IR Basics: Evaluation

정보검색 성능 평가 (IR Evaluation)

- Effectiveness
  - 사용자 만족도
    - e.g.) 정확률, 재현율, F지표
- Efficiency
  - 시간, 공간 복잡도
    - e.g.) 문서 당 평균 색인 속도, 검색 소요 시간

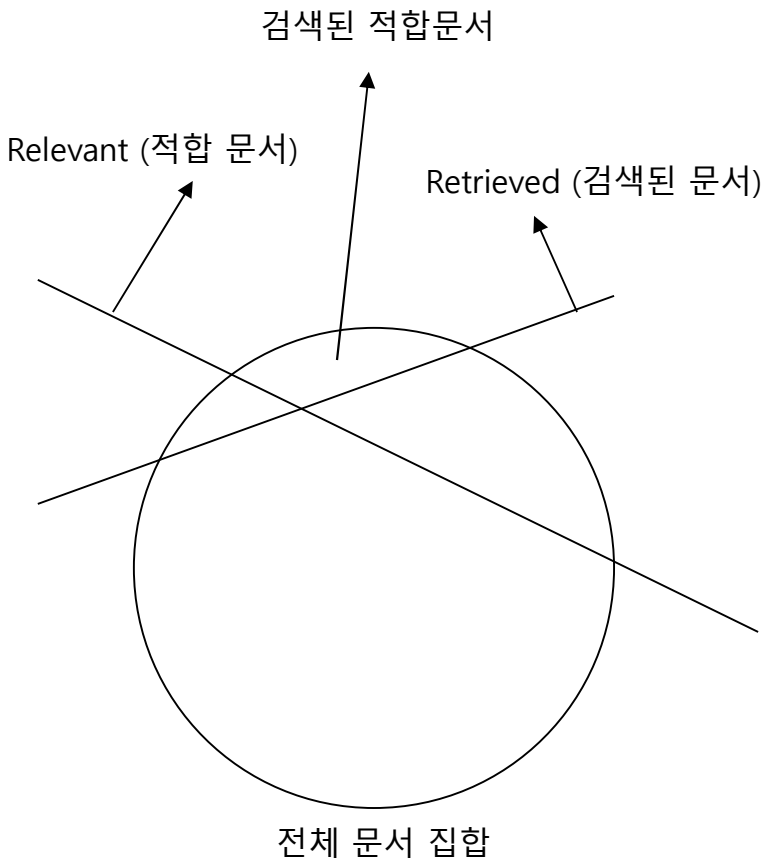
정확률, 재현율

- 정확률 (Precision)

$$Precision = \frac{\text{검색된 문서중 적합문서의 수}}{\text{검색된 문서의 수}}$$

- 재현율 (Recall)

$$Recall = \frac{\text{검색된 문서중 적합문서의 수}}{\text{적합문서의 수}}$$



정확률, 재현율

- Contingency table (발생가능상황에 대한 분할표)

	Relevant	non-Relevant
Retrieved	True Positives (TP)	False Positives (FP)
Not Retrieved	False Negatives (FN)	True Negatives (TN)

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

정확률, 재현율

Contingency table

	적합문서	부적합문서	합
검색 문서	10	90	100
검색되지 않은 문서	40	860	900
합	50	950	1,000

$$Precision = \frac{10}{10+90} = \frac{10}{100} = 0.1$$

$$Recall = \frac{10}{10+40} = \frac{10}{50} = 0.2$$

$$Accuracy = \frac{10+860}{1{,}000} = \frac{870}{1{,}000} = 0.87$$

	적합문서	부적합문서	합
검색 문서	10	90	100
검색되지 않은 문서	40	999,860	999,900
합	50	999,950	1,000,000

$$Precision = \frac{10}{10+90} = \frac{10}{100} = 0.1$$

$$Recall = \frac{10}{10+40} = \frac{10}{50} = 0.2$$

$$Accuracy = \frac{10+999{,}860}{1{,}000{,}000} = \frac{999{,}870}{1{,}000{,}000} = 0.99987$$

정확률, 재현율

- Accuracy
  - IR 문제에 부적절한 지표임
    - e.g.) Accuracy 99.9% IR 시스템 만들 수 있음
- Recall, Precision
  - 둘 다 고려되어야 함
    - e.g.) Recall만 고려할 경우 Recall 100% IR 시스템 만들 수 있음
  - 상황에 따라 어느 한 쪽이 더 중요할 수는 있음
    - e.g.) 웹 검색의 경우 첫 페이지 검색 결과에 적합 문서가 많아야 함
      - Recall보다 Precision이 더 중요
    - e.g.) 전문 분야 검색의 경우 가능한 모든 적합 문서를 찾을 필요 있음
      - Precision보다 Recall이 더 중요

F 지표 – 정확률, 재현율의 결합

- F 지표 (F measure)

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha},\ \alpha \in [0,1],\ \beta^2 \in [0,\infty]$$

$\beta > 1$이면 재현율을 더 강조함

$\beta < 1$이면 정확률을 더 강조함

- F1 지표 (F1 measure)
  - 정확률, 재현율을 같은 중요도로 결합

$$F_{\beta=1} = F_{\alpha=1/2} = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{1}{\frac{1}{2}\frac{1}{P} + (1-\frac{1}{2})\frac{1}{R}} = \frac{2PR}{P+R}$$

Precision, Recall, F1

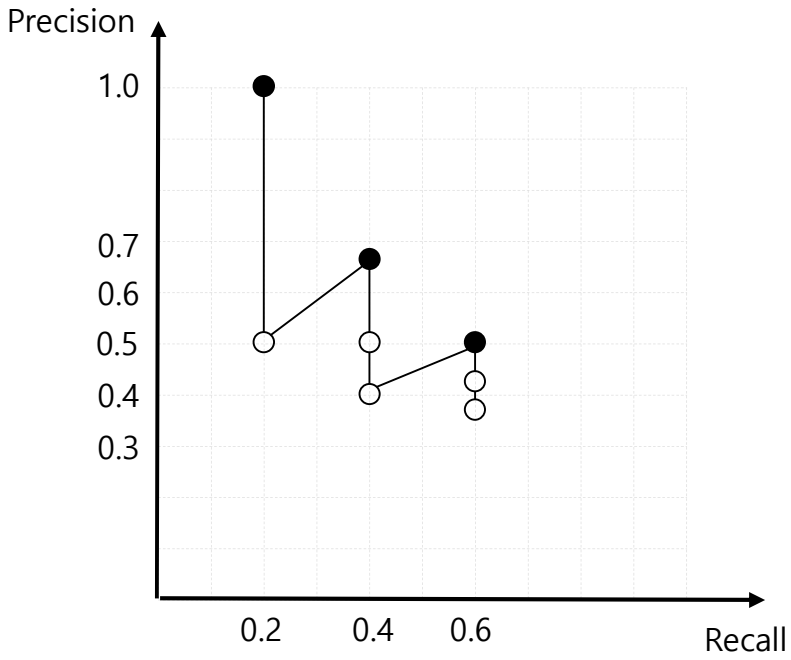
- Precision, Recall, F1
  - 검색 문서 집합 단위의 성능 평가 지표 (set-based measures)
  - 검색 문서들은 적합도 순으로 정렬되어 있지 않다고 가정 (unordered sets of documents)
  - 순위화된 형태(ranked retrieval)로 검색 문서가 제시되는 경우를 고려한 성능 평가 지표가 필요함
    - Precision-recall curve
      - 11-point interpolated average precision
    - MAP (mean average precision)
    - Precision at K
    - R-precision
    - NDCG (normalized discounted cumulative gain)

Precision-Recall Curve

순위	총 적합문서 수 = 5			
	문서번호	적합여부	재현율	정확률
1	555	O	0.2	1.00
2	888		0.2	0.50
3	111	O	0.4	0.67
4	333		0.4	0.50
5	444		0.4	0.40
6	999	O	0.6	0.50
7	222		0.6	0.43
8	666		0.6	0.38

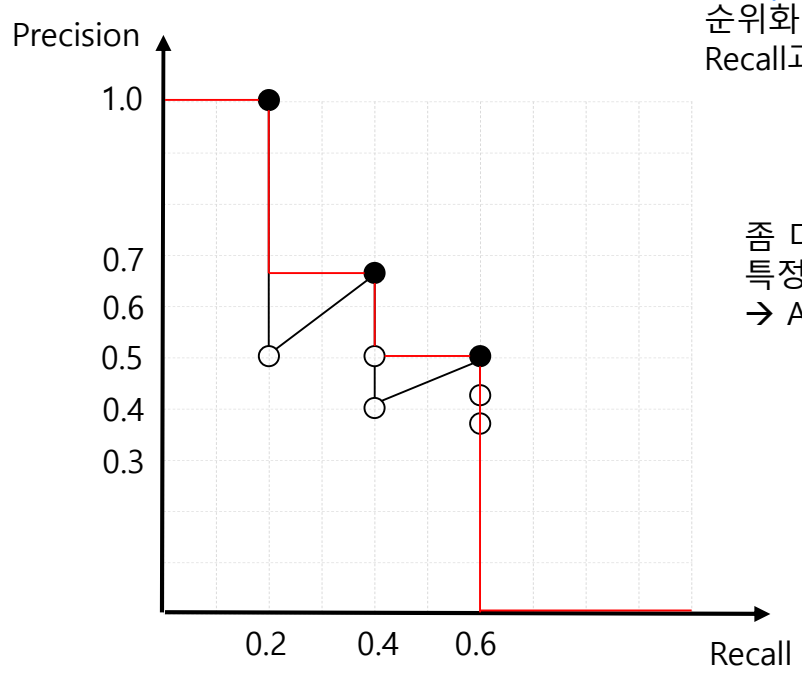
적합문서의 총 수 = 5
재현율(Recall) 0.2 →
순위화된 검색문서 리스트에서 20%의 적합문서가 검색된 지점

Precision-recall curve:
순위화된 검색 결과 리스트의 각 순위 지점(n)에 대해,
Recall과 Precision 값을 계산하여 2차원 좌표에 표시



단순 precision-recall curve의 문제→일반적으로 톱니모양 (sawtooth) 그래프 발생
m번째 적합문서와 m+1번째 적합문서 사이의 비적합문서들의
순위 지점에서 Recall은 m번째 적합문서위치와 같고 Precision은 계속 감소

Interpolated Precision-Recall Curve



Interpolated precision-recall curve:
순위화된 검색 결과 리스트의 각 순위 지점(n)에 대해,
Recall과 interpolated precision 값을 계산하여 2차원 좌표에 표시

좀 더 단순화시켜
특정 재현률 지점들에 대해서만 보간정확률을 계산할 수 있음
→ Averaged 11-point precision/recall graph

Interpolated precision (보간 정확률):
특정 recall level r에서의 보간정확률은 r이상의 모든 recall level에서의 최대 정확률로 정의된다

$$Precision_{interpolated}(r) = \max_{r' \geq r} Precision(r')$$

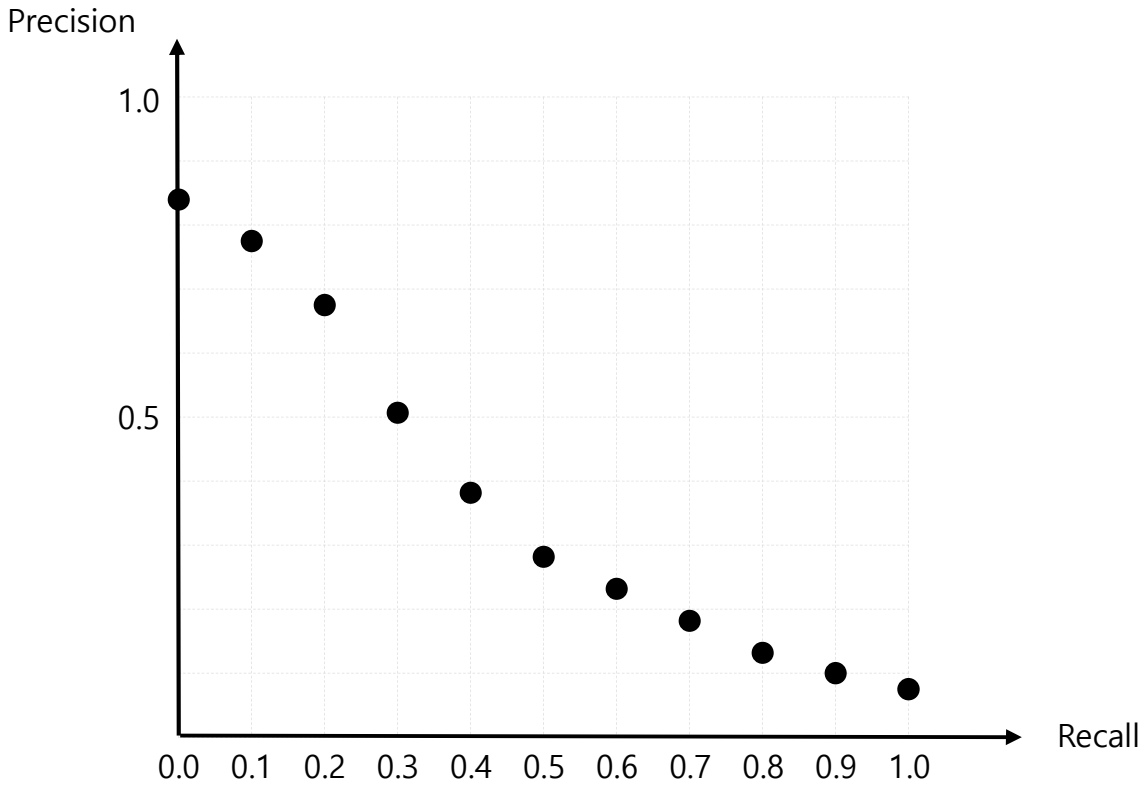
Interpolated Precision-Recall Curve

적합문서의 총 수 = 4
검색된 적합문서 순위: 1, 2, 4, 15
Exact recall points = 0.25, 0.5, 0.75, 1.0

Rank	Recall	Precision
1(R)	1/4=0.25	1
2(R)	2/4=0.5	1
3		
4(R)	3/4=0.75	3/4=0.75
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15(R)	4/4=1.0	4/15=0.27

Exact recall points를 기반으로
11개 표준 재현율 지점의 보간 정확률을 구해야 함
P(0)=P(0.1)=P(0.2)=P(0.3)=P(0.4)=P(0.5)=1.0
P(0.6)=P(0.7)=0.75
P(0.8)=P(0.9)=P(1.0)=0.27

Averaged 11-point Precision/Recall Graph



Averaged 11-point precision-recall curve:
각 query에 대해 11개 각 재현율 수준(0.0, 0.1, 0.2, …, 1.0)에서의 보간정확률들을 구하고,
같은 재현율 수준에 대해 서로 다른 질의의 보간정확률들의 평균을 구하여 2차원 좌표에 표시

평균정확률 (Mean Average Precision, MAP)

순위	총 적합문서 수 = 5			
	문서번호	적합여부	재현율	정확률
1	555	O	0.2	1.00
2	888		0.2	0.50
3	111	O	0.4	0.67
4	333		0.4	0.50
5	444		0.4	0.40
6	999	O	0.6	0.50
7	222		0.6	0.43
8	666		0.6	0.38

적합문서의 총 수 = 5

AP = (1.0+0.67+0.5) / 5

만약, 적합문서의 총 수가 8이라면
AP = (1.0+0.67+0.5) / 8

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

Average Precision

Q → query 집합
$q_j \in Q$ → j번째 query

$\{d_1,\ldots,d_{m_j}\}$ → q_j의 적합문서 집합

m_j → q_j의 총 적합문서 수
R_{jk} → q_j의 검색문서리스트 중 top에서 d_k의 위치까지의 검색문서리스트

Average Precision (AP):
하나의 query에 대해 얻어진 검색문서리스트에서
적합문서가 발견된 순위 지점에서의 Precision들을 총 적합문서에 대해 평균한 것

Mean Average Precision (MAP):
서로 다른 query들에 대해 각 query의 Average Precision들을 평균한 것

상위문서정확률 (Precision at K)

- 상위문서정확률 (*precision at k*)
  - 웹 서퍼에게 모든 적합문서의 재현 지점을 반영한 정확률은 부적절
  - 미리 정해진 상위 10, 20개 검색문서들의 정확률이 보다 유의미

순위	총 적합문서 수 = 5			
	문서번호	적합여부	재현율	정확률
1	555	O	0.2	1.00
2	888		0.2	0.50
3	111	O	0.4	0.67
4	333		0.4	0.50
5	444		0.4	0.40
6	999	O	0.6	0.50
7	222		0.6	0.43
8	666		0.6	0.38

장점: 적합문서집합의 크기를 추정할 필요 없음
단점: 평가지표들 중 가장 불안정
단점: 총 적합문서 수가 성능에 큰 영향을 미침

Pre@5 = 2/5
Pre@10 = 3/10
Pre@20 = 3/20
Pre@30 = 3/30

R-precision

- R-precision
  - 한 query에 대한 총 적합문서 수(R)와 같은 수의 검색문서집합이 얻어진 지점에서의 precision을 계산

순위	총 적합문서 수 = 5			
	문서번호	적합여부	재현율	정확률
1	555	O	0.2	1.00
2	888		0.2	0.50
3	111	O	0.4	0.67
4	333		0.4	0.50
5	444		0.4	0.40
6	999	O	0.6	0.50
7	222		0.6	0.43
8	666		0.6	0.38

총 적합문서 수 = 5

R-precision = 2/5

  - 질의마다 다른 적합문서집합의 크기에 유연하게 반응함
    - e.g.) Perfect 시스템에서 R-precision은 적합문서집합의 크기에 무관하게 항상 1.0, 반면 Pre@K는 K와 적합문서집합의 크기에 의존적
      - 총 8개 적합문서가 있을 경우 Pre@20는 0.4

R-precision

R-precision

순위	총 적합문서 수 = 5			
	문서번호	적합여부	재현율	정확률
1	555	O	0.2	1.00
2	888		0.2	0.50
3	111	O	0.4	0.67
4	333		0.4	0.50
5	444		0.4	0.40
6	999	O	0.6	0.50
7	222		0.6	0.43
8	666		0.6	0.38

총 적합문서 수 = 5

R-precision = 2/5

R-precision값은 Precision이면서 동시에 Recall임
즉, 정확률과 재현율이 같은 지점(=break-even point)이 됨

Pre@K처럼 precision-recall curve에서 하나의 점에 대응되며,
MAP처럼 curve 전체에 대한 effectiveness의 요약은 아님

그러나, R-precison은 경험적으로 MAP과의 상관관계가 크다고 알려져 있음

NDCG (Normalized Discounted Cumulative Gain)

- NDCG
  - 다중 적합도가 부여되는 상황을 위해 고안됨
    - 이진적합도 → 적합문서(relevant, 1), 부적합문서(non-relevant, 0)
    - 다중적합도 → 부적합(0), 부분적 적합(1), 적합(2), 상당부분 적합(3)
  - Pre@K처럼 상위 K개 검색문서에 대해 계산됨

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_k \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log(1+m)}$$

R(j,m) → query j에 대해 검색된 상위 m번째 문서에 대해 부여된 적합도 점수(relevance score)
Z_k → prefect ranking의 NDCG@K 값이 1이 되도록 만드는 정규화인자

NDCG (Normalized Discounted Cumulative Gain)

- CG (누적사용자만족도, Cumulative Gain)
  - 상위 K개 검색문서에 대한 누적사용자만족도
    - 상위 K개 각 문서의 사용자만족도를 단순 누적한 값
  - 단점
    - 상위 K개 리스트 내에서 각 문서의 순위를 반영하지 못함
      - 특정 만족도를 갖는 문서가 1순위이든 K순위이든 CG계산은 동일

Rank	Doc ID	Relevance Score (=Gain)
1	678	3
2	345	2
3	124	3
4	589	0
5	894	1
6	532	2

$CG_6=3+2+3+0+1+2=11$

NDCG (Normalized Discounted Cumulative Gain)

- DCG (차감누적사용자만족도, Discounted Cumulative Gain)
  - 상위 K개 검색문서에 대한 차감누적사용자만족도
    - 상위 K개 각 문서의 순위를 고려한 사용자만족도를 누적한 값
      - 특정 문서의 사용자만족도가 순위가 증가할수록 감소됨
  - 단점
    - 질의별 검색리스트/적합문서집합의 크기 차이를 반영하지 못함
      - K보다 작은 검색리스트/적합문서집합의 크기

Rank	Doc ID	Relevance Score (=Gain)	Naïve Discounted Gain	Discounted Gain
1	678	3	3 / 1	7.00
2	345	2	2 / 2	1.89
3	124	3	3 / 3	3.50
4	589	0	0 / 4	0.00
5	894	1	1 / 5	0.39
6	532	2	2 / 6	1.07
Sum				13.85

Rel_r → 순위 r에서의 Relevance Score값

$$DCG_k = \sum_{r=1}^{k} \frac{2^{rel_r} - 1}{\log_2(1+r)}$$

$DCG_6 = 13.85$

NDCG (Normalized Discounted Cumulative Gain)

이진적합도인 경우

Rank	Doc ID	Relevance Score (=Gain)	Naïve Discounted Gain	Discounted Gain
1	678	1	1 / 1	
2	345	1	1 / 2	
3	124	1	1 / 3	
4	589	0	0 / 4	
5	894	0	0 / 5	
6	532	1	1 / 6	
Sum				

Rel_r → 순위 r에서의 Relevance Score값

$$DCG_k = \sum_{r=1}^{k} \frac{2^{rel_r} - 1}{\log_2(1+r)}$$

$$\text{순위 } r = 1 \rightarrow \frac{2^{rel_1} - 1}{\log_2(1+1)} = \frac{2^1 - 1}{\log_2(1+1)} = \frac{1}{1} = 1$$

$$\text{순위 } r = 4 \rightarrow \frac{2^{rel_4} - 1}{\log_2(1+4)} = \frac{2^0 - 1}{\log_2(1+4)} = \frac{1-1}{\log_2(1+4)} = 0$$

NDCG (Normalized Discounted Cumulative Gain)

- NDCG (정규차감누적사용자만족도)
  - 상위 K개 검색문서에 대한 정규화된 차감누적사용자만족도
    - DCG_K를 $IDCG_K$(Ideal DCG_K =DCG_K의 최적값)으로 정규화함

Rank	Doc ID	Relevance Score (=Gain)	Naïve Discounted Gain	Discounted Gain	Optimal Doc ID	Optimal Relevance Score	Optimal Discounted Gain
1	678	3	3 / 1	7.00	678	3	7.00
2	345	2	2 / 2	1.89	124	3	4.42
3	124	3	3 / 3	3.50	345	2	1.50
4	589	0	0 / 4	0.00	532	2	1.29
5	894	1	1 / 5	0.39	894	1	0.39
6	532	2	2 / 6	1.07	589	0	0.00
Sum				13.85			14.60

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

$NDCG_6$ = 13.85 / 14.60 = 0.95

trec_eval

- trec_eval
  - 정보검색 성능 평가용 표준 프로그램
  - https://trec.nist.gov/trec_eval/

Test Collection for IR Evaluation

- 검색 성능 평가를 위한 *테스트 컬렉션* 구성 요소
  - 문서 집합 (a document collection)
  - 질의 집합 (a set of queries)
  - 질의-문서 간 적합성 판단 자료 (relevance judgments)

한국어 정보검색 테스트 컬렉션

- KTSET ("대용량 음성(음향)/언어/영상 DB 구축 및 표준화" 제1차 사업 발표회 보고서, 2000. URL: http://semanticweb.kaist.ac.kr/research/ksurimal/report/%C1%A4%BA%B8%B0%CB%BB%F6%C6%F2%B0%A1%B9%E6%B9%FD%B7%D0.htm)
  - 버전 1.0 (한국과학기술원, 1994) : 문서 1053, 질의 30
  - 버전 2.0(박영찬 외, 1995): 문서 4144, 질의 50
- HANTEC 2.0 (HANgul Test Collection)
  - 문서집합: 12만 건 (사회과학, 과학기술, 일반종합 분야)
  - 질의: 50개
  - http://www.kristalinfo.com/download/

영어 정보검색 테스트컬렉션

- The *Cranfield* collection (http://ir.dcs.gla.ac.uk/resources/test_collections/cran/)
  - 1,398 abstracts, 225 queries, exhaustive relevance judgments of all query-document pairs
- *CACM* collection (http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/)
  - Titles and abstracts from the journal Communication of the ACM from 1958-1979
  - 3k+ docs, 64 words/doc, 64 queries, 13 words/query, 16 reldocs/query
- *AP* collection
  - Associated Press newswire documents (1988-1990)
  - 242k+ docs, 474 words/doc, 100 queries (TREC topics 51-150), 4.3 words/query, 220 reldocs/query
- *GOV2* collection
  - Web pages crawled from .gov domain websites during early 2004
  - 25m+ docs, 1073 words/doc, 150 queries (TREC topics 701-850), 3.1 words/query, 180 reldocs/query

Pooling technique

- Pooling technique
  - Relevance judgments 구축 방법
  - 서로 다른 검색 알고리즘을 통해 얻어진 서로 다른 상위 k개 (k=50~200)의 문서들의 (중복 제거) 모음을 대상으로 수작업 적합성 판단 수행
    - 문서 모음 내 문서들은 특정한 랜덤 순서로 적합성 판단 작업자에게 제시
  - 대용량 문서 집합에 대한 적합성 판단 자료를 구축하는 현실적 대안
  - 새로운 검색 알고리즘이 기존 pool에 포함되지 않았던 많은 적합 문서들을 검색한다면?

References

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge University Press.
- Bruce Croft, Donald Metzler, Trevor Strohman. (2009). Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto. (1999). Modern Information Retrieval. Addison-Wesley Publishing Company.
- https://en.wikipedia.org/
- 박영찬, 최기선, 김영환, 김재군. 1996. 한국어 정보검색연구를 위한 시험용 데이터 모음 2.0(KTSET 2.0) 개발. 한국어정보과학회 인공지능연구회 춘계학술 발표. pp.59~65.