Performance Evaluation of Information Retrieval Systems

Why is System Evaluation Needed?

- There are many retrieval systems on the market, which one is the best?
- When the system is in operation, is the performance satisfactory? Does it deviate from the expectation?
- To fine tune a query to obtain the best result (for a particular set of documents and application)
- To provide inputs to cost-benefit analysis of an information system (e.g., time saving compared to a manual system)
- To determine the effects of changes made to an existing system (system A versus system B)

- Efficiency: speed
- Effectiveness: how good the result is?

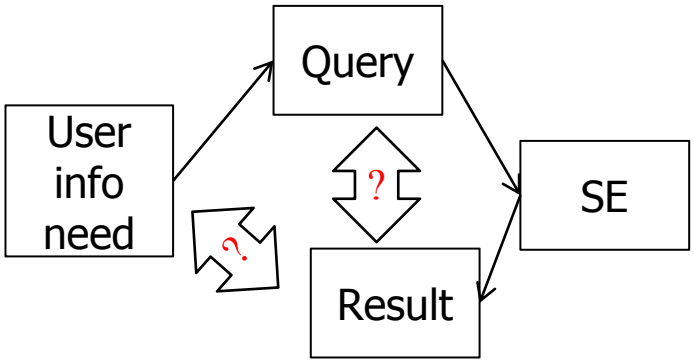
Too Kinds of Effectiveness Evaluation

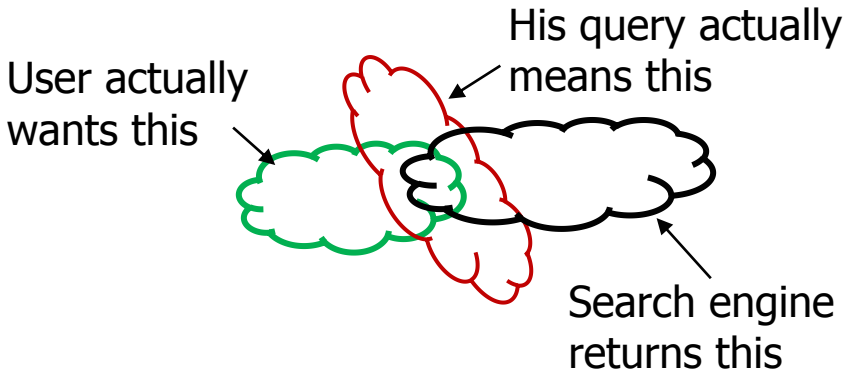
- Explicit evaluation
  - Performed by human judges: given a set of queries, the judges examine the pages and decide if the pages are relevant or non-relevant
  - Offline evaluation, artificial, rely on the judges and proper choice of queries and test data
  - Document-based: Which documents are relevant, not how happy I am
- Behaviorial evaluation
  - Examine real life user behavior logged by the search engine
    - Query -> ranked result set -> clicks + timestamp of each action
  - Assume: click behavior reveals user satisfaction
  - Online evaluation; reflects user satisfaction/experience that depends on many factors, e.g., snippets, UI, user's background, search context, that go beyond result relevancy
  - Metrics: Total number of clicks, average rank of user clicks (ARUC)

Explicit Evaluation

Are Human Judges best for Evaluation?

- Judges are not the persons who created the queries
  - They interpret a query according to their expert knowledge and make relevance judgement but short queries typically have broad scope (e.g., what does the user want for JavaScript?)
  - Judges do not have consistent judgement among themselves!
  - The query intent and information need only exist in the user's head
- Users typically identify fewer but high-quality results



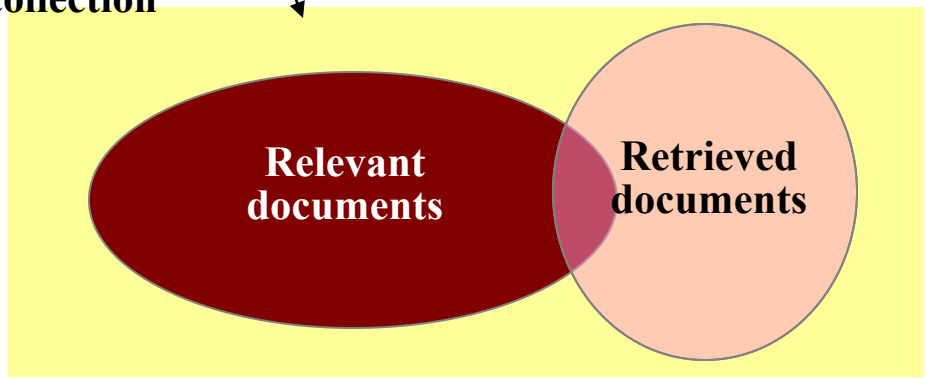


- Should we compare Result to Query or to Info Need?

Difficulties in Explicit Evaluation

- Effectiveness is based on ***relevancy*** of items retrieved
- Relevancy is not a binary evaluation but a continuous function
- Even when relevancy judgment is binary, it is difficult to make a judgment
- Relevancy, from a human judgment standpoint, is
  - subjective - depends upon a specific user's judgment
  - situational - relates to user's requirement
  - temporal - changes over time

Retrieval Effectiveness - Precision and Recall



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

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

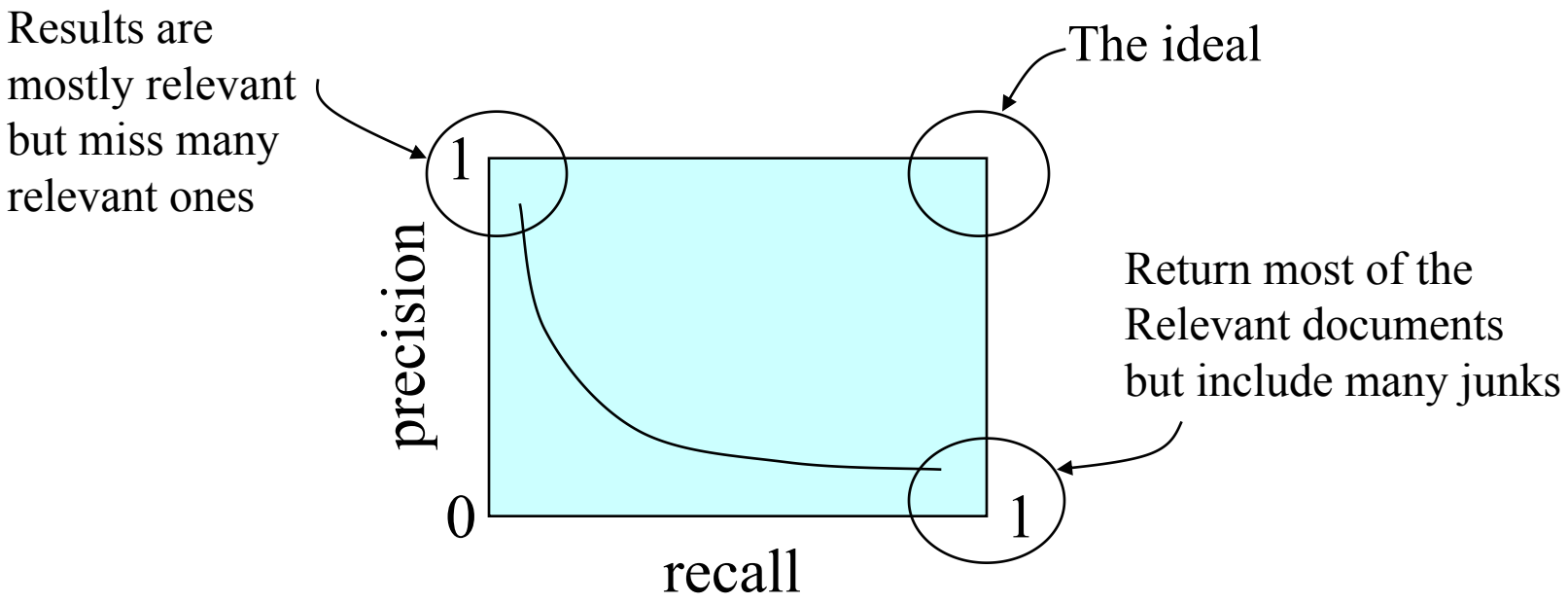
$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{total\ Number\ of\ documents\ retrieved}$$

Precision and Recall

- Precision
  - evaluates the correlation of the query to the database
  - an indirect measure of the completeness of indexing algorithm
- Recall
  - the ability of the search to find all of the relevant items in the database
- Among three numbers,
  - only two are always available
    - *total number of items retrieved*
    - *number of relevant items retrieved*
  - *total number of relevant items* is usually not available

Relationship between Recall and Precision

- Unfortunately, precision and recall affect each other in the opposite direction! Given a system:
  - Broadening a query or increasing the number of returned documents will increase recall but lower precision
  - Fine-tuning the weighing formula often makes one parameter better but the other worse



Fallout Rate

- Problems with precision and recall:
  - A query on "Hong Kong" will return most relevant documents but it doesn't tell you how good or how bad the system is! (What is the chance that a randomly picked document is relevant to the query?)
  - 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

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

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

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

- A good system should have high recall and low fallout

Fallout (cont)

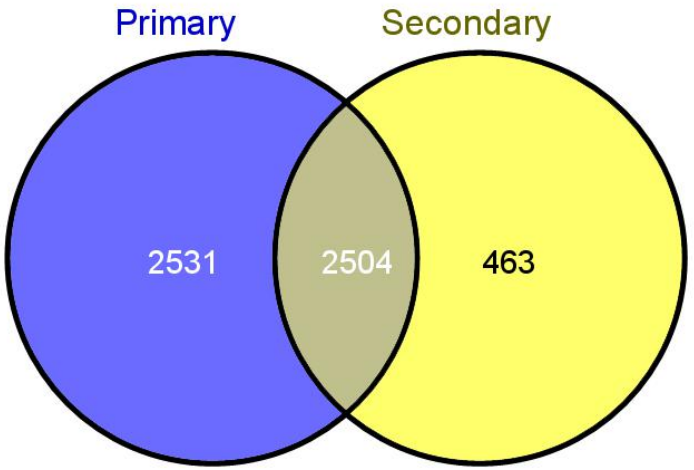
- Fallout can be viewed as the inverse of recall
- It is very unlikely to have situation as 0/0
  - the number of non-relevant items in a collection can be safely be assumed to be non-zero.
- It is the probability that a retrieved item is non-relevant. (Recall: the probability that a retrieved item is relevant)
- Among three measures, precision, recall and fallout, fallout is least sensitive to the accuracy of the search process because it reflects the overall relevance of the collection to the query
  - E.g., search "Hong Kong" in Hong Kong Government website
- A good system should have high recall and low fallout

How to obtain Relevance Judgment?

- Use human judges to evaluate a query against every document and decide if the document if relevant to the query or not
- While human evaluation appears to be ideal for obtaining relevance judgment, it is:
  - Very time consuming
  - Prone to human error and inconsistency

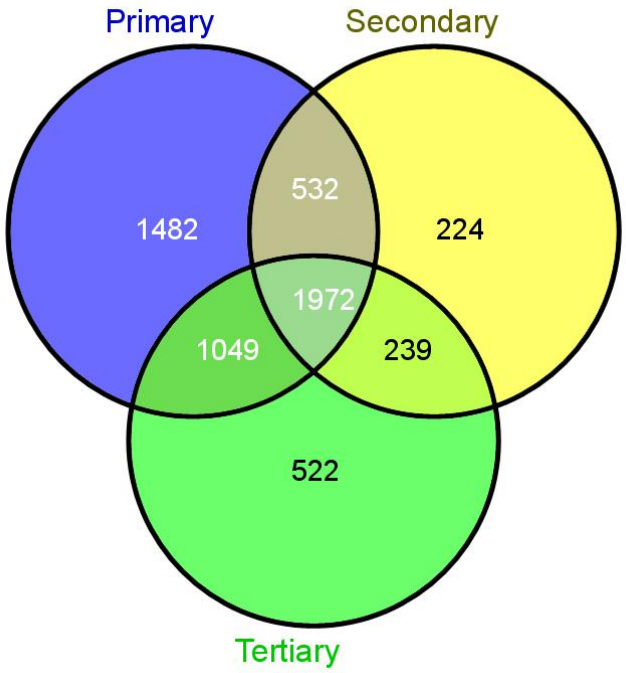
Human Judgments are Inconsistent

- Two assessors, primary and secondary, evaluated the same set of documents
- 2,504 documents are assessed as relevant by both assessors
- Totally, 2,531 + 2,504 + 463 = 5,498 documents are evaluated as relevant
- Overlap = 2,504 ／ 5,498 = 45.5%.



More Human Assessors Disagree Even More!

- Three assessors evaluated the same set of documents
- 1,972 documents are assessed as relevant by all assessors
- Totally, 1,482 + 532 + 224 + 1,972 + 1,049 + 239 + 522 = 6,020 documents are evaluated as relevant
- Overlap = 1,972 / 6,020 = 32.8%.



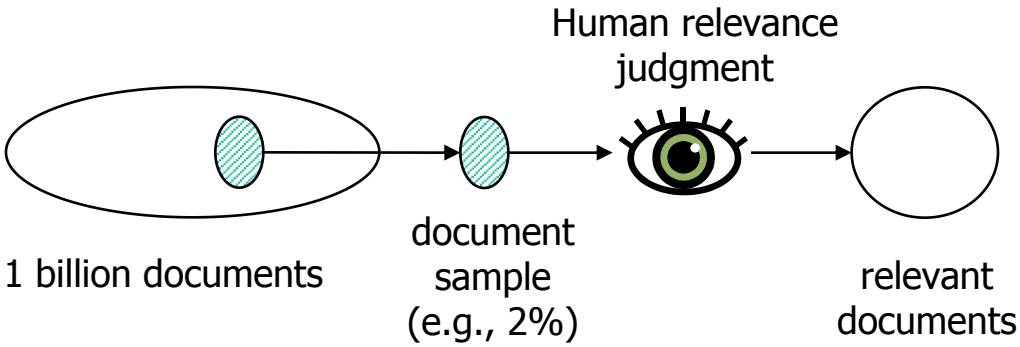
How to handle the Inconsistency?

- Identify one assessor, e.g., the primary assessor, as always correct; then, why need the secondary or tertiary assessors?
- Take the majority vote; consider only the intersection as relevant documents
- Ask an additional, authoritative judge to evaluate the inconsistent evaluations
- In summary, human evaluation is inaccurate and expensive!

- (Better) alternatives:
  - Use automatic methods to filter out relevant documents, then use human judgment on the filtered results, e.g., search engine pooling
  - Use the human judged results to train an automatic method, and iterative the process

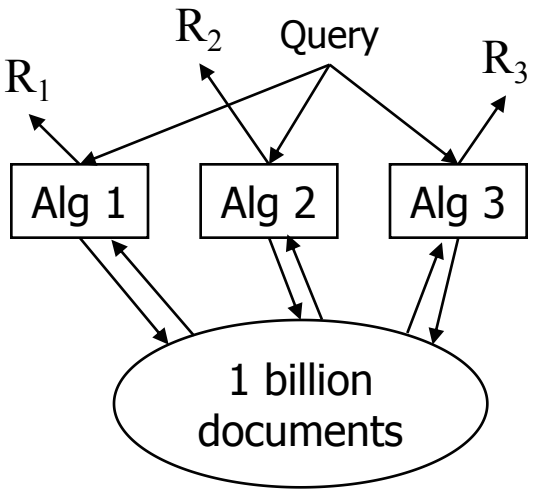
How to Find the Total Number of Relevant Items?

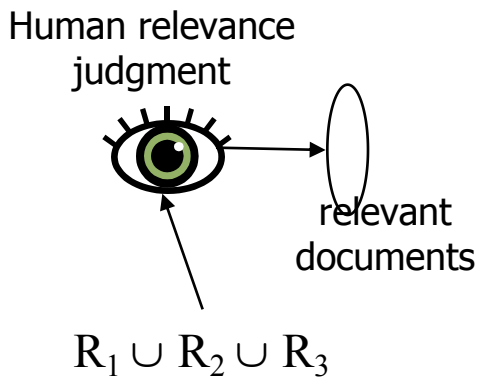
- In an infinitely large collection (e.g., the web), it is unknown.
- Note that we only need to know the number of relevant documents in a collection; no need to find the list of relevant documents
- Two possible approaches to get an estimate:
  1. Sampling the collection and perform relevance judgment on the samples



How to Find the Total Number of Relevant Items?

- Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total number of relevant documents in the collection (biased sampling)





Consideration of Ranking

- No far we have not considered ranking of results!

Computation of Recall and Precision

n	doc #	relevant	Recall	Precision
1	588	x	0.2	1.00
2	589	x	0.4	1.00
3	576		0.4	0.67
4	590	x	0.6	0.76
5	986		0.6	0.60
6	592	x	0.8	0.67
7	984		0.8	0.57
8	988		0.8	0.50
9	578		0.8	0.44
10	985		0.8	0.40
11	103		0.8	0.36
12	591		0.8	0.33
13	772	x	1.0	0.38
14	990		1.0	0.36

Suppose:
total no. of relevant docs = 5

R=1/5=0.2; p=1/1=1

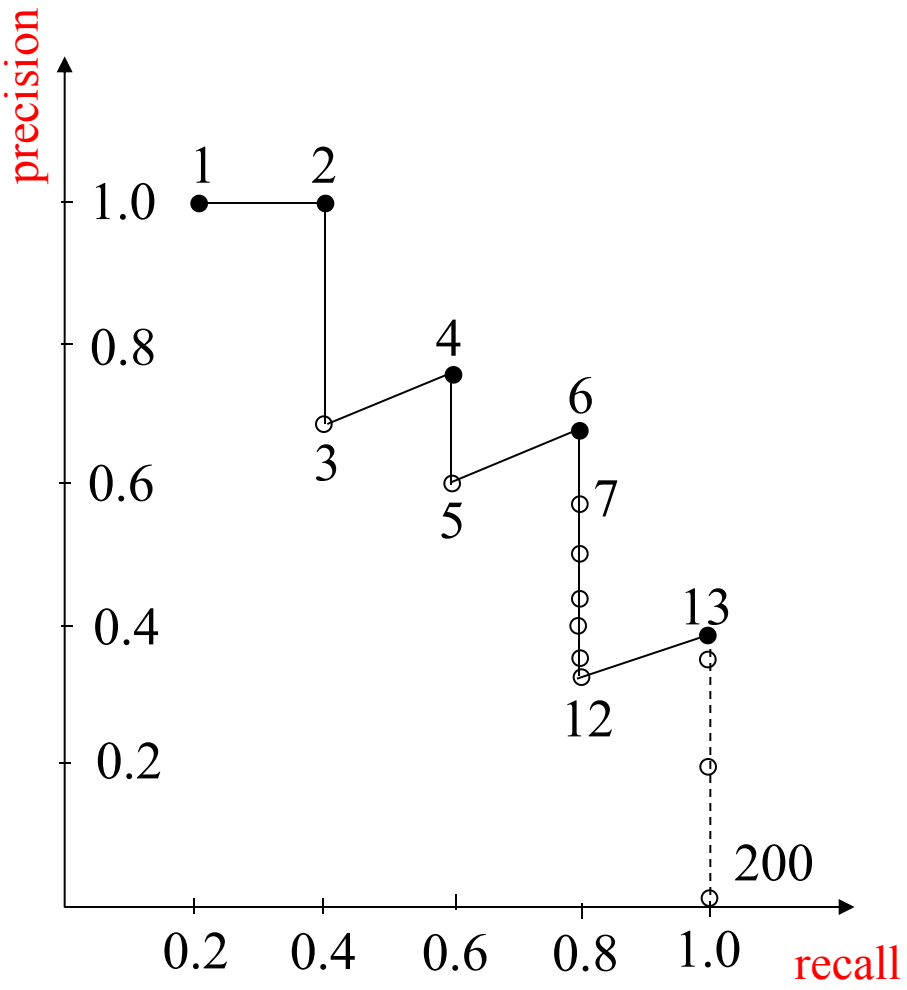
R=2/5=0.4; p=2/2=1

R=2/5=0.4; p=2/3=0.67

R=5/5=1; p=5/13=0.38

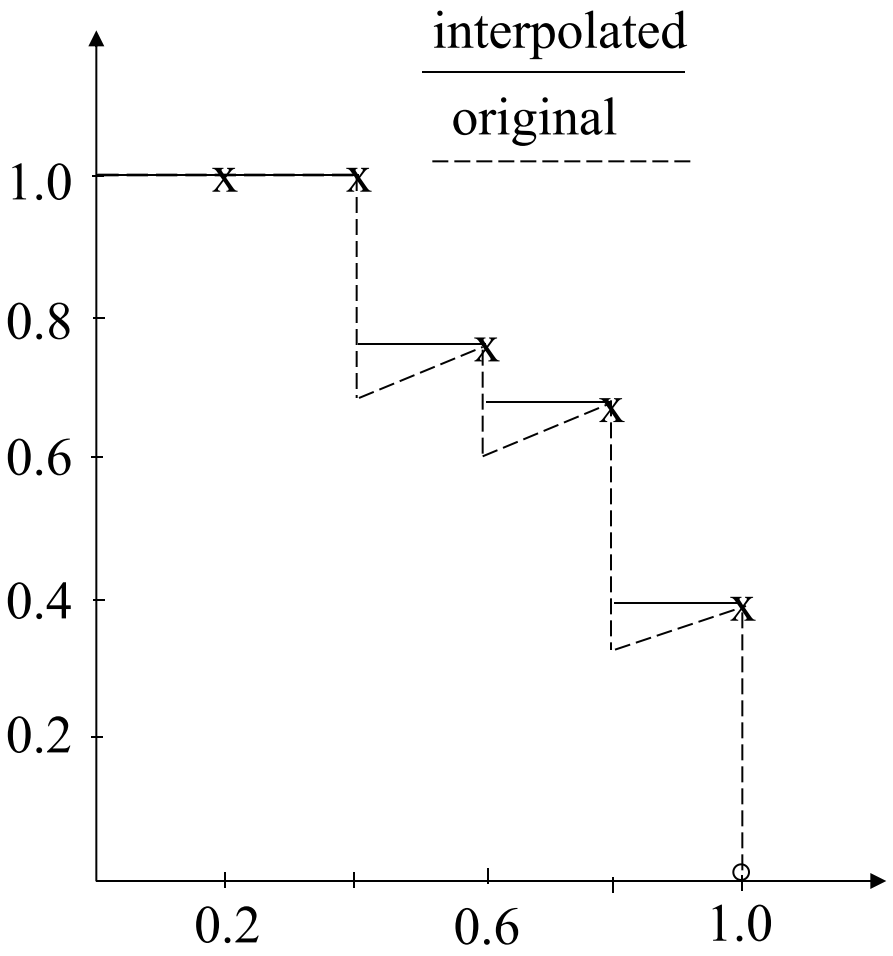
Computation of Recall and Precision

n	Recall	Precision
1	0.2	1.00
2	0.4	1.00
3	0.4	0.67
4	0.6	0.76
5	0.6	0.60
6	0.8	0.67
7	0.8	0.57
8	0.8	0.50
9	0.8	0.44
10	0.8	0.40
11	0.8	0.36
12	0.8	0.33
13	1.0	0.38
14	1.0	0.36



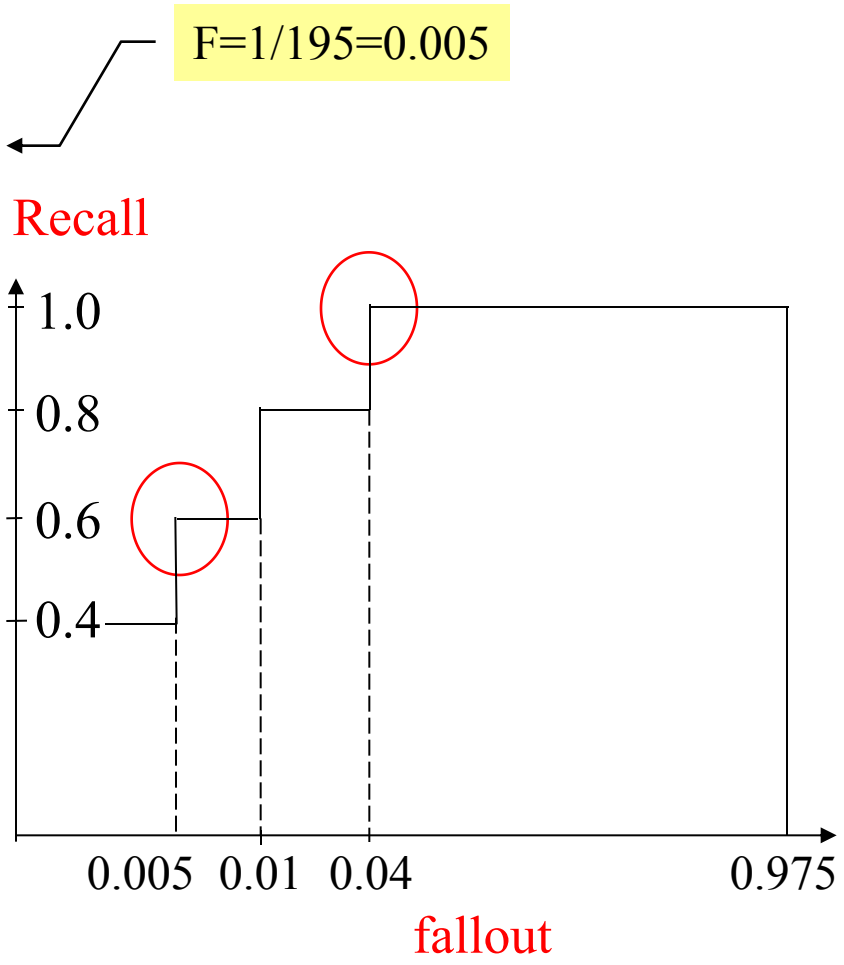
Interpolated Recall-Precision Graph

- For a certain recall level, precision could be missing
- Use interpolation: the best performance a user can achieve
  - The precision at recall level i is the highest precision from recall levels j>=i;
  - At 0.5 recall, the highest precision when recall>=0.5 is 0.76 at recall=0.6
  - Interpolation gives precision at recall=0
  - Precision is obtained for the "standard 11 recall levels", 0, 0.5, … 0.95, 1.0
- For more than one curve (query), average the (interpolated) precisions at each recall level



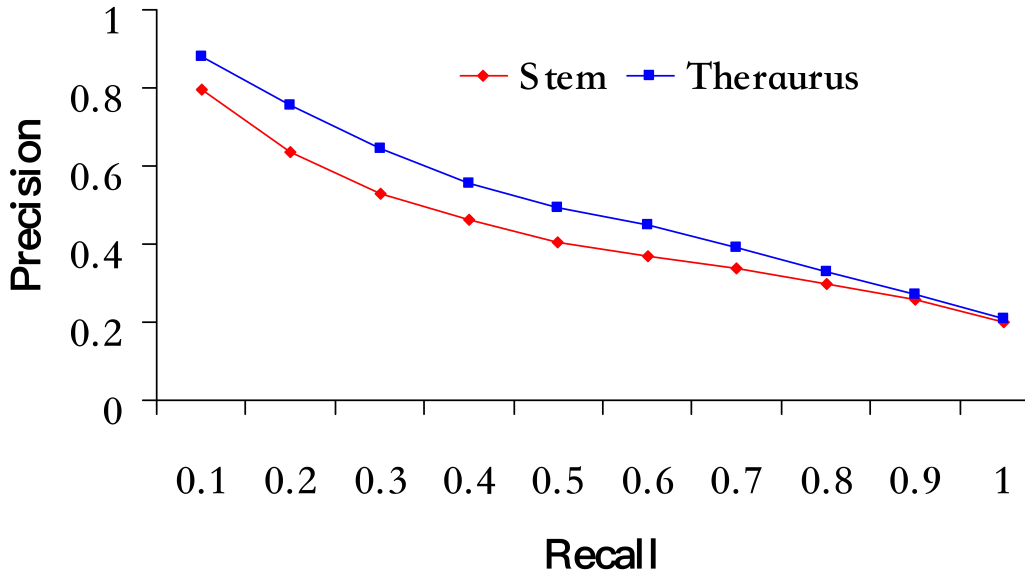
Recall-Fallout Graph

n	doc #	relevant	Recall	fallout
1	588	x	0.2	0.000
2	589	x	0.4	0.000
3	576		0.4	0.005
4	590	x	0.6	0.005
5	986		0.6	0.010
6	592	x	0.8	0.010
7	984		0.8	0.015
8	988		0.8	0.021
9	578		0.8	0.026
10	985		0.8	0.031
11	103		0.8	0.036
12	591		0.8	0.041
13	772	x	1.0	0.041
14	990		1.0	0.046
20			1.0	0.077
50			1.0	0.231
100			1.0	0.487
200			1.0	1.000



Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance
- When a system is better than the other system in one segment (e.g., at low recall) but worse in another segment (e.g., at high recall), is it better or worse than the other system?
- Are the ranks of the returned relevant documents considered in precision and recall computation?



- Tool to produce P/R graph:
  http://trecvid.nist.gov/trecvid.tools/trec_eval_video/README

More Performance Measures

Single-Valued Measures

- Desirable properties for an ideal effectiveness measure
  - Reflect retrieval effectiveness alone
  - Independent of any particular retrieval cutoff, e.g., the number of documents retrieved in a search
  - A single number if possible
- E-measure:
  - Large values of the recall (R) and precision (P) correspond to small values of E; when P = R = 1, E = 0
  - α = 1 means user is interested in P only

$$E = 1 - \frac{1}{\alpha(1/P) + (1-\alpha)(1/R)}$$

- F-measure (harmonic mean) or F_1: E-measure at α=0.5

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

$$F_\beta = \frac{(1+\beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

$F_\beta = 1 - E$ where $\alpha = 1 / (\beta^2 + 1)$.

Limitation of Precision and Recall

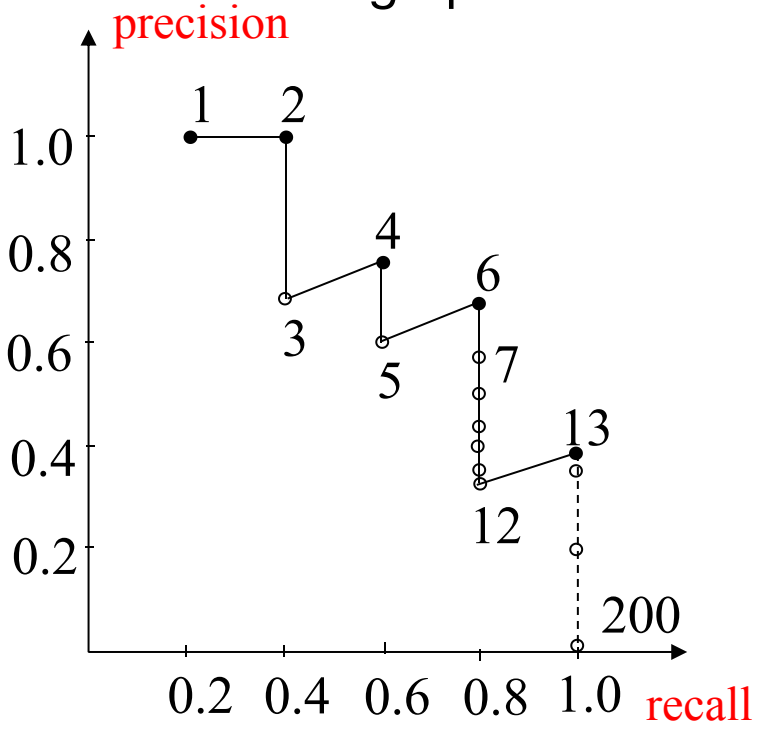
- Precision and recall consider ALL of the relevant document and ALL of the retrieved documents
  - They do not consider the ranks of the relevant documents
  - Ranking are important: a result ranked in the first position and one in the eleventh position make a lot of difference
- The way we compute precision and recall at every rank and plot a precision-recall curve solves the problem (partially)

Top-k Precision

- Precision at k or Top-k precision
  - Number of relevant documents / k
  - Notations:
    - Top-1 precision, Top-5 precision, Top-10 precision
    - precision@1, precision@5, precision@10, etc.

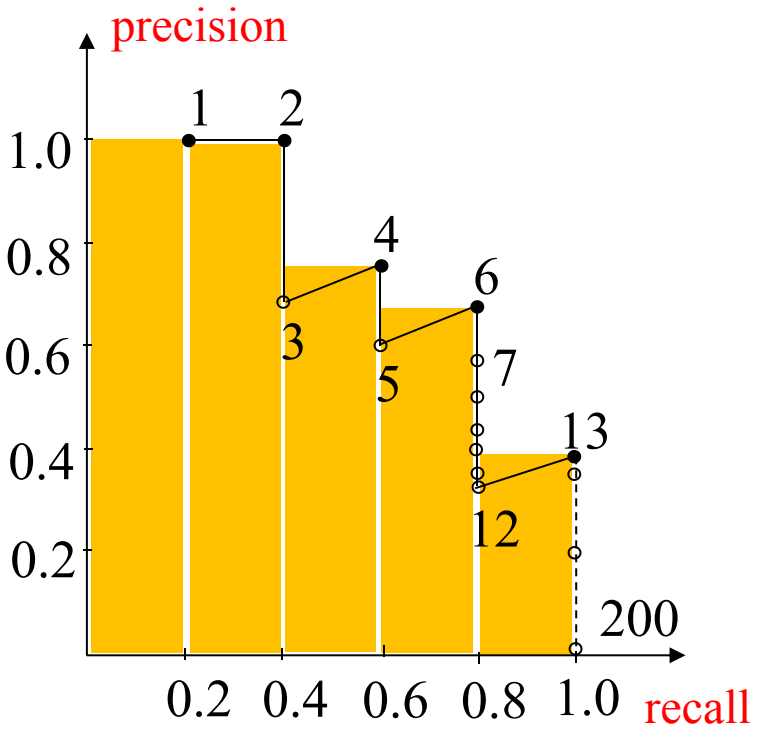
Average Precision (AP)

- Precision values are typically different at different recall level
- Average precision computes the area under the precision-recall curve
- The line graph has wiggles; interpolation smooths the graph because:
  - It is hard to compute the area under segments 3-4, 5-6 and 12-13; interpolation makes all segments horizontal or vertical
  - Queries have precisions at different recalls; the example has precisions at recall=0.2, 0.4, …, 1.0; others may have precisions at recall=0.33, 0.67, 1.0; interpolation makes all queries have precisions at fixed recall levels, commonly use is the 11-point precision values (i.e., 0, 0.1, .. 1.0)



Example for AP and AUC

k	Rel?	Recall R_k	Precision P_k
1	√	0.2	1.00
2	√	0.4	1.00
3		0.4	0.67
4	√	0.6	0.76
5		0.6	0.60
6	√	0.8	0.67
7		0.8	0.57
8		0.8	0.50
9		0.8	0.44
10		0.8	0.40
11		0.8	0.36
12		0.8	0.33
13	√	1.0	0.38
14		1.0	0.36



- Average Precision (AP) = $\Sigma_{k=1 \to 200} (P_k * \Delta R_k)$ *** $\Delta R_k = R_k - R_{k-1}$
  = 0.2*1 + 0.2*1 + 0.2*0.76 + 0.2*0.67 + 0.2*0.38 = 0.762
- Same as Area Under the P/R Curve, or AUC of the P/R curve

Mean Average Precision (MAP)

$$\text{Average Precison}(q) = \frac{\sum_{k=1}^{\text{Number of retrieved documents}} \text{Precision@k} \times \text{rel(k)}}{\text{Number of relevant documents}}$$

$rel(k) = 1$ when document at rank k is relevant

$rel(k) = 0$ when document at rank k is not relevant

rel(k) is a relevance indicator

It ignores precisions at non-relevant documents (D_2, D_3, D_6, D_8) when computing AP

Result(q):	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
rel(k)	1	0	0	1	1	0	1	0	1	1

- AP(q) = (1 + 2/4 + 3/5 + 4/7 + 5/9 + 6/10) / 6 = 0.638

Is this formula the same as in the preceding slide?

- Mean Average Precision (MAP)

$$MAP = \frac{\sum_{q=1}^{\text{Number of queries}} \text{Average Precision(q)}}{\text{Number of queries}}$$

Other Commonly Used Measures

- Discounted Cumulative Gain (DCG)
  - Relevance score is not binary: irrelevant (0), somewhat relevant (1), relevant (2), very relevant (3), etc.
  - Measure the usefulness (or gain) of a document at a certain rank
  - Documents that are highly relevant but have low rank should be penalized logarithmically proportional to k

- $DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$
  - rel_i: relevance score (e.g., 0..3) of the document at rank i
  - p: the rank at which DCG (or NDCG) is computer
- $nDCG_p = \frac{DCG_p}{IDCG_p}$
  - $IDCG_p$: Ideal DCG at rank p, i.e., results sorted by relevance score
  - $nDCG$ ranges from 0 to 1.0

Example on Computing NDCG [Wikipedia]

Documents	D_1	D_2	D_3	D_4	D_5	D_6
Relevance scores (rel_i)	3	2	3	0	1	2

- $CG_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$ (ref only; not used in NDCG)
- $DCG_6 = \sum_{i=1}^{6} \frac{rel_i}{\log_2 (i+1)}$

$= 3 + 2/\log_2 (3) + 3/\log_2 (4) + 1/\log_2 (6) + 2/\log_2 (7)$

$= 3 + 1.26 + 1.5 + 0 + 0.39 + 0.71 = 6.86$

Example on Computing NDCG (Cont.)

- Assuming no relevant documents after D_6, the ideal ranking of the results are obtained by sorting rel_i :

D_1	D_3	D_2	D_6	D_5	D_4
3	3	2	2	1	0

- $IDG_6 = \sum_{i=1}^{6} \frac{rel_i}{\log_2(i+1)}$

  $= 3 + 3/\log_2(3) + 2/\log_2(4) + 2/\log_2(5) + 1/\log_2(6)$

  $= 7.14$

- $NDG_6 = \frac{DCG_6}{IDG_6} = \frac{6.86}{7.14} = 0.96$

Subjective Relevance Measure

- *Novelty Ratio*: the proportion of items retrieved and judged relevant by the users of which they had not been aware prior to receiving the search output
  - usefulness of the results
- *Coverage Ratio*: the proportion of relevant items retrieved out of the total relevant documents ***known*** to users prior to the search
  - a very essential requirement: e.g., I want to locate a document which I have seen before (e.g., the budget report for Year 2000)
- *Sought recall*: the total number of documents examined by the user following a search, divided by the total number of relevant documents which the user would like to examine.
  - Precision and user interface of the system: e.g., consider the web, how many pages do I have to click into before I am satisfied with the search

Other Factors to Consider

- *user effort*: intellectual or physical, required from the users in formulating the queries, conducting the search, and screening the output
  - user interface, query language, etc.
- *response time*: the time interval between receipt of a user query and the presentation of system responses
  - tradeoff between response time and retrieval effectiveness
- *form of presentation* of the search output which influence the users' ability to utilize the retrieved materials
  - user interface
- *collection coverage*: the extent to which all relevant items are included in the system
  - quality of the collection; the best retrieval algorithm will return junk on collections that contain only junk! (important for web-based search engines)

Benchmarking

Experimental Setup for Benchmarking

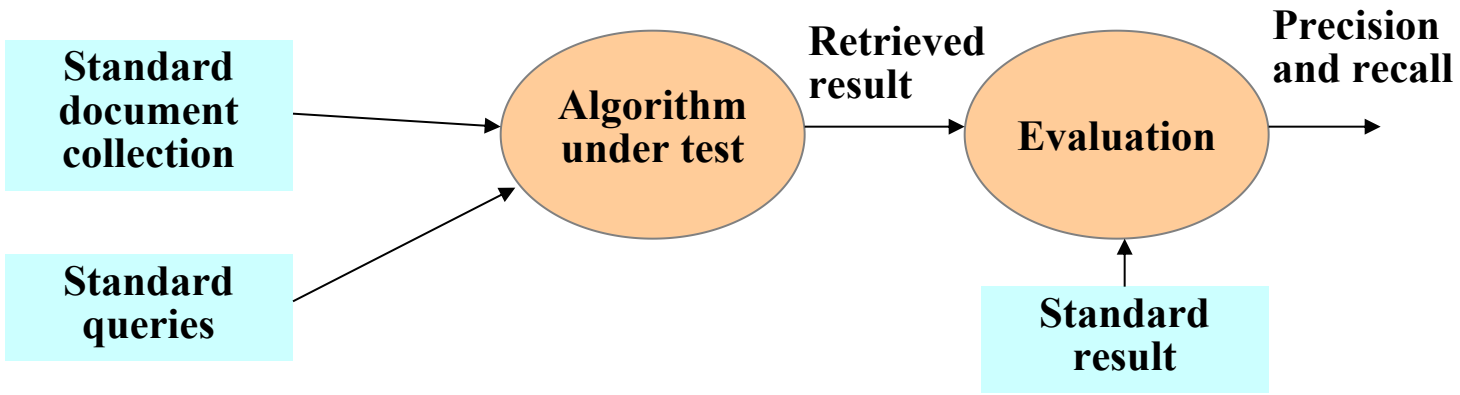
- It is very difficult to obtain ***analytical*** performance (of retrieval effectiveness) for document retrieval systems, because many characteristics of the documents such as relevance, distribution of words, etc., are difficult to describe with mathematical formula.

- Performance is measured by ***benchmarking***. That is, the retrieval effectiveness of a system is evaluated on a *given set of documents*, *queries*, and *relevance judgment*. This is analogous to benchmarking of computing systems.

- Performance data is valid only for the environment under which the system is evaluated.

Benchmarking – The Ideal

- A benchmark collection contains:
  - A set of standard documents and queries
  - 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
- Extremely resource consuming (large document collection)
- No good benchmark for the web
- No good benchmark for Asian languages
- May expose your system's weaknesses to your customers!

TREC and Earlier Test Collections

Early Test Collections

- Previous experiments were based on the SMART collection which is very 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
- 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 queries in different stages for testing and training
- Participants submit the P/R values on the final document and query set and present their results in the conference

http://trec.nist.gov/

The TREC Objectives

- Give 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

TREC Advantages

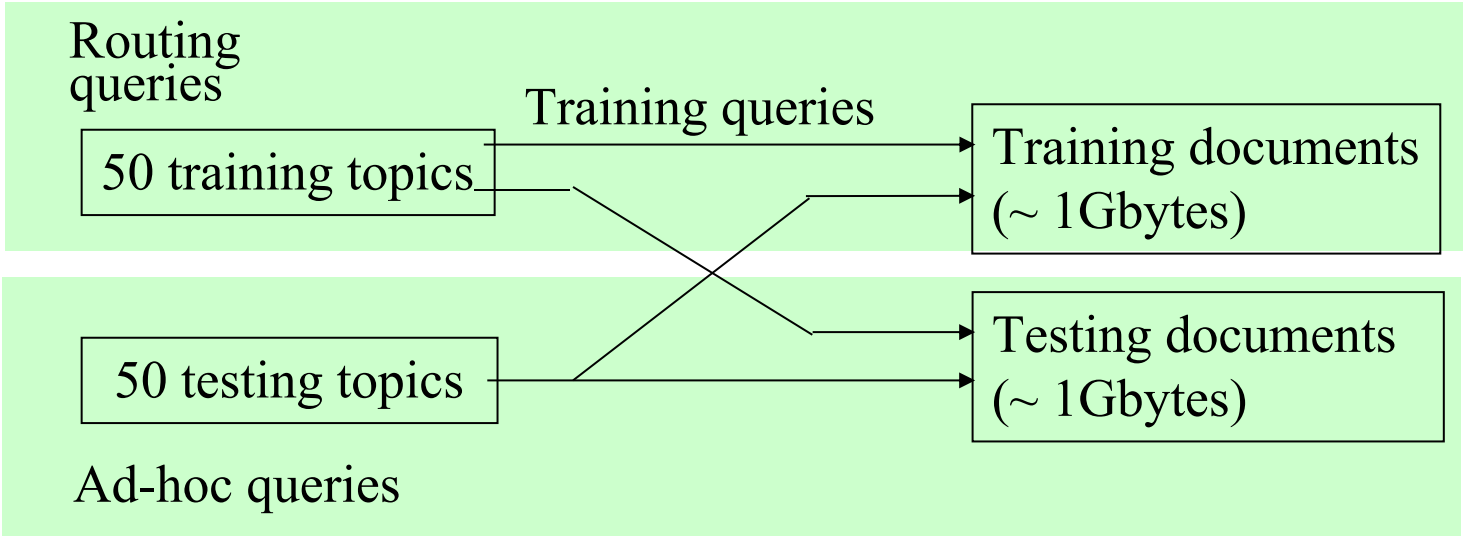
- Large scale (compared to a few Mbytes in the Cornell Collection)
- Relevance judgments provided
- Under continuous development and with support from US Government
- Wide participation: TREC 1 (’92): 28 papers 360 pages; TREC 4: 37 papers 560 pages; TREC 7: 61 papers 600 pages; TREC 8: 74 papers; TREC 9: 64 papers; TREC 10 (’01): 79 papers; TREC 2004: 97 papers; TREC 2005: 125 papers; TREC 2006: 112 papers … …

TREC Tasks

- Ad hoc - new questions are being asked on the static set of data. (library catalog searching)
- Routing - same questions are being asked, but the 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 Gbytes of 7.5 million documents)

TREC Collections

- TREC evaluates both Ad hoc and routing queries and provides both training and test collections:
  - 50 training topics + 1 Gbytes of training documents + relevance judgement
  - 50 test topics + 1 Gbytes of test documents
  - Adhoc: runs the 50 test topics on ***both*** training and test documents
  - Routing: trains the system with training queries and documents and then tests it on the unseen test documents.



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 -- Marked Up in 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 &amp;amp; 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 -- Marked Up 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, font …
<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definitions(s):
</top>

A Difficult Topic Example

<num> Number: 351

<title> Falkland petroleum exploration

<desc> Description:
What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

<narr> Narrative:
Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.

Sample Query -- Marked Up with SGML

- 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 in querying and ranking, but if it generates poor queries from the topic, its final P/R would be poor

Relevance Judgment

- Exhaustive evaluation:
  - 100 topics * 742611 documents = over 74 million judgements
- Sampling:
  - a topic has on average 200 and maximum 900 relevant documents, the sample size is still too large
- Polling:
  - Combine the retrieved documents from each system under test and perform relevance judgment on the combined documents
  - 33 runs of 200 top documents: 2398 documents per topic
  - 22 runs of 100 top documents: 1932 documents per topic (only 450 documents less than the larger combination)

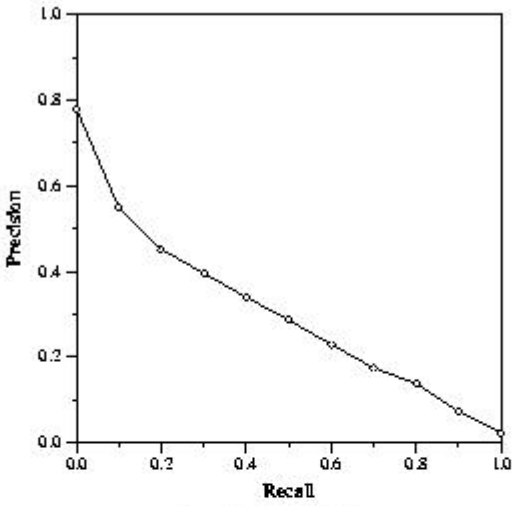
Evaluation

- 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 0.5 increment)
- Document level average: average precision when 5, 10, .., 100, … 1000 documents are retrieved
- Average precision histogram: average precision for each topic against the medium precision of all systems for that topic

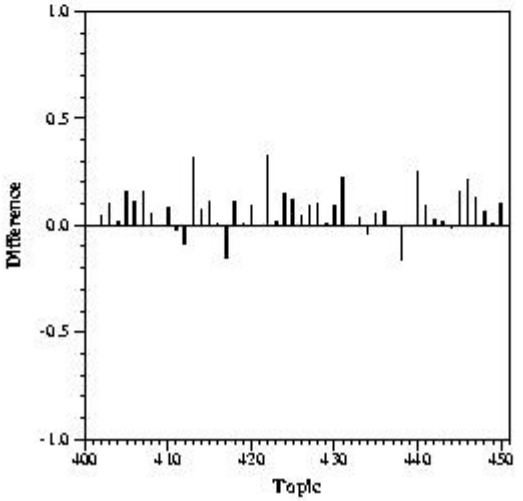
Summary Statistics	
Run Number	Flab8atd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

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

Summary

- Evaluation is essential
- Precision and Recall (and Fallout) are still the main objective measurements although relevance judgment is rather difficult and a number of other metrics of effectiveness have been proposed
- The Text Retrieval Conferences provide a source of a large database of documents, search statements and expected results from searches essential to evaluating algorithms
- Acceptable retrieval systems must be ease to use, reliable, effective and inexpensive.