Performance Evaluation of Information Retrieval Systems

This material was prepared by Diana Inkpen, University of Ottawa, 2005, updated 2021. Some of the 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?

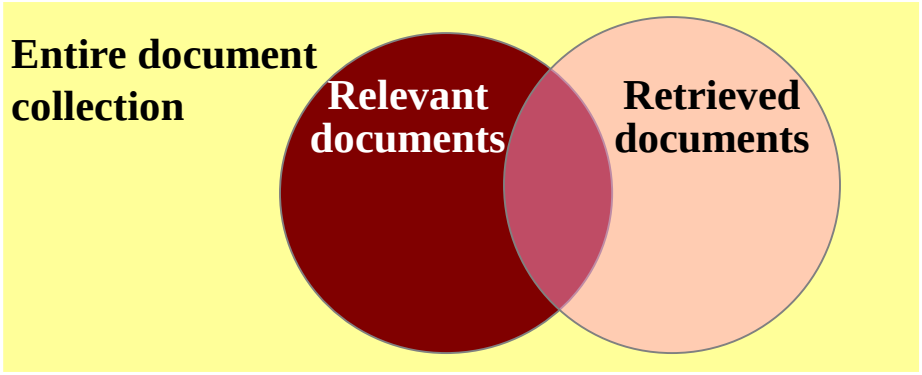
Difficulties in Evaluating IR Systems

- 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



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

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

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

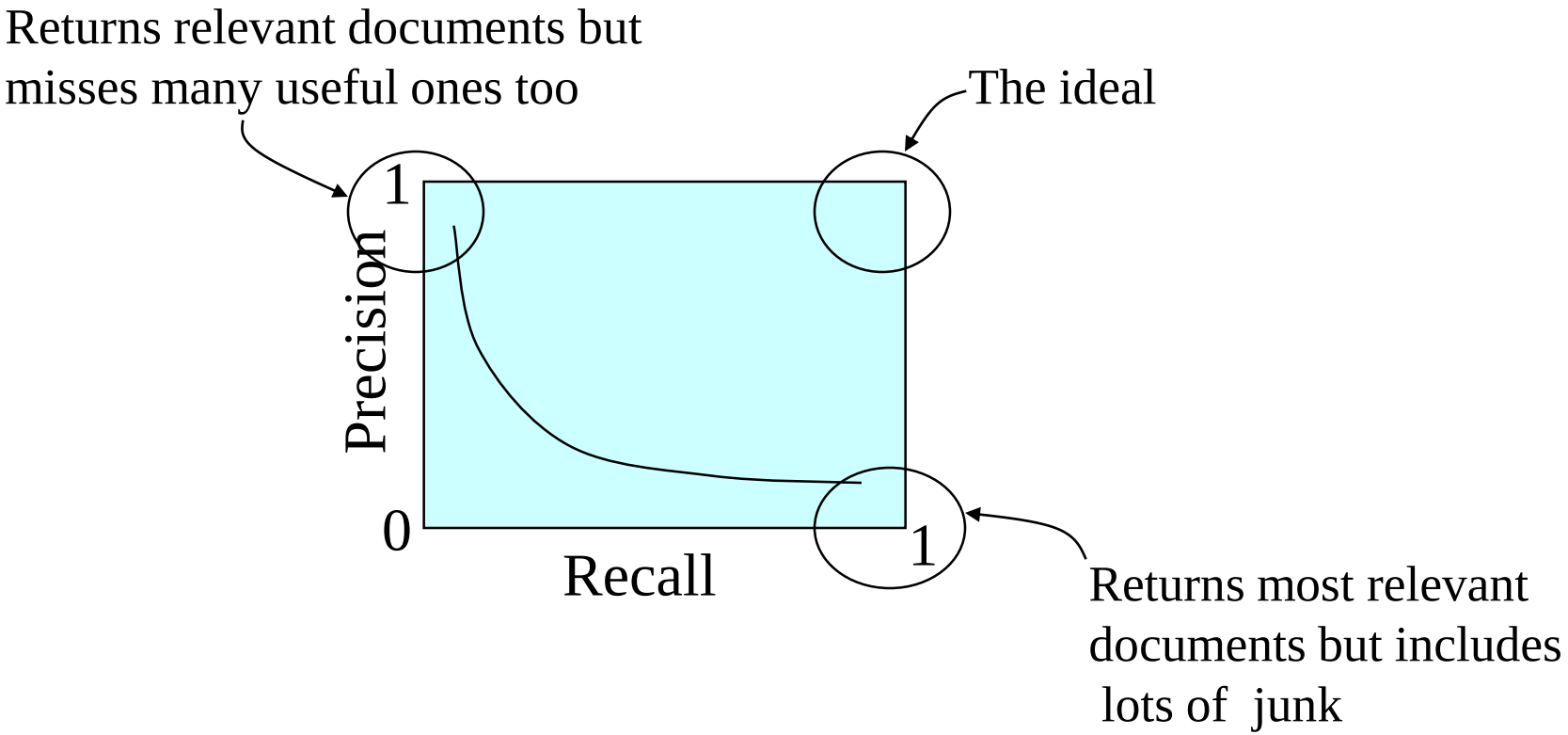
Precision and Recall

- 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



Computing Recall/Precision Points

- For a given query, produce the ranked list of retrievals.
- Adjusting a threshold on this ranked list produces different sets of retrieved documents, and therefore different recall/precision measures.
- Mark each document in the ranked list that is relevant according to the gold standard.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.

Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document. Never reach 100% recall

Average Precision

- AveP = $\frac{\sum_k P(k) * rel(k)}{number\ of\ relevant\ documents}$
- rel(k) is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

For the previous query:

AveP = (1+1+0.75+0.667+0.38)/6 = 0.632

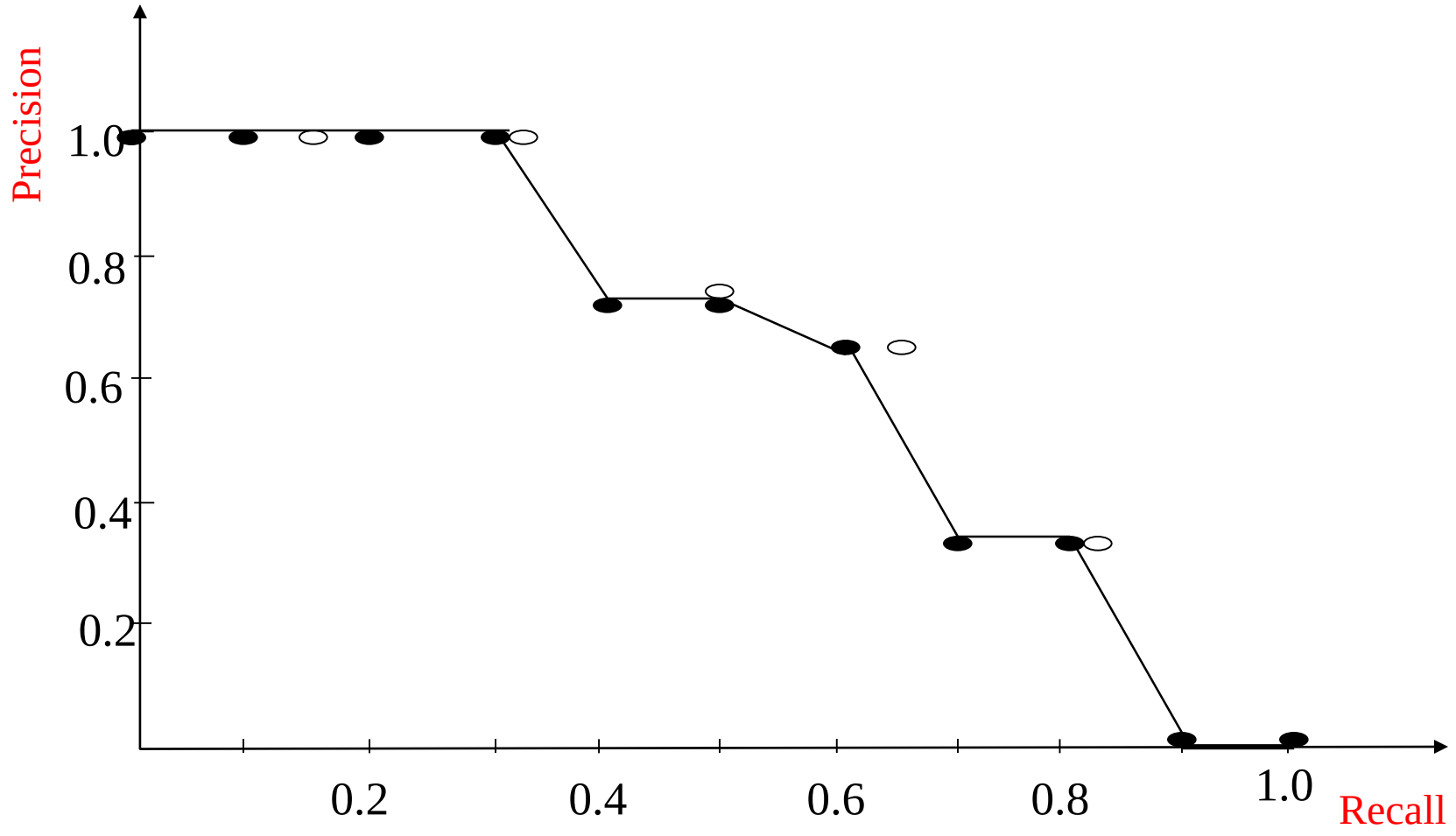
We need averages over all queries in the test set.

Interpolating a Recall/Precision Curve

- For a query, interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0, r_1 = 0.1, \ldots, r_{10}=1.0$
- The interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level between the *j*-th and (*j* + 1)-th level:

$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

Interpolating a Recall/Precision Curve: An Example

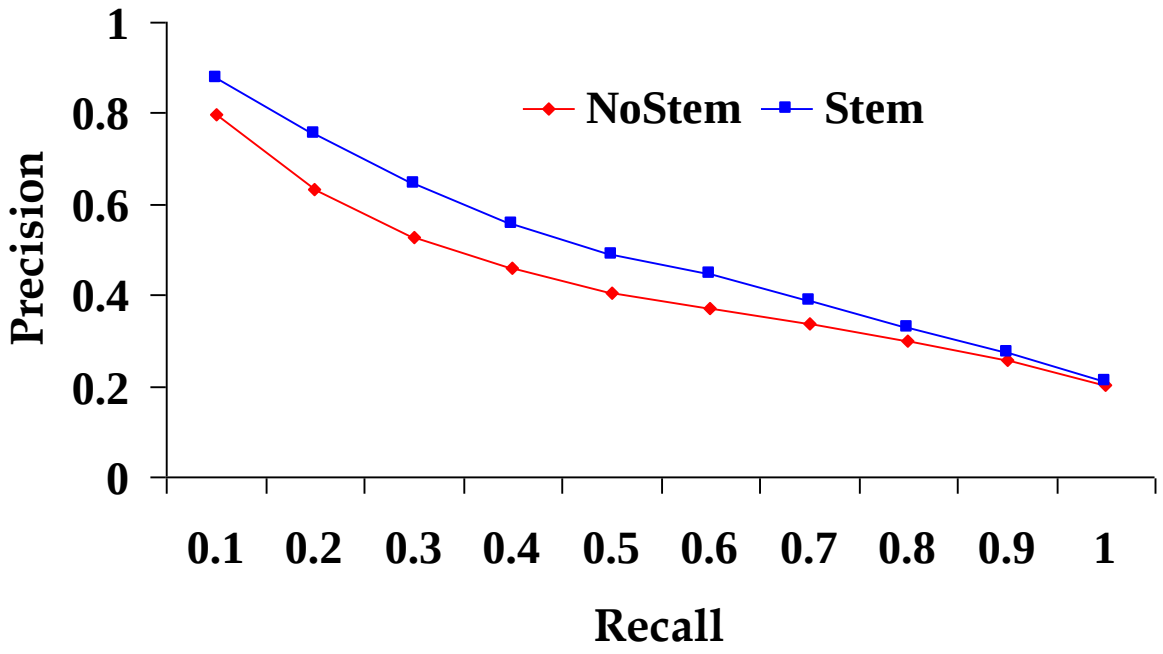


Average Recall/Precision Curve

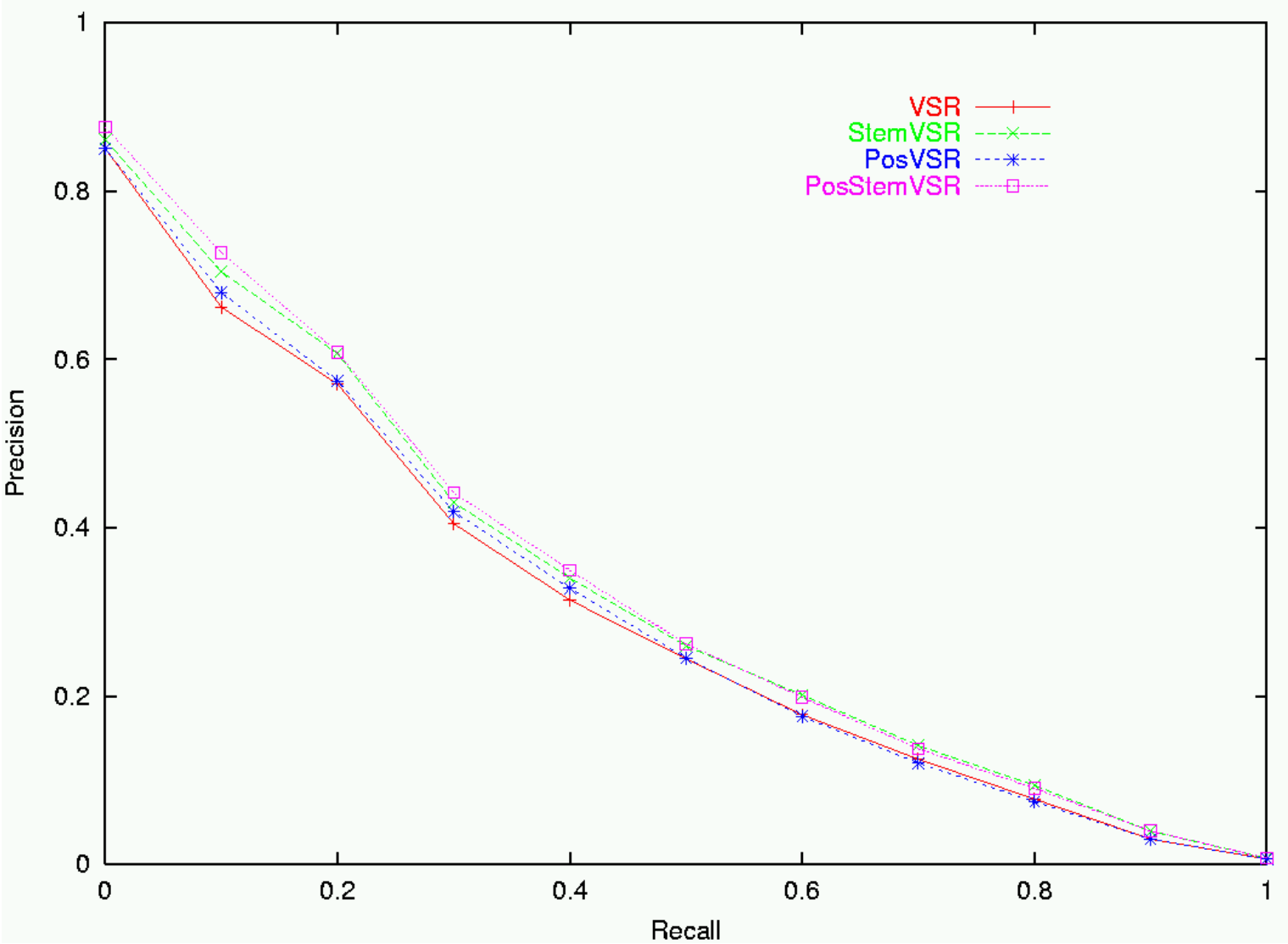
- Typically average performance over a large ***set*** of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

Compare Two or More Systems

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



Sample RP Curve for CF Corpus



Mean Average Precision (MAP score)

- Mean average precision for a set of Q queries is the mean of the average precision scores for each query (uninterpolated).
- $\text{MAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$

R- Precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

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.

E Measure (parameterized F Measure)

- A variant of F measure that allows weighting emphasis on precision over recall:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- Value of β controls trade-off:
  - β = 1: Equally weight precision and recall (E=F).
  - β < 1: Weight precision more.
  - β > 1: Weight recall more.

Fallout Rate

- Problems with both precision and recall:
  - Number of irrelevant documents in the collection is not taken into account.
  - Recall is undefined when there is no relevant document in the collection.
  - Precision is undefined when no document is retrieved.

$$Fallout = \frac{no.of\ \ nonrelevant\ items\ retrieved}{total\ no.of\ \ nonrelevant\ items\ in\ the\ collection}$$

Subjective Relevance Measure

- *Novelty Ratio*: The proportion of items retrieved and judged relevant by the user and of which they were previously unaware.
  - Ability to find ***new*** information on a topic.
- *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

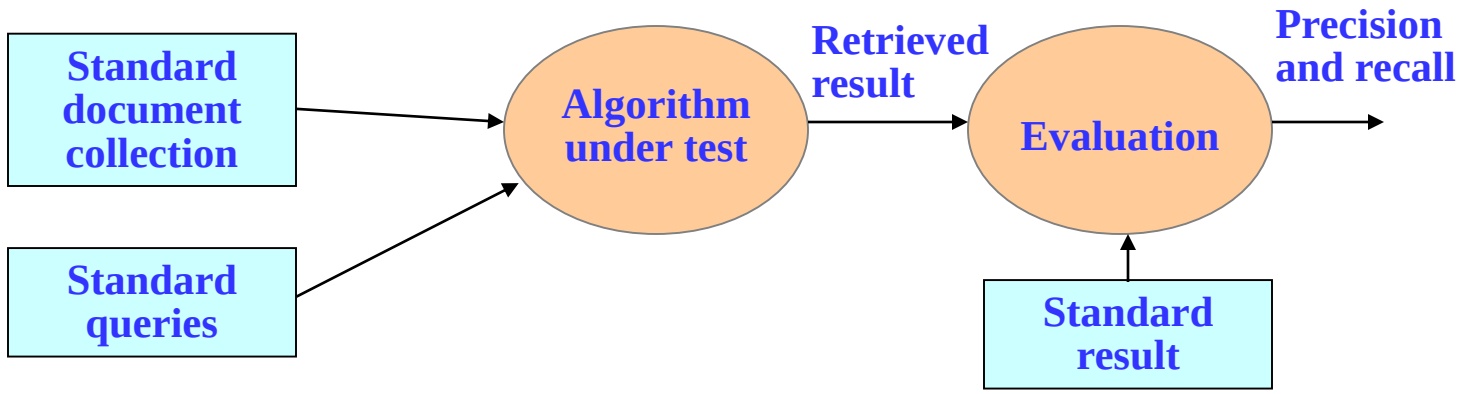
- *User effort*: Work required from the user in formulating queries, conducting the search, and screening the output.
- *Response time*: Time interval between receipt of a user query and the presentation of system responses.
- *Form of presentation*: Influence of search output format on the user's ability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

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/smart
  - 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 are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing.

- Participants submit the P/R values for the final document and query corpus and present their results at the conference.

The TREC Objectives

- Provide a common ground for comparing different IR techniques.
  - Same set of documents and queries, and same evaluation method.
- Sharing of resources and experiences in developing the benchmark.
  - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

TREC Advantages

- Large scale (compared to a few MB in the SMART Collection).
- Relevance judgments provided.
- Under continuous development with support from the U.S. Government.
- Wide participation:
  - TREC 1: 28 papers 360 pages.
  - TREC 4: 37 papers 560 pages.
  - TREC 7: 61 papers 600 pages.
  - TREC 8: 74 papers.

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

More Details on Document Collections

- Volume 1 (Mar 1994) - Wall Street Journal (1987, 1988, 1989), Federal Register (1989), Associated Press (1989), Department of Energy abstracts, and Information from the Computer Select disks (1989, 1990)
- Volume 2 (Mar 1994) - Wall Street Journal (1990, 1991, 1992), the Federal Register (1988), Associated Press (1988) and Information from the Computer Select disks (1989, 1990)
- Volume 3 (Mar 1994) - San Jose Mercury News (1991), the Associated Press (1990), U.S. Patents (1983-1991), and Information from the Computer Select disks (1991, 1992)
- Volume 4 (May 1996) - Financial Times Limited (1991, 1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), and the Federal Register (1994).
- Volume 5 (Apr 1997) - Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990).

TREC Disk 4,5

TREC Disk 4	Congressional Record of the 103rd Congress approx. 30,000 documents approx. 235 MB
	Federal Register (1994) approx. 55,000 documents approx. 395 MB
	Financial Times (1992-1994) approx. 210,000 documents approx. 565 MB
TREC Disk 5	Data provided from the Foreign Broadcast Information Service approx. 130,000 documents approx. 470 MB
	Los Angeles Times (randomly selected articles from 1989 & 1990) approx. 130,000 document approx. 475 MB

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>

TREC Properties

- Both documents and queries contain many different kinds of information (fields).
- Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system.
  - A system may be very good at querying and ranking, but if it generates poor queries from the topic, its final P/R would be poor.

Two more TREC Document Examples

ZIFF Communications Company

<DOC>
<DOCNO> ZF109-706-077 </DOCNO>
<DOCID>09 706 077.&O;</DOCID>
<JOURNAL>Business Week Dec 31 1990 n3194 p93(12).&M;
</JOURNAL>
<TITLE>Fujitsu means business for America. (Special Advertising Section by Fujitsu Ltd.) (includes related articles on the company's business relationships with Pepsi-Cola, Convex Computer, Greenville EMS, and Sequent Computer Systems)&M;
</TITLE>
<TEXT>
<ABSTRACT>In establishing itself as a major manufacturer in the computer hardware market, Fujitsu Ltd boasts a long list of corporate customers.&P; The company's client base includes: MCI Telecommunications Corp., Page Composition, Johns Hopkins Hospital, Tiara Computer Systems Inc., Pepsi-Cola, Convex Computer, Greenville EMS, and Sequent Computer Systems Inc. The company stresses its good customer relations and product development aspects, as well as its telecommunications products.&O;
</ABSTRACT>
</TEXT>
<DESCRIPT>
Company: Fujitsu Ltd. (Marketing).&O;
Topic: Marketing Strategy
Customer Relations
photograph.&M;
</DESCRIPT>
</DOC>

San Jose Mercury News

<DOC>
<DOCNO> SJMN91-06364024 </DOCNO>
<ACCESS> 06364024 </ACCESS>
<CAPTION> Photo; PHOTO: Associated Press; ANOTHER TURNOVER – Kansas City's Leonard
Griffin (98) closes in on Raiders quarterback Todd Marinovich, who fumbled on
the play. Marinovich also threw four interceptions.
</CAPTION>
<DESCRIPT> PROFESSIONAL; FOOTBALL; PLAYOFF; GAME; RESULT; BRIEF </DESCRIPT>
<LEADPARA> Too much excitement on top of too much cold medication may have caused the rapid heartbeat that forced Kansas City linebacker Derrick Thomas out of the
...
reliable place-kicker, kicked an 18-yard field goal at 10:26 of the fourth quarter, but he missed two field goals in the first half, from 33 and 47 yards.
...
</TEXT>
<FEATURE> PHOTO </FEATURE>
<STATE> CA </STATE>
<WORD.CT> 539 </WORD.CT>
<DATELINE> Sunday, December 29, 1991 00364024,SJ1
</DATELINE>
<COPYRGHT> Copyright 1991, San Jose Mercury News
</COPYRGHT>
<LANGUAGE> ENG </LANGUAGE>
</DOC>

Another Example of TREC Topic/Query

<top>
<head> Tipster Topic Description
<num> Number: 101
<dom> Domain: Science and Technology
<title> Topic: Design of the "Star Wars" Anti-missile Defense System
<desc> Description:
Document will provide information on the proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system.
<narr> Narrative:
proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system. The design and technology to be used in the anti-missile defense system advocated by the Reagan administration, the Strategic Defense Initiative (SDI), also known as "star wars." Changes of constituent technologies, are also relevant documents.
<con> Concept(s):
1. Strategic Defense Initiative, SDI, star wars, peace shield
2. kinetic energy weapon, kinetic kill, directed energy weapon, laser, particle beam, ERIS (exoatmospheric reentry-vehicle interceptor system), phased-array radar, microwave
3. anti-satellite (ASAT) weapon, spaced-based technology, strategic defense technologies
<fac> Factor(s):
<nat> Nationality: U.S.
</nat>
<def> Definition(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

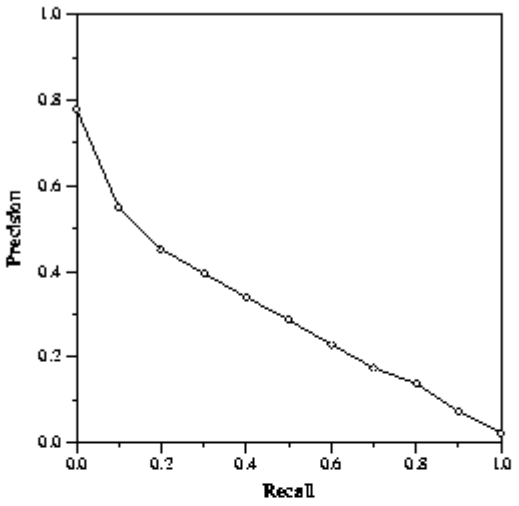
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

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

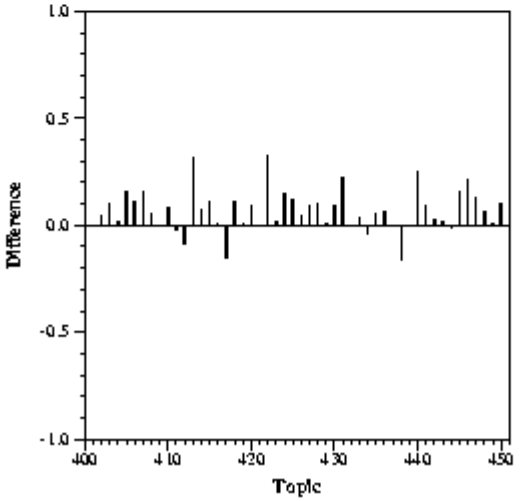
Average precision over all relevant docs	
non interpolated	0.2930

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



Recall-Precision Curve



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

Preprocessing for VSR Experiments

- Separate file for each document with just:
  - Author
  - Title
  - Major and Minor Topics
  - Abstract (Extract)
- Relevance judgment made binary by assuming that *all* documents rated 1 or 2 by *any* expert were relevant.