p(x_m | R) = \prod_{x_i \in d} p(x_i | R)$$

* Odds

$$\begin{aligned} \text{odds} &= \frac{p(y=1|x)}{1-p(y=1|x)} = \frac{p}{1-p}, \quad p = \frac{\text{odds}}{1+\text{odds}} \\ &= \frac{p(1)}{p(0)} \end{aligned}$$

* Ranking by Odds

- 주어진 q 에 대해서, 각 d 에 대한 $p(R=1|q, d)$ 계산 해야 함.

$$\begin{aligned} - O(R|d, q) &= \frac{p(R=1|d, q)}{p(R=0|d, q)} = \frac{\frac{p(d|R=1, q)}{p(d|q)} \cdot \frac{p(R=1|q)}{p(R=0|q)}}{\frac{p(d|R=0, q)}{p(d|q)} \cdot \frac{p(R=0|q)}{p(R=1|q)}} \\ &= \frac{p(R=1|q)}{p(R=0|q)} \cdot \frac{p(d|R=1, q)}{p(d|R=0, q)} \end{aligned}$$

q 가 일정하면
constant.

- Using independence 가정

$$\begin{aligned} \cdot \frac{p(d|R=1, q)}{p(d|R=0, q)} &= \prod_{x_i \in d} \frac{p(x_i|R=1, q)}{p(x_i|R=0, q)} \\ \cdot O(R|d, q) &= O(R|q) \cdot \prod_{x_i \in d} \frac{p(x_i|R=1, q)}{p(x_i|R=0, q)} \end{aligned}$$

- $X_i = 0$ or 1

$$O(R|d, q) = O(R|q) \cdot \prod_{x_i=1} \frac{p(x_i=1|R=1, q)}{p(x_i=1|R=0, q)} \prod_{x_i=0} \frac{p(x_i=0|R=1, q)}{p(x_i=0|R=0, q)}$$

$$- p_i = p(x_i=1|R=1, q), \quad u_i = p(x_i=1|R=0, q)$$

관련 있는 문서 중 단어 x_i 가 나온 확률
(문서의 개수)

관련 없는 문서 중 단어 x_i 가 나온 확률

$$O(R|d, q) = O(R|q) \cdot \prod_{x_i=1} \frac{p_i}{u_i} \prod_{x_i=0} \frac{(1-p_i)}{(1-u_i)}$$

• p_i 가 높을수록 더 중요한 term임.

u_i 가 높을수록 덜 중요한 term임.

$\rightarrow p_i, u_i$ 둘다 높을수록 있음.

* Retrieval Status Value (RSV)

- query에 존재하지 않는 단어에 대한 가중치.

• If $q_i = 0$, then $p_i = u_i$

• query 내에 있는 단어들만 고려하자.

$$- O(R|d, q) = O(R|q) \cdot \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{u_i} \prod_{\substack{x_i=0 \\ q_i=1}} \frac{p_i}{u_i} \prod_{\substack{x_i=0 \\ q_i=0 \\ p_i=1}} \frac{1-p_i}{1-u_i} \prod_{\substack{x_i=0 \\ q_i=0 \\ p_i=0}} \frac{1-p_i}{1-u_i}$$

$$= O(R|q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{u_i} \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-u_i}$$

변형

$$- O(R|d, q) = O(R|q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{u_i} \prod_{\substack{x_i=1 \\ q_i=1}} \left(\frac{1-u_i}{1-p_i} \cdot \frac{1-p_i}{1-u_i} \right) \cdot \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-u_i}$$

$$= O(R|q) \underbrace{\prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i(1-u_i)}{u_i(1-p_i)}}_{\text{Only term for doc.}} \cdot \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-u_i}$$

문서 차트와 관련없음.
query 차트의 값.

* 정리

$$- O(R|d, q) = O(R|q) \cdot \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i(1-u_i)}{u_i(1-p_i)} \cdot \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-u_i}$$

Constant → only 중요.

$$- RSV = \log \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i(1-u_i)}{u_i(1-p_i)} = \sum_{q_i=x_i=1} \log \frac{p_i(1-u_i)}{u_i(1-p_i)} = C_i$$

- 문제 : p_i (관련있는 문서 중 x_i 가 발생할 확률) 계산도 어렵다.

- 변형

$$RSV = \sum_{q_i=x_i=1} C_i, \quad C_i = \log \frac{p_i(1-u_i)}{u_i(1-p_i)} = \log \frac{p_i}{1-p_i} + \log \frac{1-u_i}{u_i}$$

TF (중요성) IDF (Penalty)

* 문제 풀이

- f_i 에 있는 단어 x_i 에 대해

$$\textcircled{1} \quad p_i = p(x_i=1 | R=1, f_i) \text{ 구하고, } \log \frac{p_i}{1-p_i} \text{ 구하기.}$$

$$\textcircled{2} \quad u_i = p(x_i=1 | R=0, f_i) \text{ 구하고, } \log \frac{1-u_i}{u_i} \text{ 구하기.}$$

u_i 에 각 단어의 중요도

$$\textcircled{3} \quad C_i = \log \frac{p_i}{1-p_i} + \log \frac{1-u_i}{u_i} \text{ 구해서, 각 단어에 대해 } C_i \text{ 구하기.}$$

- $p(d | R=1, f_i)$ 구하기 ranking

$C_i > 0$: 중요
 $C_i < 0$: 순위

\textcircled{1} d 에 각 단어에 대해, C_i 가 포함되면 C_i 를 모두 더하기.

* 실2) p_i, u_i 계산법 (Heuristic)

	$R=1$	$R=0$	total	
$x_i=1$	s	$df_i - s$	df_i	
$x_i=0$	$S-s$	$(N-df_i) - (S-s)$	$N-df_i$	$p_i = \frac{s}{S}, \quad u_i = \frac{df_i - s}{N-S}$
total	S	$N-S$	N	$s \ll N$ $S \ll N$

• $\log \frac{1-u_i}{u_i} = \log \frac{(N-df_i) - (S-s)}{df_i - s} \approx \log \frac{N-df_i}{df_i} \approx \log \frac{N}{df_i} : IDF$,

2) p_i

- Heuristic 도 어렵다.
- Constant 가능성도 있음. $p_i = 0.5$. 또는 $p_i = \frac{df_i}{N}$.

* Iterative Approach

- User의 의해 relevance 나뉨.

- $UR = \{d \in R : R_{d,f_i} = 1\}$

$VNR = \{d \in R : R_{d,f_i} = 0\}$

) OI 해석기.

$$\cdot \quad p_i = \frac{|UR_i|}{|URI|} \xrightarrow{\text{smoothing}} p_i = \frac{|UR_i| + \frac{1}{2}}{|URI| + 1} \quad \text{or}$$

$$p_i^{(k+1)} = \frac{|UR_i| + k p_i^{(k)}}{|URI| + k}, \quad k=5$$