Performance Evaluation

Issues in IR

Motivation

- One can use several models, e.g., boolean or vector, different indexing structures, different user-interfaces, etc.
- Which combination is the best one ?
- What is the measuring criteria ?
- The DBMS community is usually concerned with quality as it relates to time (query time, update time, availability time, etc).
- The IR community is **also** concerned with quality as it relates to usefulness of the answer, i. e., whether it fulfills the information needs of the user.

Motivation

- Given any query, an IR system will return a <u>set</u> of documents as the answer
- Among the returned documents some will be relevant and some (hopefully not many) will be irrelevant
- Given a query I and its <u>relevant</u> set *R* and the (returned) <u>answer</u> set *A*, let $|R|$ and $|A|$ denote the cardinality of these sets. Further, let *D* denote the set of all docs

Recall

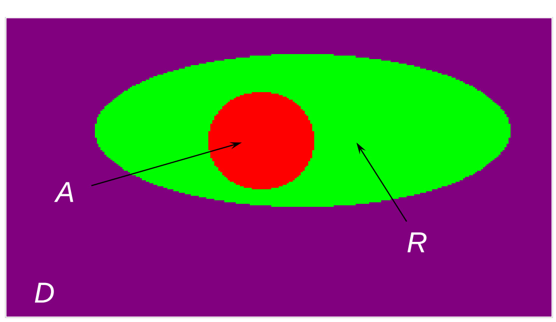
- Recall is the fraction of the relevant documents which were retrieved:
  - Recall = $|R \cap A| / |R|$
  - $0 \leq$ Recall ≤ 1
- Do we want 100% Recall ?
- If we get 100% Recall does it mean our search was very successful ?

Precision

- Precision is the fraction of retrieved documents which were relevant:
  - Precision = $|R \cap A| / |A|$
  - $0 \leq$ Precision ≤ 1
- Do we want 100% Precision ?
- How are Precision and Recall related ?

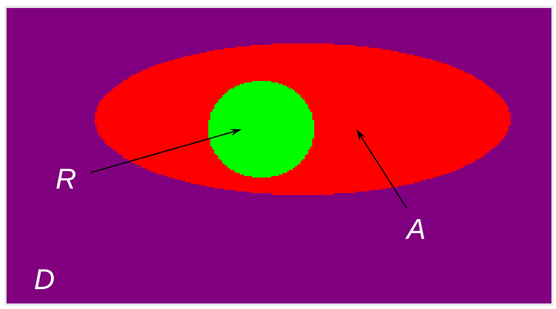
Recall and Precision ?

- $0 \leq$ Recall ≤ 1 and
- Precision = 1



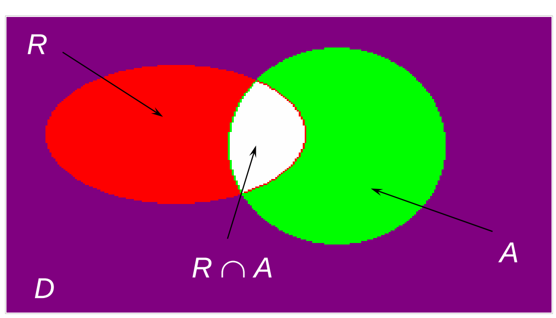
Recall and Precision ?

- Recall = 1 and
- 0 ≤ Precision ≤ 1



Recall and Precision ?

- 0 ≤ Recall ≤ 1 and
- 0 ≤ Precision ≤ 1



Recall and Precision as a Measure

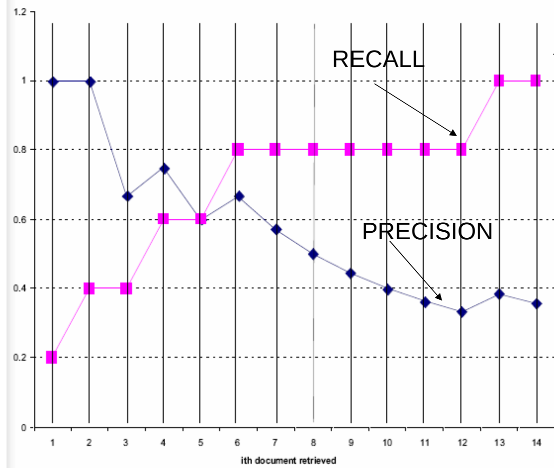
- Consider a query for which the relevant set is *R* = { d1, d2, d3, d4, d5} out of 10 docs
- Let us assume that a given IR system returned *A* = { d3, d43, d1, d4 }
- Recall = 3/5 = 60% and Precision = 3/4 = 75%
- How do we visualize this relationship between Recall and Precision considering ranking ?

Recall and Precision as a Measure

R = { d1, d2, d3, d4, d5}
A = { d3, d43 d1, d4 }

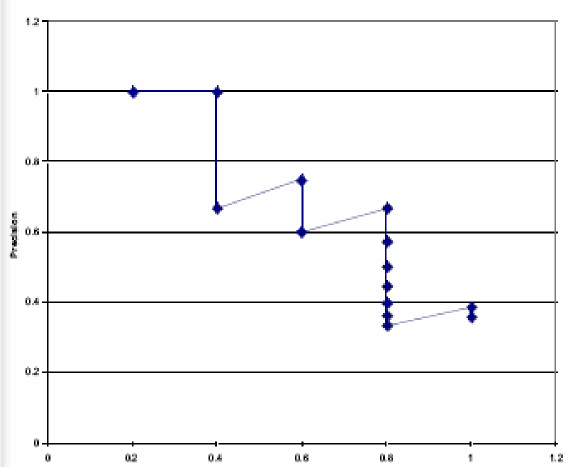
- {d3} yields 100% Precision at 20% Recall
- {d3, d43} yields 50% Precision at 20% Recall, (yes, two precision values are possible for a single recall value)
- {d3, d43, d1} yields 66% Precision at 40% Recall (yes, precision can go up and down)
- {d3, d43, d1, d4} yields 75% Precision at 60% Recall

Precision and Recall Graphs



	R/notR	Pre	Rec
1	1	1	0.2
2	1	1	0.4
3	0	0.666666667	0.4
4	1	0.75	0.6
5	0	0.6	0.6
6	1	0.666666667	0.8
7	0	0.571428571	0.8
8	0	0.5	0.8
9	0	0.444444444	0.8
10	0	0.4	0.8
11	0	0.363636364	0.8
12	0	0.333333333	0.8
13	1	0.384615385	1
14	0	0.357142857	1

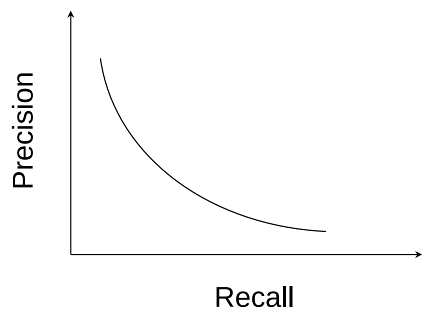
Precision-Recall Graph



	R/notR	Pre	Rec
1	1	1	0.2
2	1	1	0.4
3	0	0.666666667	0.4
4	1	0.75	0.6
5	0	0.6	0.6
6	1	0.666666667	0.8
7	0	0.571428571	0.8
8	0	0.5	0.8
9	0	0.444444444	0.8
10	0	0.4	0.8
11	0	0.363636364	0.8
12	0	0.333333333	0.8
13	1	0.384615385	1
14	0	0.357142857	1

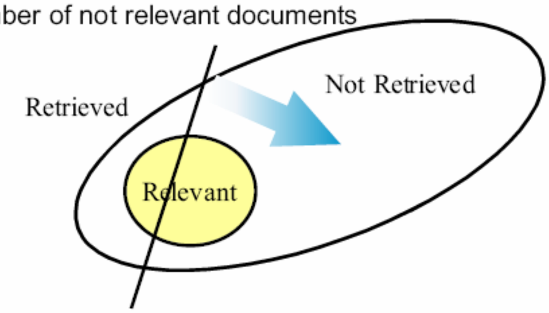
Recall and Precision Relationship

- Usually the relationship between Recall and Precision turns out to be shaped like this:



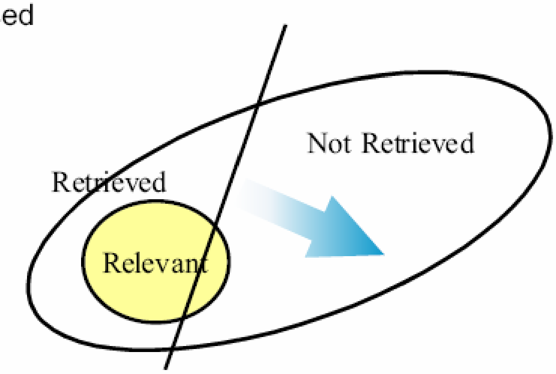
What happen when we increase the number of documents retrieved?

- At **low retrieval volumes** when we increase the number of documents retrieved , the number of relevant documents increase more rapidly than the number of not relevant documents



What happen when we increase the number c documents retrieved?

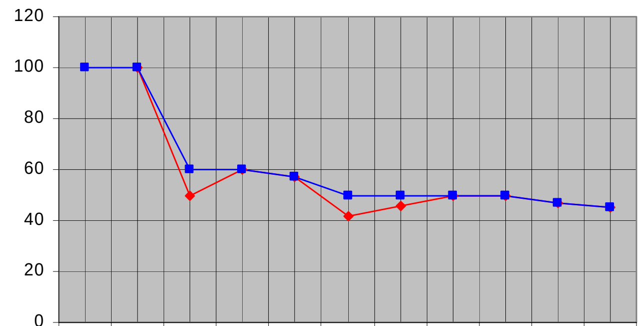
- At **high retrieval volumes** when we increase the number of document retrieved the situation is reversed



Interpolated P x R

- Different queries may yields different points of recall, thus making an average computation complicated
- Usual procedure: use of 11 standard points of recall: 0%, 10%, 20%, …, 100%
- The precision at any point is the higher precision value at any later recall value. This can be cascaded and guarantees the curve is non-increasing

Interpolated P x R



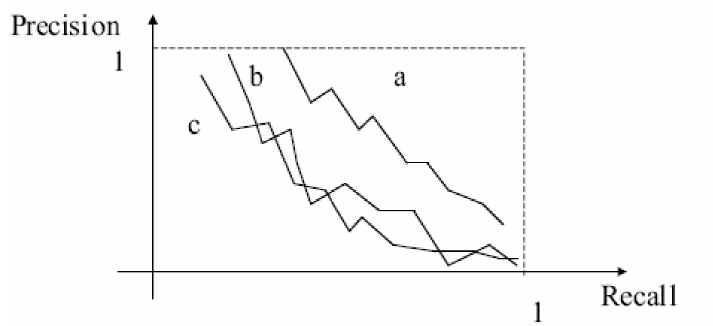
	0	10	20	30	40	50	60	70	80	90	100
Precision		100	50	60	57	42	46	50	50	47	45
Interpolated Precision	100	100	60	60	57	50	50	50	50	47	45

Comparing P x R curves

- What if one wants to compare different P x R curves, i.e., precision values at the same values of recall for different approaches ?
- Consider R = {d1, d2, d3, d4, d5} and
  - A1 = {d3, d5, d1, d4, d2, …[5 docs]}
  - A2 = {[5 docs]…, d1, d2, d3, d4, d5}
  - A3 = {d3, d5, d43}
- Which one is best ?
  - Both A1 and A2 have 50% precision at 100% recall
  - A3 has 66% precision at 40% recall (better ?)
  - Try sketching P x R curves

Precision-Recall Graph

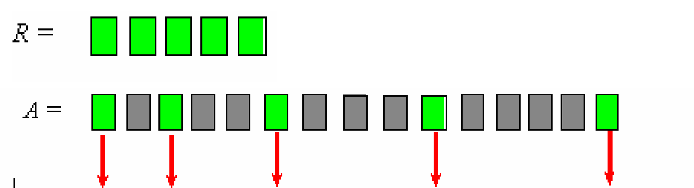
- The system has the best performance but what about system b and c, which one is the best?



Single Value Measures

- Average precision at seen relevant docs: Compute the precision every time a relevant doc is found and report the overall average
  - Few low-ranked docs shouldn't affect performance too much if most relevant docs are retrieved early
  - It is an "optimistic" measure
- R-precision
  - The precision of the lowest ranked relevant doc
  - Unlike the previous case, this is a "pessimistic" measure

Average precision at seen relevant docs



$$p = \frac{|Ra|}{|A_{seen}|}$$

$\frac{1}{1} = 1$ $\frac{2}{3} = 0.66$ $\frac{3}{6} = 0.5$ $\frac{4}{10} = 0.4$ $\frac{5}{15} = 0.33$

Average precision at seen relevant docs $\frac{1+0.66+0.5+0.4+0.33}{5} = 0.57$

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.
- Introduced by van Rijbergen, 1979
- Harmonic mean of recall and precision:

$$F_j = \frac{2P_jR_j}{P_j + R_j} = \frac{2}{\frac{1}{R_j} + \frac{1}{P_j}}$$

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

Single Value Measures

- Given the j-th doc in the ranking, its recall rj and precision pj, van Rijsbergen (see his book on-line) proposed the following measure:
  - $E_j = 1 - (1 + b^2)/(b^2/r_j + 1/p_j)$
  - b is a parameter set by the user
- If b = 1, $E_j = 1 - 2/(1/r_j + 1/p_j)$
  - docs with high precision and high recall have a low E value, whereas docs with low precision and low recall have a high E value (yes, sort of counter-intuitive)

Single Value Measures

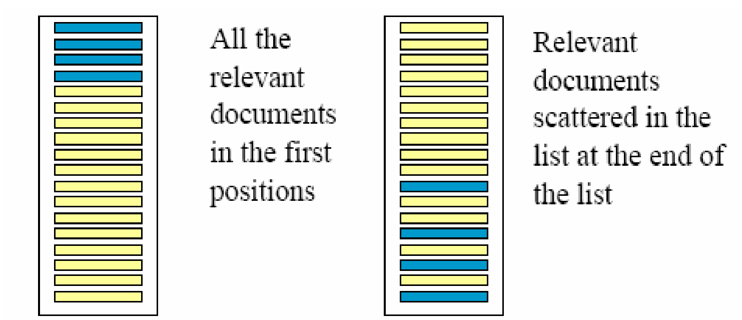
- $E_j = 1 - (1 + b^2)/(b^2/r_j + 1/p_j)$
  - If b > 1, then the emphasis would be on precision, conversely
  - if b < 1, then the user would be more interested in recall
- The main aspect of the measure E is that it evaluates each ranked document, not the whole document set, thus "anomalies" can be seen

Measuring Performance w/ Ranks

- Thus far we were not concerned explicitly with the rank (position) of the relevant docs
- Ideally the i^{th} relevant doc should be ranked i, yielding recall = $i/|R|$
- Unfortunately, such good behavior is not the typical case
- While it is true that the i^{th} relevant doc will yield recall = $i/|R|$, its rank will (usually) be $RANK_i > i$
- Can we use this to assess performance ?

Measuring performance w/ Ranks

- Two systems can give a very different perception if they just organize the same documents in a different way:



Measuring Performance w/ Ranks

- Normalized Recall:
  - NRecall = 1- $\Sigma_{i = 1,\ldots, |R|}$ ($RANK_i$ - i) / |R|(|*D*| - |*R*|)
  - Note that (|D| - |R|) is a normalizing factor and the “1 -” is only to make 1 the best case and 0 the worst case and not vice-versa.
  - Notice:
    - NRecall=1 when $RANK_i$=i (ideal case: all relevant documents first in the ranking)
    - NRecall=0 when $RANK_i$=(|*D*| - |*R*|)+i (worst case: all irrelevant documents first in the ranking)

Average Recall/Precision Curve

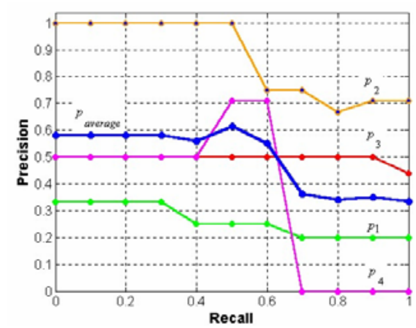
$$\overline{P}(r)=\sum_{i=1}^{N_q}\frac{P_i(r)}{N_q}$$

$\overline{P}(r)$ is the average Precision at Recall level r

N_q is the number of queries

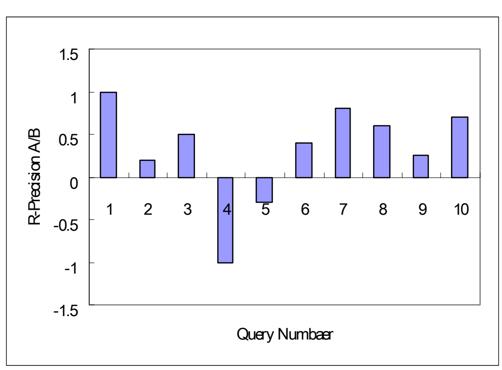
$P_i(r)$ is the Precision at Recall level r for the i-th query

- 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.



Precision Histograms

- Use R-precision measures to compare the retrieval history of two algorithms through visual inspection
- *RPA/B(i)=RPA(i)-RPB(i)*



Document Cutoff Levels

- Another way to evaluate:
  - Fix the number of relevant documents retrieved at several levels:
    - top 5
    - top 10
    - top 20
    - top 50
    - top 100
    - top 500
  - Measure precision at each of these levels
- This is a way to focus on how well the system ranks the first k relevant documents.

Fallout Measure

- Recall = $|R \cap A| / |R|$
  - What is the recall when there are no relevant docs to be retrieved ?
- Precision = $|R \cap A| / |A|$
  - What is the precision if no docs are retrieved ?
- Both recall and precision are concerned with retrieved relevant docs
- <u>Fallout</u> is concerned with retrieved but non-relevant docs
  - F = |A - R| / |D - R|

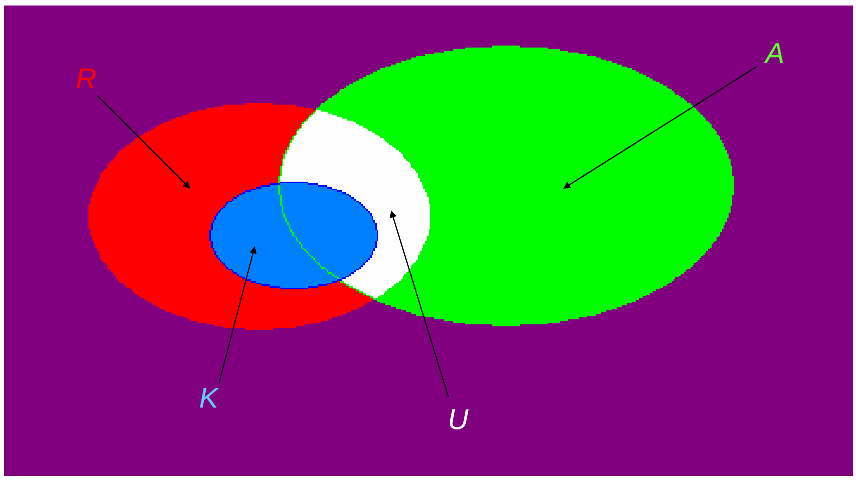
F x R vs P x R

- Typically |*D* - *R*| >> |*R*| thus Fallout varies less than Precision as a function of recall
- P x R is non-increasing whereas F x R is non-decreasing
- P x R is user-oriented
  - P helps to measure how well the system found good docs. Users are interested in usefulness of what they obtain
- F x R is systems-oriented
  - F helps to measure how well the system rejected bad docs. Implementors are interested in the robustness of their systems

User-oriented Measure of Performance

- It is also important to take into account what the (different) users feel about the answer sets
- Users may consider the same answer set of different usefulness, this is specially true if they know (in different degrees) the answers they "should" obtain
- In addition to *R* and *A* let us also consider the following subsets of *R*:
  - *K*: set of answers which are known to the user and,
  - *U*: set of answers which were not known by the user and were retrieved

User-oriented Measure of Performance



User-oriented Measure of Performance

- C = |*A* ∩ *K*|/|*K*| is the <u>coverage</u> of the answer
  - A high coverage ratio means that the system is finding most of what the user was expecting
- N = |*U*|/(|*K*| + |*U*|) is the <u>novelty</u> of the answer
  - A high novelty ratio means indicates that the user is finding many new docs which were not known and are relevant

- This is useful (?) in the context of investigating whether a new/improved systems is actually improving the search for end-users (actual use is non-trivial though)

Summary Table Statistics

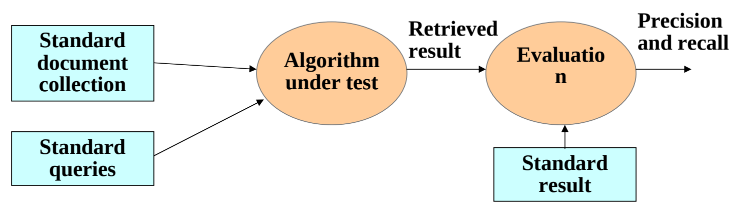
- the number of queries ,
- total number of documents retrieved by all queries,
- total number of relevant documents were effectively retrieved when all queries are considered
- total number of relevant documents could have been retrieved by all queries…

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/



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

- Most collections available from http://www.sigir.org

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>

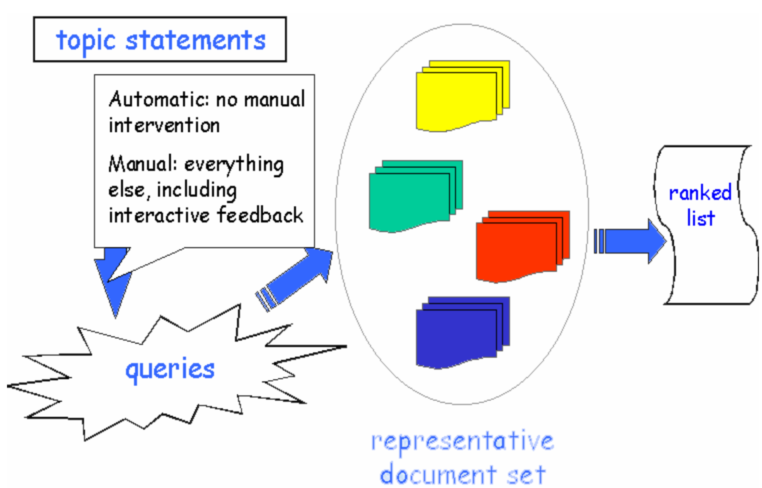
Sample Query (with SGML)

<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 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 and ranked. (news clipping, library profiling).
- Secondary tasks added after TREC 4:
  - Chinese: documents and topics in Chinese
  - Filtering: routing with no ranking
  - Interactive: evaluation of interactive systems
  - Natural Language Processing
  - Cross Language: documents and topics in different language
  - High precision: retrieval of ten documents answering a given information request within five minutes
  - Spoken document retrieval: retrieval techniques of spoken documents
  - Very large corpus: retrieval from collections of size 20 gigabytes

Creating a test collection for an ad hoc task



Obtaining Relevance Judgments

- Exhaustive assessment can be too expensive
  - TREC has 50 topics for >2 million docs each year
- Random sampling won't work either
  - If relevant docs are rare, none may be found!
- IR systems can help focus the sample
  - Each system finds some relevant documents
- Different systems find different relevant documents
  - Together, enough systems will find most of them

Pooled Assessment Method

- Each system submits top 100 documents
- All are placed in a single pool
- Duplicates are eliminated
- Placed in an arbitrary order to avoid bias
- Evaluated by the person that wrote the topic
- Assume un-evaluated documents not relevant

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.