Chap 9
Search Engine Evaluation

데이터 마이닝 연구실

G201849026
신승엽

목차

1. What to measure?
2. Evaluation of Set Retrieval
3. Evaluation of a Ranked List
4. Evaluation with Multi-level Judgements

무엇을 평가?

1.Effectiveness or accuracy : 검색 결과가 얼마나 정확한지

2.Efficiency : 검색이 얼마나 빠르게 되는지

3. Usability : 검색이 얼마나 유용한지

=> 이 중 주요하게 다룰 것은 1.Effectiveness or accuracy

System A vs System B

Q1과 관련 있는 문서의 개수가 총 10개라고 가정.

System A

-> 3개 중 2개가 맞음. 10개 중 2개를 찾음

System B

-> 5개 중 3개가 맞음. 10개중 3개를 찾음

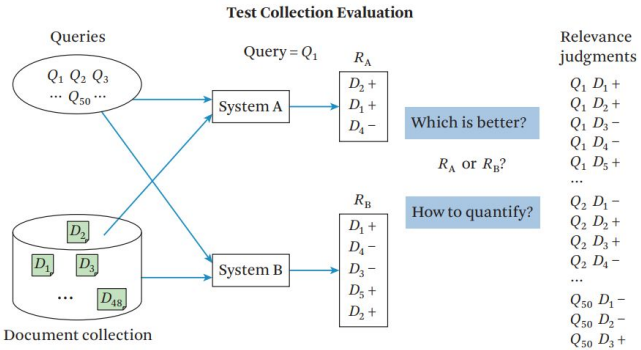


Figure 9.1 Illustration of Cranfield evaluation methodology.

Evaluation of Set Retrieval

Set Retrieval

검색 결과의 순서는 고려하지않고 검색 결과 관련 문서가 있는지만 고려.

Precision and Recall

		Action: Retrieved	Action: Not retrieved
Doc	Relevant	*a*	*b*
	Not relevant	*c*	*d*

$$\text{Precision} = \frac{a}{a+c}$$

Ideal results: precision = recall = 1.0

$$\text{Recall} = \frac{a}{a+b}$$

In reality, high recall tends to be associated with low precision

System A

Precision : 2/3 Recall : 2/10

System B

Precision : 3/5 Recall : 3/10

A'Precison > B'Precision

A'Recall < B'Recall

Precision VS Recall

Precicison 과 Recall 중 더 중요한 것은 상황에 따라 다름.

일반적으로 정보 검색에서 사용자는 Ton N개의 결과만 보기 때문에 Top N에 대한 Precision이 중요함.

하지만 특정 상황의 경우 정보 검색 결과가 관련 문서를 최대한 많이 보여주는 것이 좋은 경우가 있음.

F measure

Precision과 Recall의 Tradeoff 관계를 고려해 둘을 한번에 계산하는 방법.

$$F_\beta = \frac{1}{\frac{\beta^2}{\beta^2+1}\frac{1}{R}\frac{1}{\beta^2+1}\frac{1}{P}} = \frac{(\beta^2+1)P * R}{\beta^2 P + R}$$

$$F_1 = \frac{2PR}{P+R}$$

where P = precision, R = recall, β = parameter (often set to 1)

F-1 measure

β가 1인 경우 Precision과 Recall의 조화평균

Precision과 Recall의 산술평균을 사용하지 않는이유

=> 큰 값에 영향을 많이 받기 때문.

두가지 값을 동시에 고려하기 위한 지표인데 하나의 값에 영향을 크게 받는 지표는 좋지 않음

Evaluation of a Ranked List

Ranked List

검색 결과 나오는 문서의 순서도 고려한다.

Precision–Recall (PR) Curve

사용자가 검색 결과를 얼마나 조회하냐에 따라

Precision과 Recall 값이 달라진다.

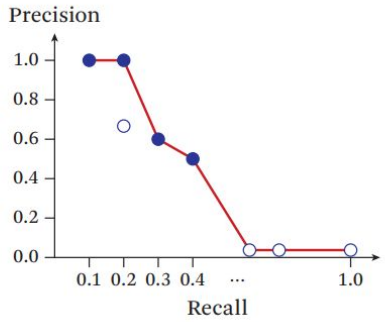
조회를 많이할수록 Recall은 올라가고 Precision은

떨어진다.

Evaluating Ranking: Precision–Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D_1 +	1/1	1/10
D_2 +	2/2	2/10
D_3 –	2/3	2/10
D_4 –		
D_5 +	3/5	3/10
D_6 –		
D_7 –		
D_8 +	4/8	4/10
D_9 –		
D_{10} –	?	10/10



System A VS System B

그림 1

System A > System B

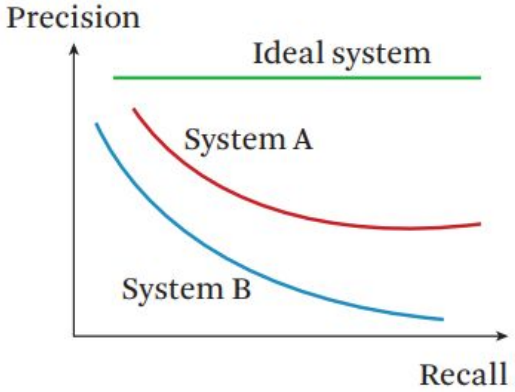
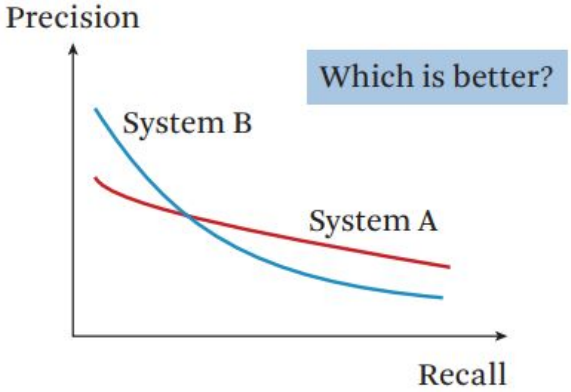


그림 2

상황에 따라 다름



Average Precision

Mathematically, we can define **average precision** on a ranked list L where $|L| = n$ as

$$\text{avp}(L) = \frac{1}{|Rel|} \sum_{i=1}^{n} p(i), \tag{9.1}$$

|Rel| : 관련 있는 문서의 총 개수

p(i) 는 문서의 랭킹/검색 결과의 순서 ex) 세번째로 랭크된 문서가 검색 결과 7번째 나온경우 p(7) = 3/7

Average Precision

i	Rel	$p(i)$
1	+	$\frac{1}{1} = 1.0$
2	+	$\frac{2}{2} = 1.0$
3	−	0.0
4	−	0.0
5	+	$\frac{3}{5} = 0.6$
6	−	0.0
7	−	0.0
8	+	$\frac{4}{8} = 0.5$
⋮	−	0.0
sum		3.1
avp		$\frac{3.1}{10} = 0.31$

Average Precision은 검색 결과 나오는 문서의 순서에 따라 다른 값을 가지므로 Ranked List를 평가 가능

Mean Average Precision

정보 검색 시스템을 평가할 때 하나의 쿼리에 대해서만 평가하는 것은 올바른 평가 방법이 아니다.

특정 쿼리에 대해 유난히 좋은 결과를 가지는 시스템이 있을 수 있기 때문에 여러 쿼리로 평가를 해야함.

앞서 사용한 Average Precision은 하나의 쿼리에 대한 평가이므로 Mean Average Precision이 필요.

m : 쿼리 개수

L : 검색 결과 Raked Lists

$$MAP(\mathcal{L}) = \frac{1}{m} \sum_{i=1}^{m} \text{avp}(\mathcal{L}_i).$$

Geometric Mean Average Precision,

앞에서 F1 measure 구하는 식이 산술평균이 아니었던 이유와 마찬가지로 MAP는 큰 값의 영향을 크게 받는다.

즉, 쉬운 검색(높은 AP값)에 영향을 크게 받는 값이다.

쉬운 검색을 잘하는 지 보고 싶은게 아니라 어려운 검색을 잘하는지 보기 위해서 gMAP를 사용

$$\mathrm{gMAP}(\mathcal{L}) = \left(\prod_{i=1}^{m} \mathrm{avp}(\mathcal{L}_i) \right)^{\frac{1}{m}},$$

or in log space as

$$\mathrm{gMAP}(\mathcal{L}) = \exp\left\{ \frac{1}{m} \sum_{i=1}^{m} \ln \mathrm{avp}(\mathcal{L}_i) \right\}.$$

Reciprocal Rank

관련 있는 문서가 단 하나 존재하는 경우(찾고자 하는 사이트가 하나)

검색 내에서 관련 있는 문서의 위치(Rank)의 역수 : 1/r

MRR : 여러 쿼리의 Reciprocal Rank의 평균

단순히 Rank가 아니라 Rank의 역수를 취하는 이유

=> MRR은 산술평균으로 앞서 말했듯이 큰 값에 영향을 받음

r이 크다는 것은 정보 검색 결과가 좋지 않음을 뜻함.

역수를 취해줌으로써 검색 결과가 좋을수록 큰 값을 가짐.

Evaluation with Multi-level Judgements

Multi-level

쿼리와 문서간의 관계 유무에서 나아가 관계 정도를 고려

Gain,CG,DCG,NDCG

	Gain	Cumulative gain	Discounted cumulative gain
D_1	3	3	3
D_2	2	3 + 2	3 + 2/log **2**
D_3	1	3 + 2 + 1	3 + 2/log **2** + 1/log **3**
D_4	1	3 + 2 + 1 + 1	⋯
D_5	3	⋯	
D_6	1		
D_7	1		
D_8	2		
D_9	1		
D_{10}	1		

$$\text{Normalized DCG} = \frac{\text{DCG@10}}{\text{IdealDCG@10}}$$

$$\text{DCG@10} = 3 + 2/\log \mathbf{2} + 1/\log \mathbf{3} + \cdots + 1/\log \mathbf{10}$$

$$\text{IdealDCG@10} = \mathbf{3} + \mathbf{3}/\log 2 + \mathbf{3}/\log 3 + \cdots + \mathbf{3}/\log 9 + \mathbf{2}/\log 10$$

Relevance level: $r = 1$ (non-relevant), 2 (marginally relevant), 3 (very relevant)

Gain,CG,DCG,NDCG

$$CG(L) = \sum_{i=1}^{n} r_i \quad DCG(L) = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i} \quad NDCG(L) = \frac{DCG(L)}{IDCG}$$

CG: 검색 결과 얻는 Gain값을 단순히 모두 더 한것

DCG : 문서의 Gain 값을 문서가 검색 결과의 위치에 따라 조정한 것. 검색 결과에 늦게 보여질수록 얻게 되는 Gain 값이 작아짐.

NDCG : 서로 다른 쿼리에 대한 비교를 위해 0과 1사이의 값을 갖도록 정규화 한 것. 특정 쿼리에 대하여 DCG 값이 가장 커지는 경우를 IDCG라 정하고 이를 통해 DCG 값을 나눈 값

결론

Recall, Precision, F measure, Average Precision, MAP, gMAP, Reciprocal Rank, MRR, CG, DCG, NDCG

위의 개념들이 나온 배경과 정보 검색 평가에 어떻게 쓰이는 지를 알아보았다.