Evaluation Metrics

국민대학교 소프트웨어학부
강 승 식

Contents

1. Introduction

- 정보검색 시스템 성능 평가
  - 시스템의 성능 평가:
    - 시간, 공간
  - 검색 결과의 정확성 평가

- 검색엔진 평가 collections
  - 문헌의 컬렉션
  - 사용자 정보요구 집합(질의 집합)
  - 정보요구에 연관된 문헌 집합

Retrieval performance evaluation

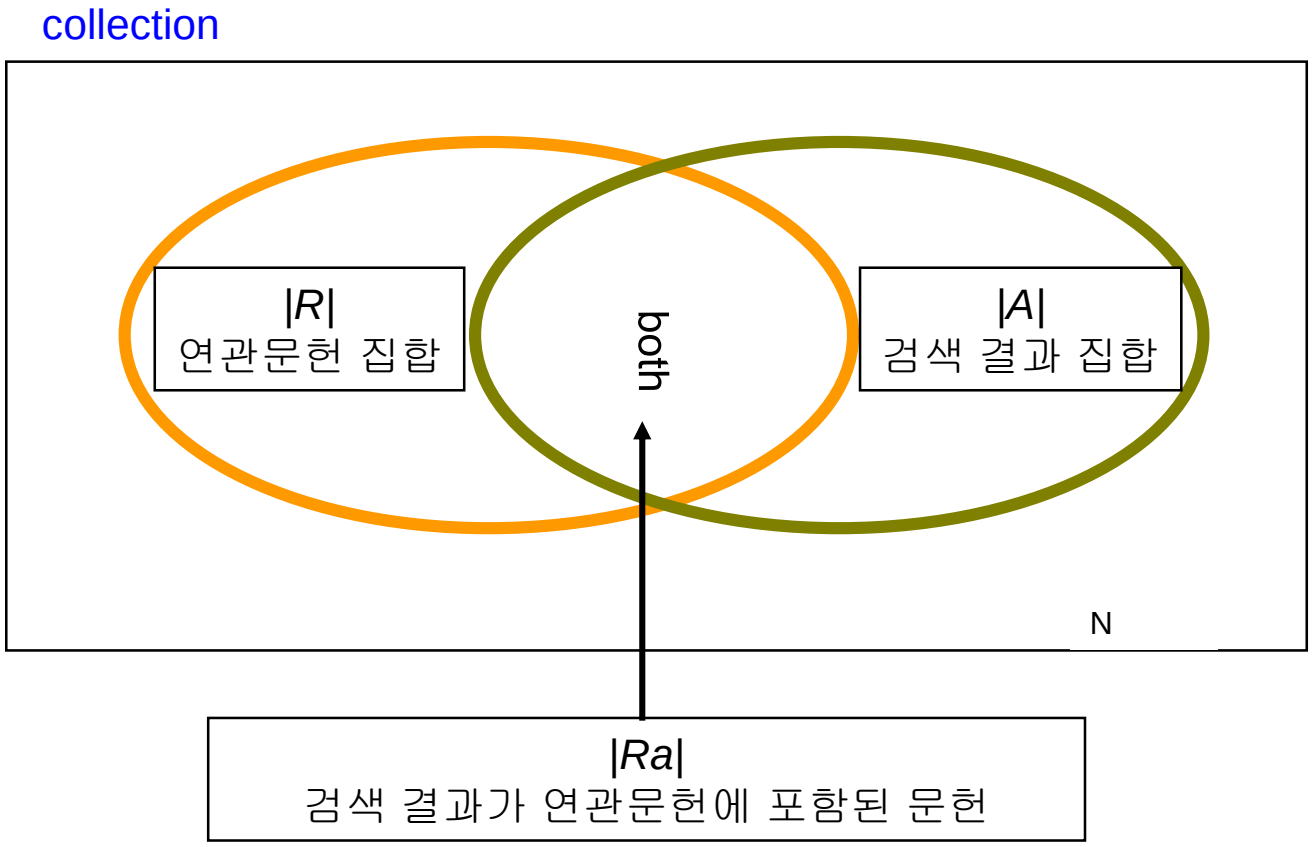
- 성능평가 고려요인 (검색 작업에 따른 평가)
  - 일괄처리 작업
    - 전통적인 검색결과 수집의 형태
    - 평가 요인 : 응답 집합의 품질
  - 대화형 작업
    - 검색과정을 interactive하게 전개해 나가는 방법
    - 평가 요인
      - 사용자의 노력
      - 인터페이스 특성
      - 시스템 제공 안내
      - 세션의 길이

Recall and Precision

- Recall(재현율)
  - 전체 적합문헌에 대한 검색된 적합문헌의 비율
  - Recall = |Ra| / |R|
    - |R| - 컬렉션에서 적합 문헌의 수
    - |Ra| - 검색된 적합 문헌의 수

- Precision(정확률)
  - 전체 검색문헌에 대한 검색된 적합문헌의 비율
  - Precision = |Ra| / |A|
    - |A| - 질의에 의해서 검색된 문헌 수

Recall and Precision



Recall and Precision

- R_q : 질의 q에 대한 연관 문헌 집합 10개인 경우

  $R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{389}, d_{123}\}$

  (1) 질의 q에 의해 검색된 문헌의 ranking : **d_{123}**, d_{84}, **d_{56}**
  - Precision(정확률) : 0.66 (= 2/3)
  - Recall(재현율) : 0.20 (= 2/10)

  (2) 질의 q에 의해 검색된 문헌의 ranking : **d_{123}**, d_{84}, **d_{56}**, d_6, d_8, **d_9**
  - Precision(정확률) : 0.50 (= 3/6)
  - Recall(재현율) : 0.30 (= 3/10)

Recall and Precision

- 질의 q에 대해 검색된 문헌 15개의 ranking 예

(100%,10%) 1. **d_{123}*!** 6. **d_9*!** (50%,30%) 11. d_{38}

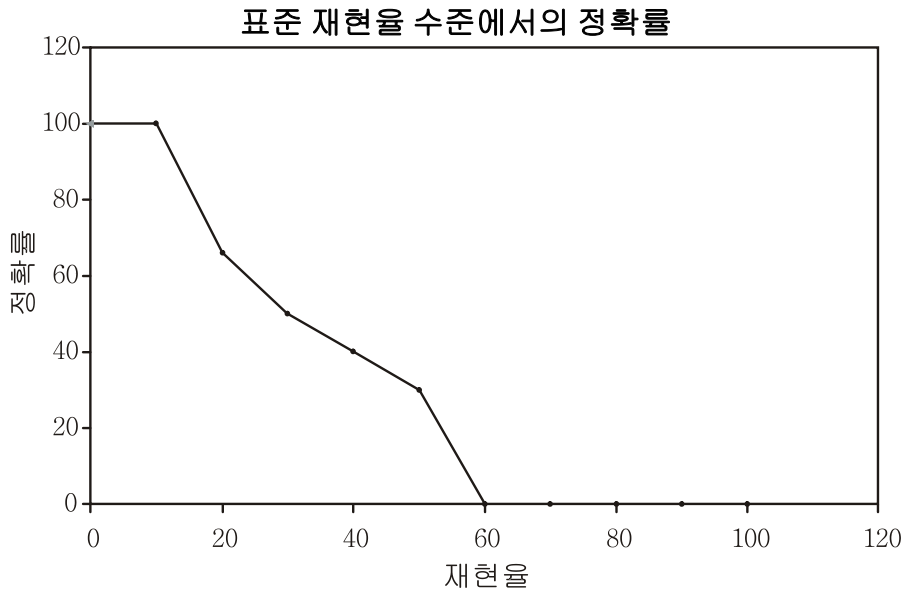
2. d_{84} 7. d_{511} 12. d_{48}

(66%,20%) 3. **d_{56}*!** 8. d_{129} 13. d_{250}

4. d_6 9. d_{187} 14. d_{113}

5. d_8 10. **d_{25}*!** (40%,40%) 15. **d_3*!** (33%,50%)

- 1: (1/1, 1/10)
  2: (1/2, 1/10)
  3: (2/3, 2/10)
  4: (2/4, 2/10)
  5: (2/5, 2/10)
  6. (3/6, 3/10)
  ….



Recall and Precision

- 검색 알고리즘 평가
  - 여러 개의 질의를 수행하여 평가
- 평균 정확률(average precision)

$$\overline{P}(r) = \sum_{i=1}^{Nq} \frac{P_i(r)}{Nq}$$

- P(r): 재현율 level r에 따른 평균 정확률
- N_q: 질의 개수
- $P_i(r)$: i번째 질의에 대한 재현율 level r에 따른 정확률

Recall and Precision

- $R_q=\{d_3, d_{56}, d_{129}\}$ 인 경우에 (precision, recall)

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
(33.3%,33.3%) 3. d_{56} ●	8. d_{129} ● (25%,66.6%)	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3 ● (20%,100%)

- 연관문헌 개수가 3개이므로 재현율 level(33%, 67%, 100%)에 따른 정확률로 계산해야 함
- 재현율 level(0%, 10%, 20%, 30%, ...)에 따른 정확률 계산이 어렵다

Single Value Metrics

- 재현율-정확률을 한 개 수치로 표현 방법
  - 평균 정확률
  - R-정확률
  - 정확률 히스토그램
  - 요약 테이블 통계치
  - F-measure

평균 정확률

- **검색된 연관 문헌에서의 평균 정확률**
  - 새로운 연관 문헌이 검색될 때마다 정확률을 계산하여 평균을 구하는 방법

예) 연관 문헌 수:	1	2	3	4	5
정확률	1	0.66	0.5	0.4	0.3

$$\mathbf{P}_{\mathbf{avg}} = (1 + 0.66 + 0.5 + 0.4 + 0.3) / 5 = 0.57$$

- 연관 문헌들이 상위 rank되는 시스템이 좋은 성능

R-Precision (R-정확률)

- R 번째 검색 순위에서 정확률 계산

  R : 각 질의에 대한 연관 문헌 개수

- R-정확률 = 0.4

  (R=10일 때, 순위 10위 안에 연관문헌 4개일 경우)

- R-정확률 = 0.33

  (R=3 일 때, 순위 3위 안에 연관문헌 1개일 경우)

정확률 히스토그램(Precision histogram)

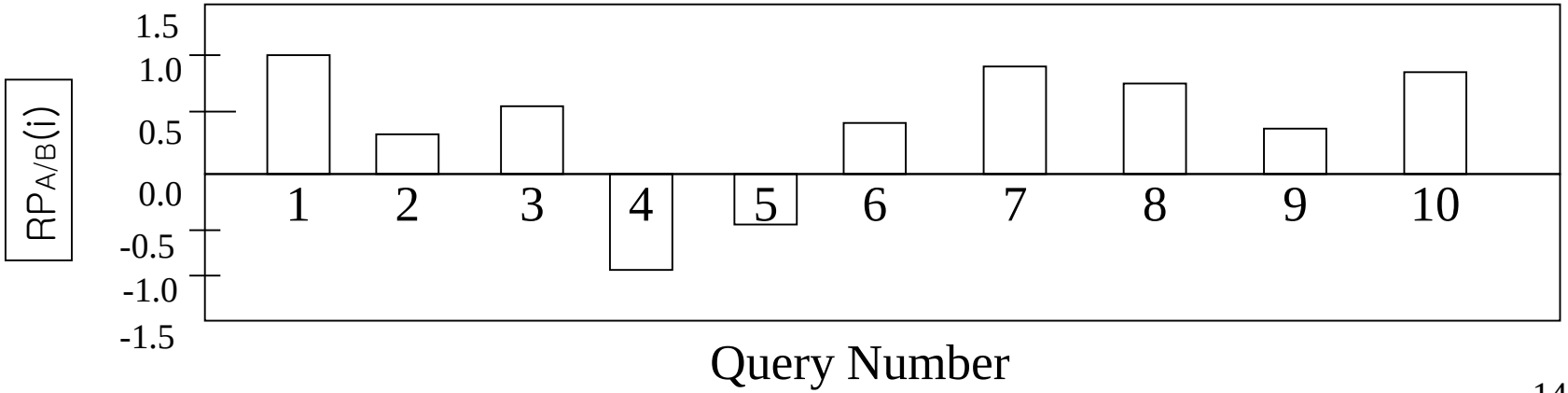
- 두 알고리즘에 대한 R-정확률 차이를 그린 막대 그래프
- 어떤 두 알고리즘의 검색 이력을 비교하는데 사용

  예제)

  $RP_{A/B}(i) = RP_A(i) - RP_B(i)$

  $RP_A(i)$: i번째 질의에 대한 검색 알고리즘 A의 R-정확률

  $RP_B(i)$: i번째 질의에 대한 검색 알고리즘 B의 R-정확률
- 두 알고리즘의 성능 차이를 시각적으로 확인할 수 있다.



요약 테이블 통계치 (Summary table statistics)

- 모든 질의들에 대한 단일 수치를 테이블로 작성
  예)
  - 검색 작업에 사용된 질의 수
  - 전체 질의에 의해 검색된 문헌 수
  - 전체 질의에 의해 검색될 수 있는 연관 문헌의 수
  - 모든 질의를 고려할 때, 검색된 연관 문헌의 수

Recall and Precision

- 재현율과 정확률의 문제점
  - 높은 재현율을 얻기 위해서 컬렉션에 있는 모든 문헌에 대한 지식이 필요하다.
    - 대규모 컬렉션일 경우에는 불가능하다.
  - 질의의 개별적인 특성은 관찰하기 위해서 재현율과 정확률을 사용하는 것은 적절하지 않다.
    - 재현율과 정확률은 시스템의 전체 성능은 관찰할 수 있다.
    - 단일 수치를 사용하는 것이 바람직하다.
  - 대화형 검색에서 재현율과 정확률은 적합하지 못하다.
    - 최근의 대부분 검색 시스템은 대화형 검색 시스템이다.
  - 검색 결과를 순위화하지 않을 때, 재현율과 정확률을 사용하는 것은 적합하지 않다.

Alternative Measures

- 조화 평균(Harmonic mean)
- E 척도(E-measure)
- User-oriented measure

조화 평균(Harmonic mean)

- 조화 평균(Harmonic mean) : F(j)
  - 재현율과 정확률이 모두 높아야 조화평균이 높다
  - r(j) : j번째 순위의 문헌의 재현율
  - p(j) : j번째 순위의 문헌의 정확률
  - F(j) = 0 : 연관된 문헌이 하나도 검색되지 않음
  - F(j) = 1: 연관된 문헌이 모두 검색됨

$$F(j) = \frac{2}{\frac{1}{R(j)} + \frac{1}{P(j)}}$$

$$F = \frac{2 \times P \times R}{P + R}$$

E-measure, F-measure

- E 척도(E-measure) : E(j)
  - 재현율에 더 관심이 있는지, 정확률에 더 관심이 있는지를 명시
  - b : 재현율과 정확률의 중요도를 조절하는 매개변수

    b = 1 : E(j) 척도는 조화평균 F(j)의 보수(complement)

    b > 1 : 정확률을 강조

    b < 1 : 재현율을 강조

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{R(j)} + \frac{1}{P(j)}}$$

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$

User-oriented Measures

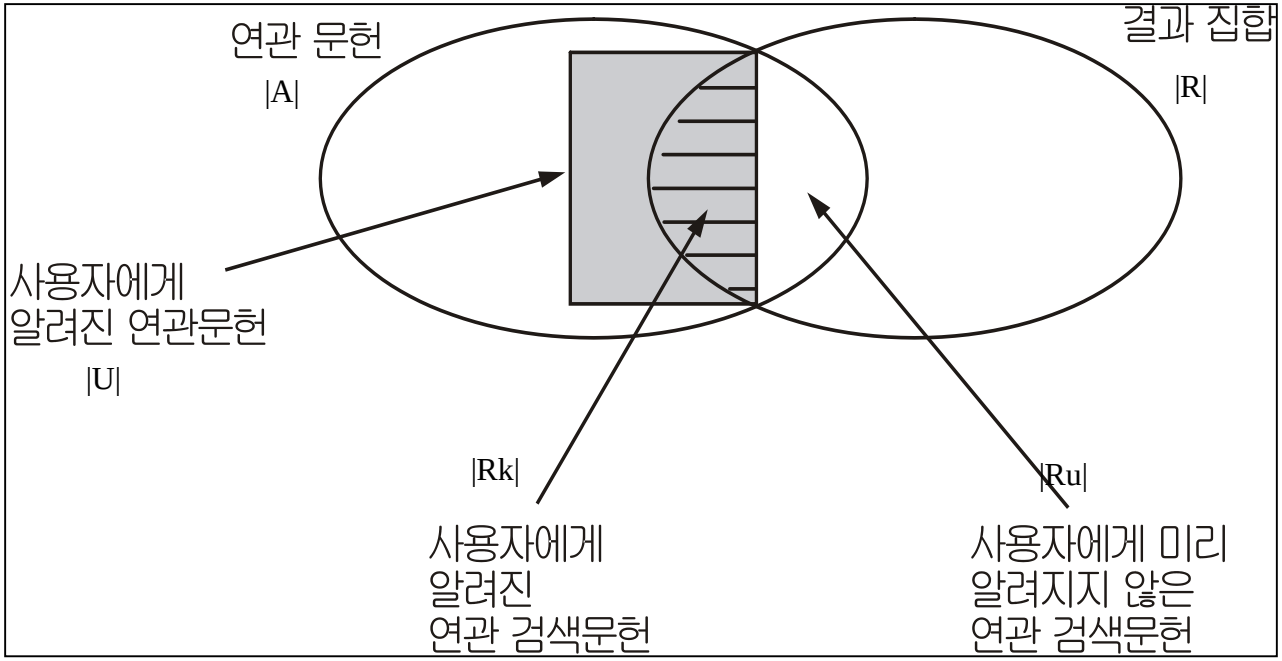
- 가정: 사용자에 따라 연관 문헌이 서로 다르다는 관점의 척도

- 커버율, 적용율(coverage ratio) = |Rk| / |U|
  - 사용자에게 미리 알려진 연관문헌 중에서 실제로 검색된 연관 문헌의 비율
  - 높은 적용율 : 검색 시스템이 사용자가 기대하는 대부분의 연관문헌을 검색
- 신 문헌율(novelty ratio) = |Ru| / (|Ru| + |Rk|)
  - 검색된 연관 문헌 중에서 사용자에게 미리 알려지지 않은 문헌의 비율
  - 높은 신문헌율: 새로운 연관 문헌을 많이 검색
- 상대 재현율(relative recall)
  - 검색한 연관문헌 수와 사용자가 검색하기를 기대하는 연관문헌 수 사이의 비율

$$\frac{|R_k|+|Ru|}{|U|}$$

- 재현 노력도(recall effort)
  - 사용자가 기대하는 수의 연관 문헌을 발견하기 위해 검사해야 하는 문헌 수 사이의 비율

Alternative Measures



Alternative Measures

- 검색 결과의 요약
  - 사용자가 알고 있는 연관 문헌의 수 : 15
  - 검색된 연관 문헌의 수: 10
  - 검색된 연관 문헌 중에 알고 있는 문헌의 수: 4

  - 적용율 : 4 /15
  - 신문헌율 : 6/ 10 (새로운 관련 문헌 수: 6)

  - 상대 재현율: 3 / 5
    - 가정: 시스템이 사용자에게 20개의 문헌을 보여주었는데 그 중에 5개는 원하는 연관문헌이나 20개 중에는 3개만 연관 문헌이다.

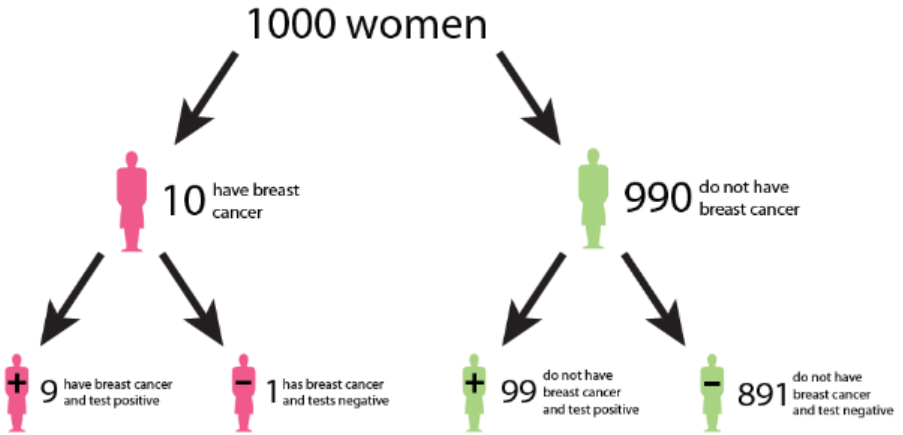
Other Metrics

- Medical testing, binary classification(ML)
  - False positives (Type I error)
  - False negatives (Type II error)

- M.T. and Summarization
  - BLEU
  - ROUGE

False positives and false negatives

- Medical testing, binary classification, etc



- Ex) Pregnant or cancer, guilty, spam email?

TP, TN, FP, FN

- True positive (진양성), True negative (진음성)

- False positive error (위양성)
  - False alarm or Type I error(음성을 양성으로 판정)
  - *Ex) Decision is a cancer/pregnant, but actually it is not.*

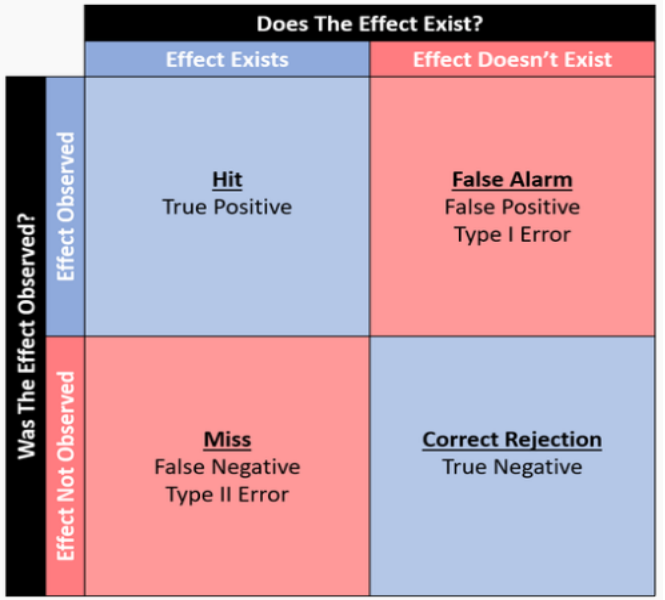
- False negative error (위음성)
  - Type II error (양성을 음성으로 판정)
  - *Ex) Decision is not a cancer/pregnant, but actually it is.*

Machine Learning

- Binary classification: spam filtering

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Confusion matrix



Commonly used terms for the cells in a confusion matrix.

$$\text{False Positive Rate(FPR)} = \frac{FP}{FP+TN}$$

$$\text{False Negative Rate(FNR)} = \frac{FN}{TP+FN}$$

$$\text{True Positive Rate(TPR)} = \frac{TP}{TP+FN}$$

$$\text{True Negative Rate(TNR)} = \frac{TN}{TN+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

BLEU, ROUGE

- Compare an automatically produced summary or translation against a reference or a set of references (human-produced).
- BLEU (Bilingual Evaluation Understudy)
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- ROUGE and BLEU are based on n-gram to measure the similar between the summaries of systems and the summaries of human.

- Bleu measures precision:
  - how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries.

- Rouge measures recall:
  - how much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries.

BLEU

- Evaluating the quality of text
  - *"The closer a machine translation is to a professional human translation, the better it is."*
  - Correspondence between machine and human

- Scores are calculated for
  - individual translated segments - generally sentences
  - by comparing them with a set of good quality reference translations

BLEU 문제점 및 해결 방법

- 문제점: 아래 예에서 P=1 (7/7)

Example of poor machine translation output with high precision

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

- 해결 방법: max count, n-grams

Comparing metrics for candidate "the the cat"

Model	Set of grams	Score
Unigram	"the", "the", "cat"	$\frac{1+1+1}{3}=1$
Grouped Unigram	"the"*2, "cat"*1	$\frac{1+1}{2+1}=\frac{2}{3}$
Bigram	"the the", "the cat"	$\frac{0+1}{2}=\frac{1}{2}$

ROUGE

- ROUGE-N:
  - Overlap of N-grams between the system and reference summaries.
  - ROUGE-1 : the overlap of unigram (each word)
  - ROUGE-2 : the overlap of bigrams
- ROUGE-L:
  - Longest Common Subsequence (LCS) based statistics.
  - LCS problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.
- ROUGE-W:
  - Weighted LCS-based statistics that favors consecutive LCSes .
- ROUGE-S:
  - Skip-bigram based co-occurrence statistics.
  - Skip-bigram is any pair of words in their sentence order.
- ROUGE-SU:
  - Skip-bigram plus unigram-based co-occurrence statistics.

Reference Collections

- TIPSTER/TREC collection
- CACM and ISI collections
- Cystic Fivrosis collection
- Korean collections

TREC Collection

- 정보 검색 연구에 대한 비판
  - 객관적인 평가 기준이 없었다
  - 일관성 있는 테스트베드와 벤치마크가 없다

- TREC 컬렉션
  - 1990년 초: NIST, Donna Harman이 Text REtrieval Conference (TREC) 창설
  - 제1회 TREC 대회: 1992년 11월
  - 구성: 문헌집합, 정보요구(질의), 연관문헌 집합
  - 6 CD-ROM : 1GB, 압축된 텍스트 형태
  - 종류:
    - WSJ : wall street Journal
    - AP : Associated Press (news)
    - ZIFF : Computer Selects (articles)
    - FR : Federal Register
    - DOE : US DOE Publications (abs)
    - SJMN : San Jae Mercury News
    - PAT : US Patents
    - FT : Financial Times
    - CR Congressional Record
    - FBIS : Foreign Broadcast Information Service
    - LAT : LA Times

Disk	*Contents*	*Size Mb*	*Number Docs*	*Words/Doc. (median)*	*Words/Doc. (mean)*
1	WSJ, 1987-1989	267	98,732	245	434.0
	AP, 1989	254	84,678	446	473.9
	ZIFF	242	75,180	200	473.0
	FR, 1989	260	25,960	391	1315.9
	DOE	184	226,087	111	120.4
2	WSJ, 1990-1992	242	74,520	301	508.4
	AP, 1988	237	79,919	438	468.7
	ZIFF	175	56,920	182	451.9
	FR, 1988	209	19,860	396	1378.1
3	SJMN, 1991	287	90,257	379	453.0
	AP, 1990	237	78,321	451	478.4
	ZIFF	345	161,021	122	295.4
	PAT, 1993	243	6,711	4,445	5391.0
4	FT, 1991-1994	564	210,158	316	412.7
	FR, 1994	395	55,630	588	644.7
	CR, 1993	235	27,922	288	1373.5
5	FBIS	470	130,471	322	543.6
	LAT	475	131,896	351	526.5
6	FBIS	490	120,653	348	581.3

TREC Collection

<doc>

<docno> WSJ880406-0090 </docno>
<hl> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </hl>
<author> Janet Guyon (WSJ Staff) </author>
<dateline> New York </dateline>

<text>
American Telephone & Telegraph Co. introduced the first of a new generation of phone
services with broad …

…

</text>

</doc>

TREC Collection (topic)

- 용도
  - 새로운 순위화 알고리즘을 실험하기 위한 집합
  - 3개의 필드로 구성되어 있음
    - (1) 제목, (2) 기술, (3) 설명

<top>

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

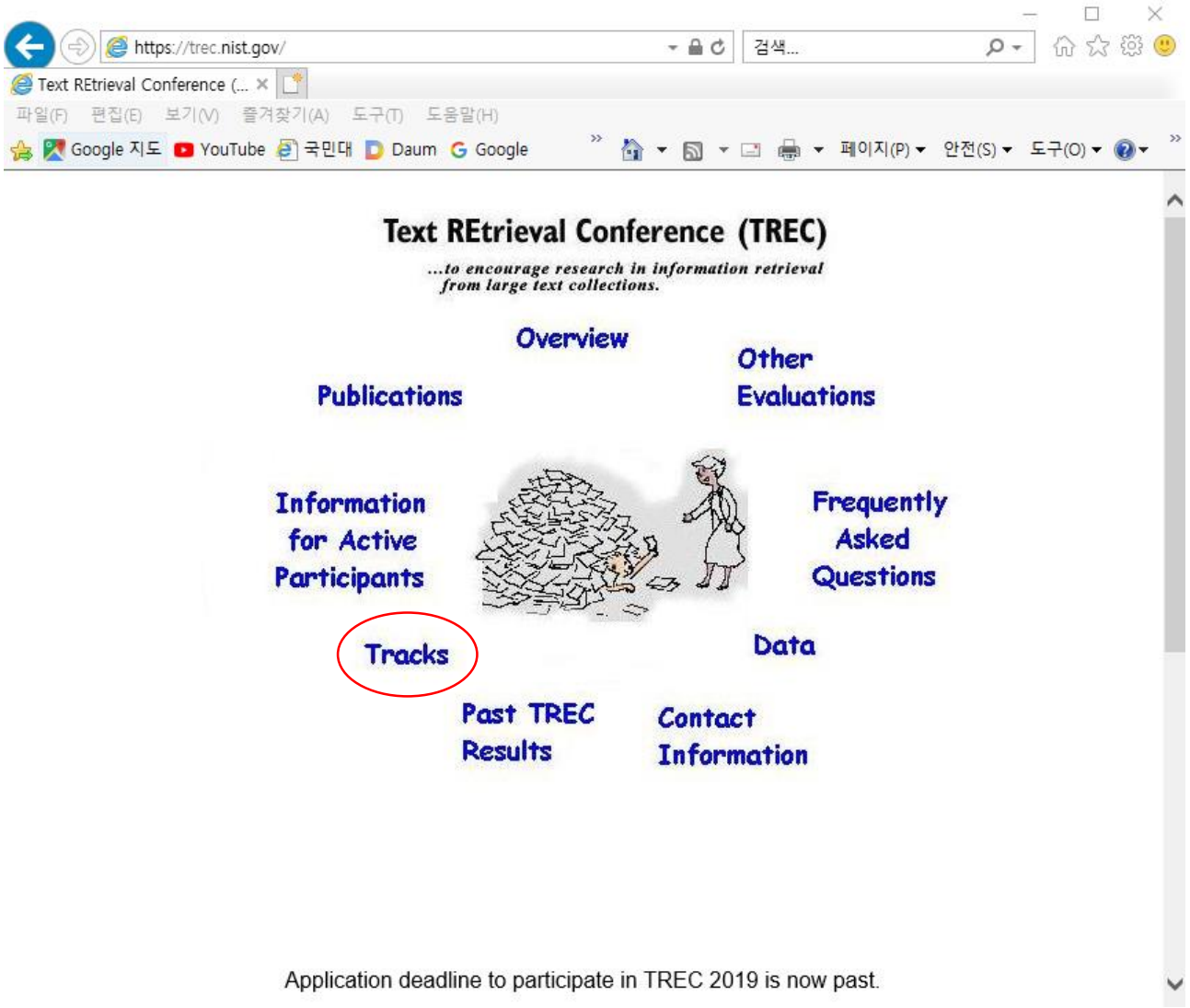
<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

</top>

TREC Collection

- 연관문헌의 선정 방법
  - 풀링 방법(pooling method)
    - 주어진 질의에 대해 검색 시스템으로부터 검색된 문헌 중 상위 K의 문헌을 하나의 풀을 생성한다.(보통 K=100)
    - 이들 풀에 속한 K개의 문헌을 전문가에 의해서 연관 여부를 결정한다.
    - 가정
      - 연관문헌의 대부분은 풀에 포함될 것이다.
      - 풀에 포함되지 않은 문헌은 비연관 문헌이다.
- TREC 학술회의(벤치마크) 작업
  - 축적 검색(ad-hoc)
    - 변하지 않는 문헌 컬렉션에 대해서 여러 질의를 적용하는 방법
  - 라우팅(routing, filtering)
    - 사용자 요구인 질의는 고정되고 문헌 컬렉션이 변하는 경우이다.
    - 같은 질의가 동적인 문헌 집합을 대상으로 실행되는 여과(filtering) 작업
      (예, 뉴스 클립핑 서비스)
    - 순수 여과 작업과는 달리 검색된 문헌은 순위화
    - 실험 정보 요구와 2개의 서로 다른 문헌 컬렉션 제공
      (1.검색 알고리즘의 학습과 튜닝, 2. 튜닝 된 알고리즘의 테스트)

TREC 경진대회: https://trec.nist.gov/



TREC Tracks 예제

- 중국어(Chinese):
  - 문헌과 토픽 모두가 중국어로 된 축적 검색 작업
- 여과(filtering):
  - 새로 도착한 문헌이 연관문헌인지 아니지만 결정하는 라우팅 작업이며, 문헌 순위화 하지 않고, 테스트 자료는 도착 순서대로 처리
- 대화(interactive):
  - 탐색자가 문헌의 연관성을 평가하기 위하여 정보 검색 시스템과 대화적으로 작업하며, 문헌은 연관 혹은 비연관 문헌으로 구분(순위화 비제공)
- 자연언어 처리(natural language):
  - 자연언어 처리에 기반을 둔 검색 알고리즘이 기존의 색인어를 이용한 검색 알고리즘에 비해 장점이 있는지 여부를 검증하기 위한 작업
- 다국어 축적 검색(cross language):
  - 문헌은 하나의 언어를 사용하나 질의는 여러 가지 다른 언어를 사용
- 높은 정확률(high precision):
  - 정보 검색 시스템 사용자가 주어진 정보 요구(이전에 알려지지 않은)에 대한 응답으로 5분 이내에 10개의 문헌을 검색하도록 하는 작업
- 구어체 문헌 검색(Spoken document retrieval):
  - 라디오 방송의 뉴스 쇼를 기록한 문헌을 검색하는 작업이며, 구어체 문헌 검색 기술에 대한 연구를 촉진하기 위한 것임
- 대용량 코퍼스(Very large corpus):
  - 축적 검색 작업으로 검색 시스템은 20 기가바이트(7500만 문헌) 크기의 컬렉션을 처리해야 함

TREC collections

https://trec.nist.gov/data.html

Data

TREC home NIST HOME

Versions of trec_eval

Ad hoc Test Collections

Web Test Collections

Blog Track

Chemical IR Track

Clinical Decision Support Track

Common Core Track

Confusion Track

Contextual Suggestion Track

Crowdsourcing Track

Dynamic Domain Track

Enterprise Track

Entity Track

Filtering Track

Federated Web Search Track

Genomics Track

HARD Track

Interactive Track

Knowledge Base Acceleration Track

Legal Track

Medical Track

Microblog Track

Million Query Track

News Track

Novelty Track

Query Track

Question Answering Track

Precision Medicine Track

Real-time Summarization Track

Relevance Feedback Track

Robust Track

Session Track

SPAM Track

Spoken Document Retrieval Track

Tasks Track

Temporal Summarization Track

Terabyte Track

Web Track

Tracks in TREC2001

- Tracks in TREC2001
  - Cross-Language Track
  - Filtering Track
  - Interactive Track
  - Question Answering Track
  - Video Track
  - Web Track
  - …

Tracks in TREC2001

- Cross-Language Track
  - 검색시스템이 언어에 관계없이 적합한 문서를 찾아내는지를 조사하는 작업

    예) 영어 질의로 아랍 문서를 찾는 작업
  - 평가: 11-point 재현율 정확률 그래프

- Filtering Track
  - Given static queries, retrieve new documents
    - A task in which the user's information need is stable
  - For each document, the system must make a binary decision as to whether the docs. must be retrieved
  - 적합률 지향적인 측정

    R / max(MinD, (R+N)) R = 검색된 연관문헌 N = 검색된 비연관문헌

    MinD = TREC-9에서는 50

- Question Answering Track
  - A track designed to take a step closer to information retrieval rather than document retrieval
  - Returning ranked list of up to 5 [document-id, response string] pairs for each of 693 fact-based
  - 평가: Mean Reciprocal Answer Rank

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i}$$

- Web Track
  - Ad hoc search task on a document set (100GB VLC2 collection)
  - 평가: 11-point recall-precision curve, R-precision

8 Tracks in TREC2019

https://trec.nist.gov/tracks.html

Complex Answer Retrieval Track

The focus of the Complex Answer Retrieval track is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus.
Track coordinators:
Laura Dietz, University of New Hampshire
Ben Gamari, Well-typed LLP
Track Web Page:
trec-car.cs.unh.edu
Mailing list:
Google group, name: trec-car

Conversational Assistance Track

The Conversational Assistance Track is a forum for building and testing systems that engage in open-domain information centric conversational dialogues.
Track coordinators:
Jeff Dalton, University of Glasgow
Chenyan Xiong, Microsoft Research
Jamie Callan, Carnegie Mellon University
Track Web Page:
CAsT web site
Twitter: @treccast
Slack: treccast.slack.com

Decision Track

The Decision Track aims to (1) provide a venue for research on retrieval methods that promote better decision making with search engines, and (2) develop new online and offline evaluation methods to predict the decision making quality induced by search results.
Track coordinators:
Christina Lioma, University of Copenhagen
Mark Smucker, University of Waterloo
Guido Zuccon, University of Queensland
Track Web Page:
Decision Track web site
Mailing list:
Google group, name: trec-decision-track

Deep Learning Track

The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.
Track coordinators:
Nick Craswell, Microsoft
Bhaskar Mitra, Microsoft and University College London
Emine Yilmaz, University College London
Daniel Campos, Microsoft
Track Web Page:
Deep Learning track web page
Mailing list:
The list will use the new TREC Slack.

Fair Ranking Track

The Fair Ranking track focuses on building two-sided systems that offer fair exposure to ranked content producers while ensuring high results quality for ranking consumers.
Track coordinators:
Asia Biega, Max Planck Institute for Informatics
Fernando Diaz, Microsoft Research Montreal
Michael Ekstrand, Boise State University
Track Web Page:
Fairness track web page
Mailing list:
Google group, name: fair-trec

Incident Streams Track

The Incident Streams track is designed to bring together academia and industry to research technologies to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators.
Track coordinators:
Richard McCreadie, University of Glasgow
Cody Buntain, NYU
Ian Soboroff, NIST
Track Web Page:
Incident Streams track web page
Track Mailing List:
Google group, name: trec-is

News Track

The NEWS track features modern search tasks in the news domain. In partnership with The Washington Post, we will develop test collections that support the search needs of news readers and news writers in the current news environment.
Track coordinators:
Shudong Huang, NIST
Donna Harman, NIST
Ian Soboroff, NIST
Track Web Page:
trec-news.org
Mailing list:
Google group, name: trec-news-track

Precision Medicine Track

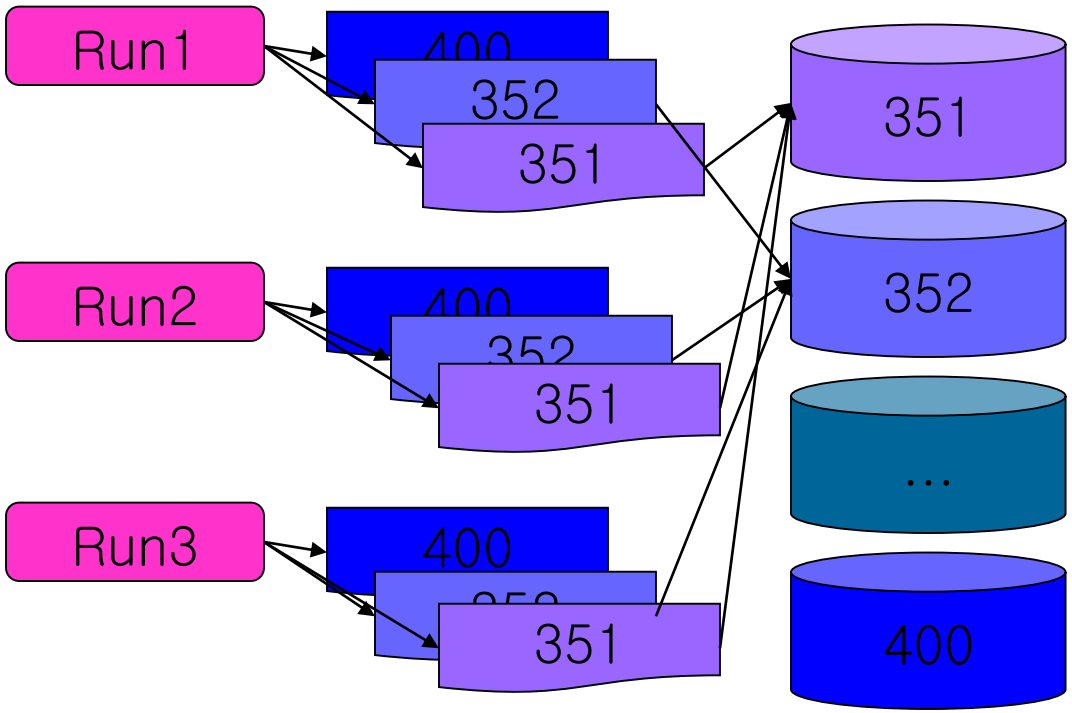
This track focuses on building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments as well as evidence-based literature to identify the most effective existing treatments.
Track coordinators:
Kirk Roberts, University of Texas Health Science Center
Dina Demner-Fushman, U.S. National Library of Medicine
Ellen Voorhees, NIST
William Hersh, Oregon Health and Science University
Alexander Lazar, University of Texas MD Anderson Cancer Center
Shubham Pant, University of Texas MD Anderson Cancer Center
Track Web Page:
www.trec-cds.org/
Mailing list:
Google group, name: trec-cds

TREC 경진대회: https://trec.nist.gov/

1. 참가자들은 검색 결과를 NIST에 보내준다.

2. NIST는 받은 결과를 모아서 Pool을 만든다.

3. NIST에서는 Pool의 각 문서에 대해서 relevance judgments를 한다. (이 과정에서 새로운 relevance judgments가 발견될 수도 있다.)



4. NIST는 relevance judgments를 이용하여 각 system을 평가한다.

TREC의 평가 측정

- 요약 테이블 통계(summary table statistics):
  - 주어진 작업에 대한 통계 값 들을 요약한 테이블.
    - 작업에 사용된 토픽(정보 요구) 수,
    - 전체 토픽에 대해 검색된 문헌의 수,
    - 전체 토픽에 대해 효과적으로 검색된 연관 문헌의 수,
    - 전체 토픽에 대해 검색했어야 할 문헌의 수

- 재현율-정확률 평균(recall-precision averages):
  - 11 표준 재현율 수준에 있어서 평균 정확률을 표시하는 그래프나 표로 구성

- 문헌 수준 평균(document level averages):
  - 전체 토픽에 대한 평균 정확률이 미리 정의된 문헌 컷 오프에서 계산

- 평균 정확률 히스토그램(average precision histogram):
  - 각 토픽에 대한 단일 수치 척도를 포함하는 그래프

CACM과 ISI 컬렉션

- CACM(Communication of the ACM) 컬렉션
  - 3204 문헌으로 구성
  - 부가 정보
    - 저자명
    - 날짜
    - 제목과 요약에서 추출된 키워드
    - 계층적 분류 체계에서 추출된 범주(Computing Review의 범주 체계)
    - 논문 사이의 직접 인용 정보
    - 서지학적 연결(bibliographic coupling) 정보
    - 두 문헌 사이에 상호 인용(co-citation) 빈도
  - 52개의 정보요구
    - 예) 1번 정보요구

      What articles exist which deals with TSS(Time Sharing System), an operating system for IBM computers

      (IBM 컴퓨터 운영체제인 TSS(시분할 시스템)에 대한 논문은 어떤 것이 있는가?)

    - 각 정보 요구에 대해, 두 개의 불리안 질의와 연관 문헌 집합을 포함한다.
    - 각 정보 요구에 대한 연관 문헌의 평균 개수는 15개 정도로 비교적 작다.
    - 정확률과 재현율 성능은 비교적 낮은 경향이 있다.

CACM과 ISI 컬렉션

- ISI(CISI) 의 1460개 문헌
  - ISI(Institute of Science Information)
  - 부가정보
    - 저자 이름
    - 제목과 요약에서 추출된 키워드
    - 각 논문 쌍에 대한 상호 인용 빈도
  - 정보요구
    - 35개의 불린 질의
    - 41개의 자연어 질의
  - 각 질의에 대한 평균 연관 문헌 수: 약 50개

Cystic Fibrosis Collection

- Cystic Fibrosis(섬유종) collection
  - 1239개의 문헌
  - 부가 정보
    - MEDLINE 병명 번호
    - 저자, 제목, 출처
    - 주요 주제, 보조 주제
    - 요약
    - 참고문헌, 인용
  - 정보요구: 100개
    - 질의 당 평균 연관 문헌의 수 : 10 ~ 30 사이의 문헌
  - 연관도(0~2까지 연관점수로 평가)
    - 0 : 연관성 없다. 1 : 연관성 중간이다. 2 : 연관성 높다

- CF(Cystic Fibrosis) 컬렉션 특징
  - 연관점수가 분야 전문가에 의해 신중한 평가 과정을 거쳐 작성된다.
  - 비교적 많은 수의 정보 요구를 포함하고 있으며, 따라서 질의 벡터들 간에 겹치는 부분이 존재한다.

Korean Collections

- KTSET
  - 제작자 : 한국 통신
  - 목적 : 정보검색 시스템의 성능 테스트
  - KTSET 2.0의 문서 구성
    - 주제 : 컴퓨터과학, 정보과학 관련 분야
    - 텍스트 구성 : 총 4,414건의 문서로 구성
      - 일반 논문 초록 1000건
      - 한국 통신 논문 초록 1414건
      - 전자신문, 잡지기사 2000건
      - 키워드 : 문서 내용에 따라 수동으로 추출한 키워드를 제공
      - 사용자 질의 : 자연어 질의와 불리언 질의를 각각 59개씩 제공
      - 질의와 문서간의 관련성(Relevancy)을 수동으로 판단하여 제공

Korean Collections

- HANTEC2.0
  - 충남대와 연구개발정보센터에서 공동 제작
  - 문서집합
    - 일반, 사회과학, 과학기술 분야에 속하는 120,000건의 다양한 크기의 문서들로 구성
  - 질의집합
    - 질의는<num>, <title>, <desc>, <narr>, <quer>의 5개로 구성되며, 과학기술(30), 사회(10), 일반(10)의 총 50개 질의로 구성
  - 적합문서집합
    - 1(부적합), 2(약산적합), 3(다소 적합), 4(적합), 5(매우 적합)의 점수부여
    - 2인의 평가자가 적합도를 평가
      - G : 평가자가 부여한 점수 중 높은 것을 채택
      - L : 평가자가 부여한 점수 중 낮은 것을 채택
      - 예) G2 : 2인의 평가자가 부여한 두 가지 점수 중에서 더 높은 점수가 2점 이상이라면 그 문서를 적합문서로 간주

Korean Collections

컬렉션	주 제	문헌수	질의수
KTSET93	전산학, 정보학	**1,000**	**30**
KTSET95	KTSET93 확장(신문기사)	**4,414**	**50**
EKSET	계몽사 백과사전	**23,000**	**46**
KRIST	과학기술 연구 보고서	**13,515**	**30**
HANTEC	일반,사회과학,과학기술	**120,000**	**50**