
Performance Evaluation of Information Retrieval Systems

Many slides in this section are adapted from Prof. Joydeep Ghosh (UT ECE) who in turn adapted them from Prof. Dik Lee (Univ. of Science and Tech, Hong Kong)

Why System Evaluation?

- There are many retrieval models/ algorithms/ systems, which one is the best?
- What is the best component for:
 - Ranking function (dot-product, cosine, ...)
 - Term selection (stopword removal, stemming...)
 - Term weighting (TF, TF-IDF,...)
- How far down the ranked list will a user need to look to find some/all relevant documents?

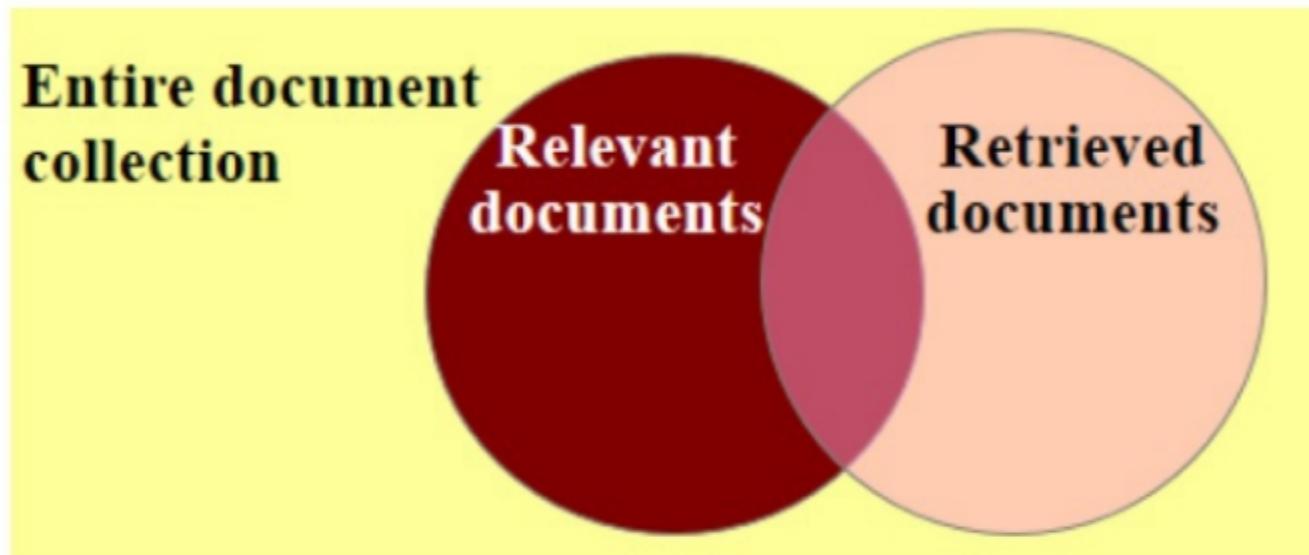
Difficulties in Evaluating IR Systems

- Effectiveness is related to the *relevancy* of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
 - Subjective: Depends upon a specific user's judgment.
 - Situational: Relates to user's current needs.
 - Cognitive: Depends on human perception and behavior.
 - Dynamic: Changes over time.

Human Labeled Corpora (Gold Standard)

- Start with a corpus of documents.
- Collect a set of queries for this corpus.
- Have one or more human experts exhaustively label the relevant documents for each query.
- Typically assumes binary relevance judgments.
- Requires considerable human effort for large document/query corpora.

Precision and Recall



		relevant	irrelevant
retrieved	retrieved & irrelevant	Not retrieved & irrelevant	
	retrieved & relevant	not retrieved but relevant	
not retrieved			

$$\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

$$\text{precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

Precision and Recall

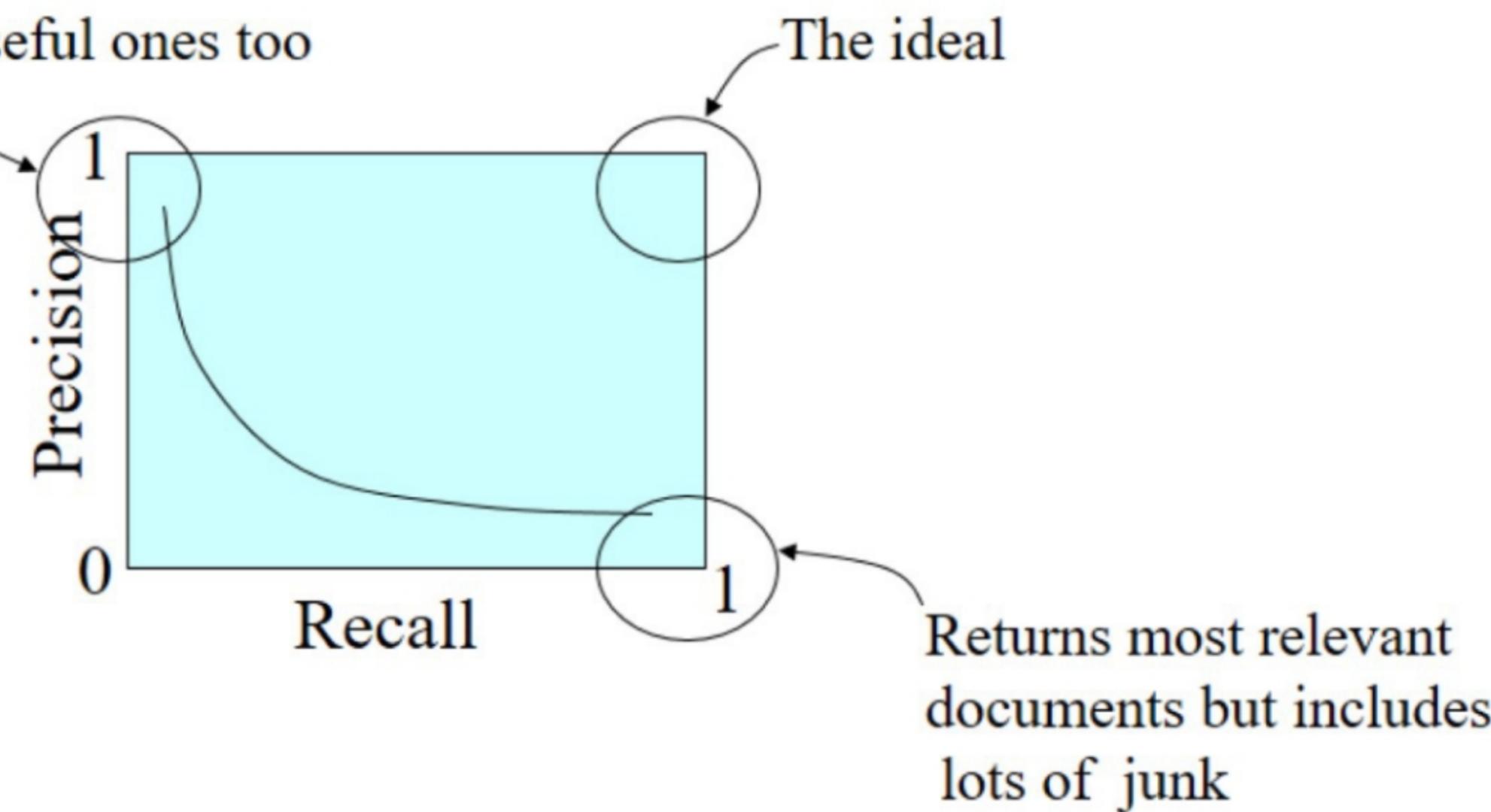
- Precision
 - The ability to retrieve top-ranked documents that are mostly relevant.
- Recall
 - The ability of the search to find *all* of the relevant items in the corpus.

Determining Recall is Difficult

- Total number of relevant items is sometimes not available:
 - Sample across the database and perform relevance judgment on these items.
 - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.

Trade-off between Recall and Precision

Returns relevant documents but misses many useful ones too

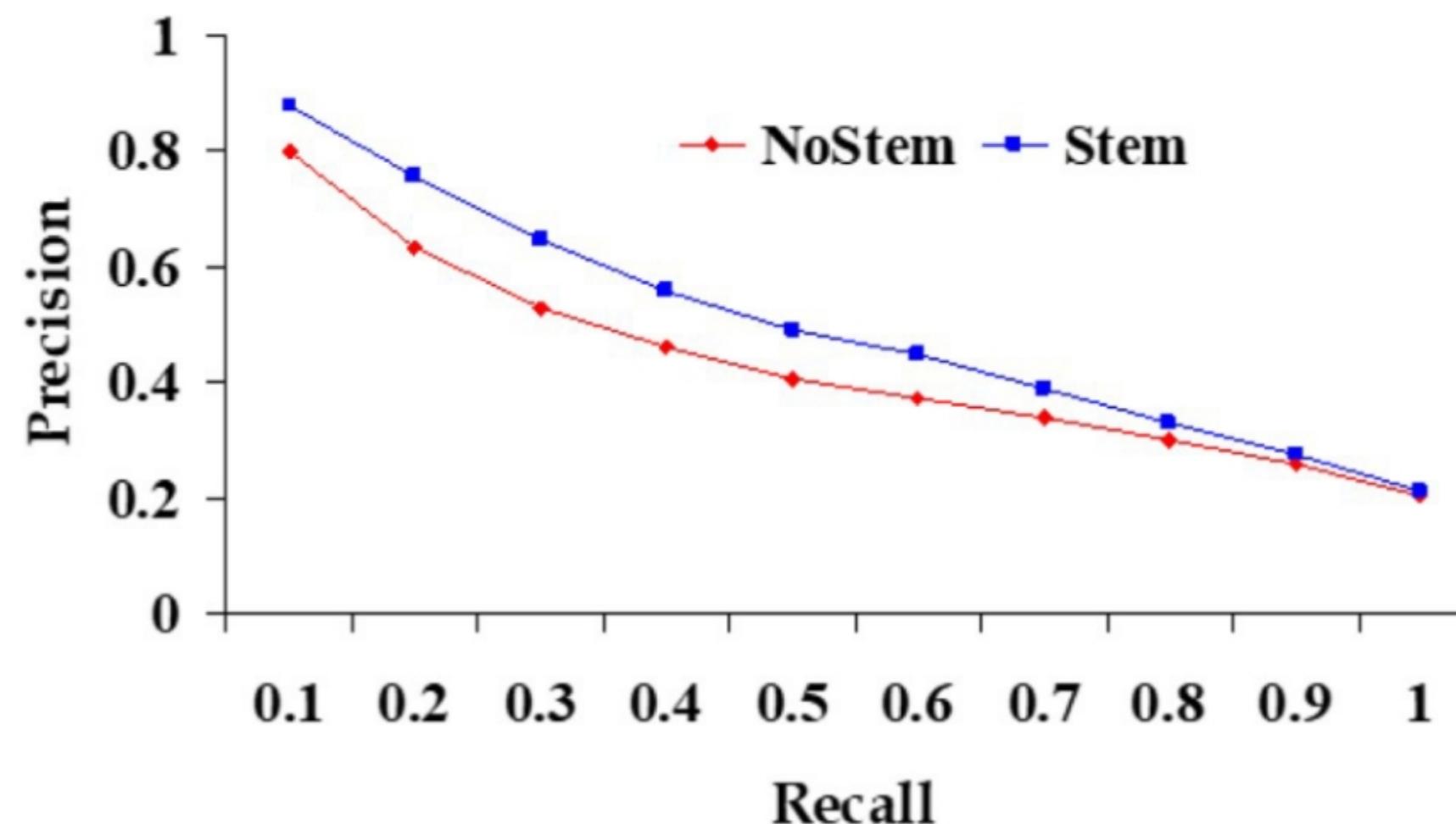


Average Recall/Precision Curve

- Typically average performance over a large *set* of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

Mean Average Precision (MAP)

- **Average Precision:** Average of the precision values at the points at which each relevant document is retrieved.
 - Ex1: $(1 + 1 + 0.75 + 0.667 + 0.38 + 0)/6 = 0.633$
 - Ex2: $(1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429)/6 = 0.625$
- **Mean Average Precision:** Average of the average precision value for a set of queries.

Issues with Relevance

- ***Marginal Relevance:*** Do later documents in the ranking add new information beyond what is already given in higher documents.
 - Choice of retrieved set should encourage **diversity** and **novelty**.
- ***Coverage Ratio:*** The proportion of relevant items retrieved out of the total relevant documents ***known*** to a user prior to the search.
 - Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000).

Other Factors to Consider

- *User effort*: Work required from the user in formulating queries, conducting the search, and screening the output.
- *Response time*: Time interval between receipt of a user query and the presentation of system responses.
- *Form of presentation*: Influence of search output format on the user's ability to utilize the retrieved materials.
- *Collection coverage*: Extent to which any/all relevant items are included in the document corpus.

A/B Testing in a Deployed System

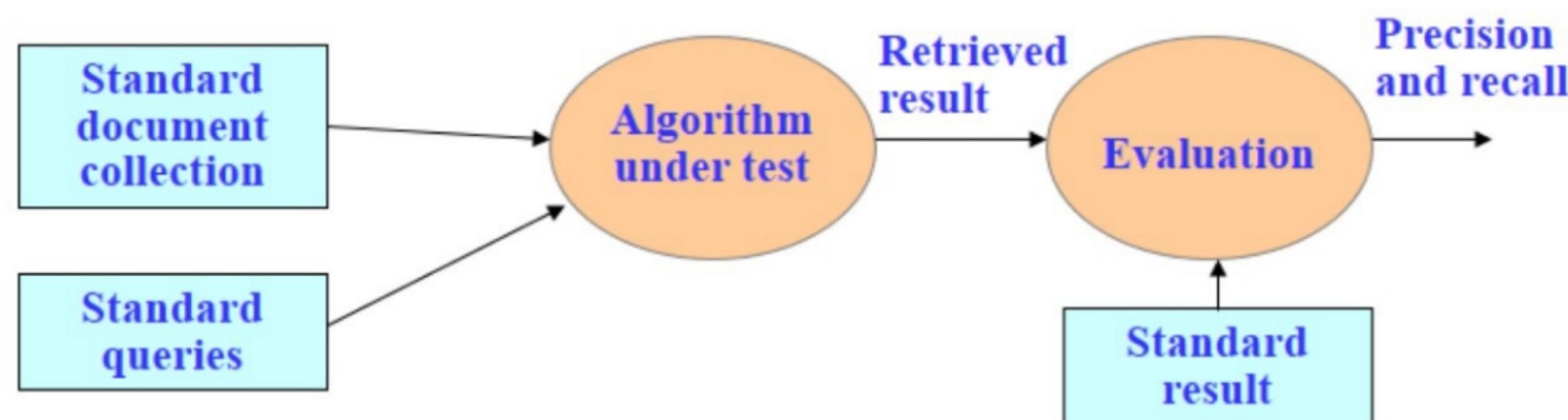
- Can exploit an existing user base to provide useful feedback.
- Randomly send a small fraction (1–10%) of incoming users to a variant of the system that includes a single change.
- Judge effectiveness by measuring change in ***clickthrough***: The percentage of users that click on the top result (or any result on the first page).

Experimental Setup for Benchmarking

- *Analytical* performance evaluation is difficult for document retrieval systems because many characteristics such as relevance, distribution of words, etc., are difficult to describe with mathematical precision.
- Performance is measured by *benchmarking*. That is, the retrieval effectiveness of a system is evaluated on a *given set of documents, queries, and relevance judgments*.
- Performance data is valid only for the environment under which the system is evaluated.

Benchmarks

- A benchmark collection contains:
 - A set of standard documents and queries/topics.
 - A list of relevant documents for each query.
- Standard collections for traditional IR:
 - Smart collection: <ftp://ftp.cs.cornell.edu/pub/smarter>
 - TREC: <http://trec.nist.gov/>



Benchmarking – The Problems

- Performance data is valid only for a particular benchmark.
- Building a benchmark corpus is a difficult task.
- Benchmark web corpora are just starting to be developed.
- Benchmark foreign-language corpora are just starting to be developed.

Early Test Collections

- Previous experiments were based on the SMART collection which is fairly small.
(<ftp://ftp.cs.cornell.edu/pub/smart>)

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

- Different researchers used different test collections and evaluation techniques.

The TREC Benchmark

- TREC: Text REtrieval Conference (<http://trec.nist.gov/>)
Originated from the TIPSTER program sponsored by
Defense Advanced Research Projects Agency (DARPA).
- Became an annual conference in 1992, co-sponsored by the
National Institute of Standards and Technology (NIST) and
DARPA.
- Participants submit the P/R values for the final document
and query corpus and present their results at the conference.

TREC Tasks

- **Ad hoc**: New questions are being asked on a static set of data.
- **Routing**: Same questions are being asked, but new information is being searched. (news clipping, library profiling).
- New tasks added after TREC 5 - Interactive, multilingual, natural language, multiple database merging, filtering, very large corpus (20 GB, 7.5 million documents), question answering.

Characteristics of the TREC Collection

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).
- Test documents consist of:

WSJ	Wall Street Journal articles (1986-1992)	550 M
AP	Associate Press Newswire (1989)	514 M
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 M
FR	Federal Register	469 M
DOE	Abstracts from Department of Energy reports	190 M

Sample Document (with SGML)

```
<DOC>
<DOCNO> WSJ870324-0001 </DOCNO>
<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>
<DD> 03/24/87</DD>
<SO> WALL STREET JOURNAL (J) </SO>
<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)
      MARKETING, ADVERTISING (MKT) TELECOMMUNICATIONS,
      BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
      John Blair & Co. is close to an agreement to sell its TV station
      advertising representation operation and program production unit to an
      investor group led by James H. Rosenfield, a former CBS Inc. executive,
      industry sources said. Industry sources put the value of the proposed
      acquisition at more than $100 million. ...
</TEXT>
</DOC>
```

Sample Query (with SGML)

```
<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language
       processing technology which is being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution
       developing or marketing a natural language processing technology,
       identify the technology, and identify one of more features of the
       company's product.
<con> Concept(s): 1. natural language processing ;2. translation, language,
       dictionary
<fac> Factor(s):
<nat> Nationality: U.S.</nat>
</fac>
<def> Definitions(s):
</top>
```

Evaluation

- **Summary table statistics:** Number of topics, number of documents retrieved, number of relevant documents.
- **Recall-precision average:** Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- **Document level average:** Average precision when 5, 10, .., 100, ... 1000 documents are retrieved.
- **Average precision histogram:** Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

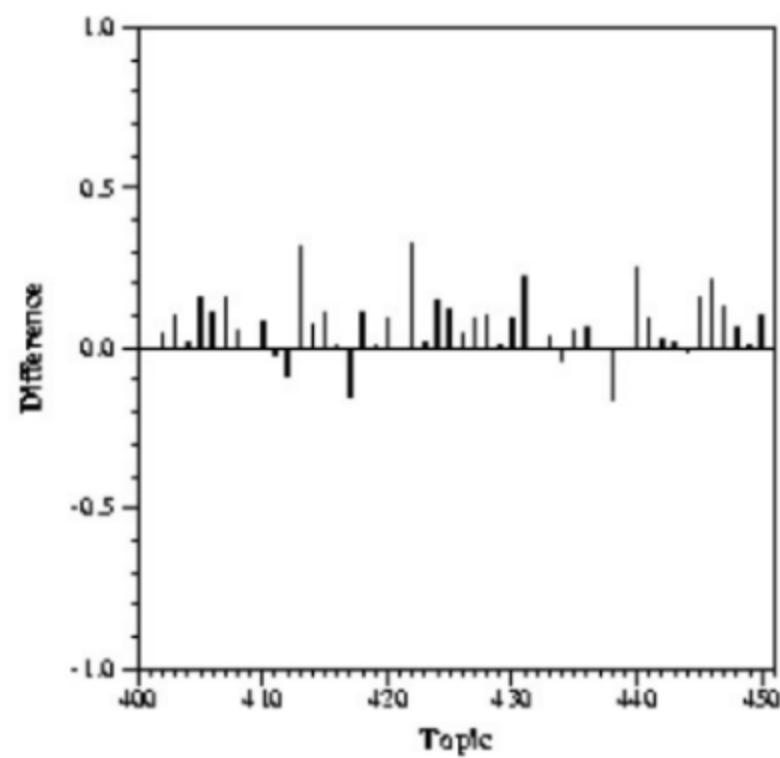
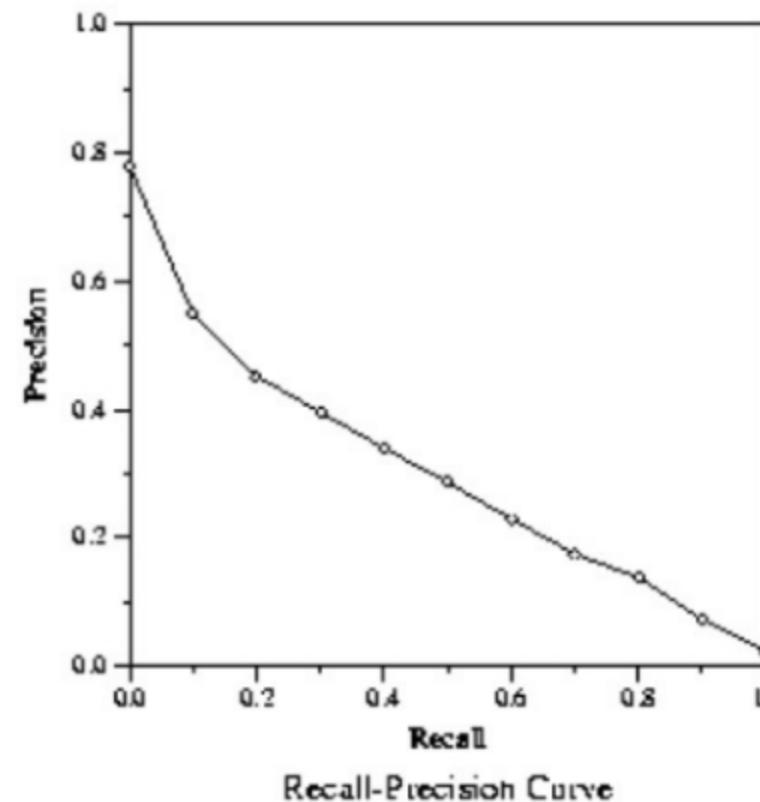
Summary Statistics	
Run Number	Flab8atd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

Recall Level Precision Averages	
Recall	Precision
0.00	0.7796
0.10	0.5490
0.20	0.4517
0.30	0.3954
0.40	0.3397
0.50	0.2863
0.60	0.2291
0.70	0.1745
0.80	0.1381
0.90	0.0720
1.00	0.0224

Average precision over all relevant docs	
non interpolated	0.2930

Document Level Averages	
	Precision
At 5 docs	0.5480
At 10 docs	0.4880
At 15 docs	0.4587
At 20 docs	0.4200
At 30 docs	0.3887
At 100 docs	0.2490
At 200 docs	0.1777
At 500 docs	0.1011
At 1000 docs	0.0598

R Precision (precision after R docs retrieved (where R is the number of relevant documents);)	
Exact	0.3203



Difference from Median in Average Precision per Topic

Cystic Fibrosis (CF) Collection

- 1,239 abstracts of medical journal articles on CF.
- 100 information requests (queries) in the form of complete English questions.
- Relevant documents determined and rated by 4 separate medical experts on 0-2 scale:
 - 0: Not relevant.
 - 1: Marginally relevant.
 - 2: Highly relevant.

CF Document Fields

- MEDLINE access number
- Author
- Title
- Source
- Major subjects
- Minor subjects
- Abstract (or extract)
- References to other documents
- Citations to this document

Sample CF Document

AN 74154352

AU Burnell-R-H. Robertson-E-F.

TI Cystic fibrosis in a patient with Kartagener syndrome.

SO Am-J-Dis-Child. 1974 May. 127(5). P 746-7.

MJ CYSTIC-FIBROSIS: co. KARTAGENER-TRIAD: co.

MN CASE-REPORT. CHLORIDES: an. HUMAN. INFANT. LUNG: ra. MALE.

SITUS-INVERSUS: co, ra. SODIUM: an. SWEAT: an.

AB A patient exhibited the features of both Kartagener syndrome and cystic fibrosis. At most, to the authors' knowledge, this represents the third such report of the combination. Cystic fibrosis should be excluded before a diagnosis of Kartagener syndrome is made.

RF 001 KARTAGENER M BEITR KLIN TUBERK 83 489 933

002 SCHWARZ V ARCH DIS CHILD 43 695 968

003 MACE JW CLIN PEDIATR 10 285 971

...

CT 1 BOCHKOVA DN GENETIKA (SOVIET GENETICS) 11 154 975

2 WOOD RE AM REV RESPIR DIS 113 833 976

3 MOSSBERG B MT SINAI J MED 44 837 977

...

Sample CF Queries

QN 00002

QU Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respiratory tract in CF?

NR 00007

RD 169 1000 434 1001 454 0100 498 1000 499 1000 592 0002 875 1011

QN 00004

QU What is the lipid composition of CF respiratory secretions?

NR 00009

RD 503 0001 538 0100 539 0100 540 0100 553 0001 604 2222 669 1010
711 2122 876 2222

NR: Number of Relevant documents

RD: Relevant Documents

Ratings code: Four 0-2 ratings, one from each expert