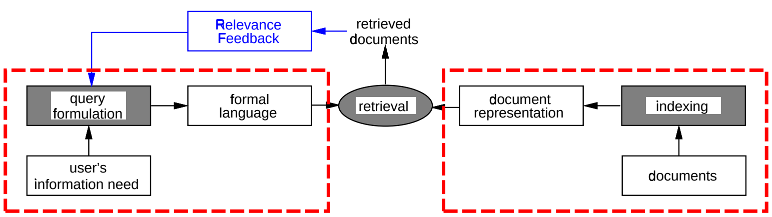
Contents

1 Automatic Indexing

1.1 Document Retrieval Model



1.2 Indexing Methods

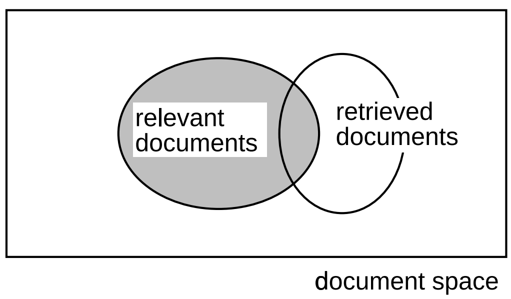
- Objective (indexing database attributes) vs Non-objective (indexing contents)
- Manual indexing vs Automatic indexing
- Controlled vocabulary (consistency) vs Uncontrolled vocabulary
- Single term indexing vs Complex term indexing
- Specificity vs Exhaustivity

1.3 Problems with Manual Indexing

- high labor cost of trained indexers
- inconsistency in selecting index terms and judging relevance.
  - thesauri created by two indexers in a given subject domain have only 60% of index terms in common
  - indexes obtained by two indexers from the same document with the same thesaurus have only 30% in common
  - documents obtained from two persons searching the same database with the same question have only 40% in common
  - relevance judgements obtained by two users on the same set of documents and the same topic have only 60% in common.

2 Performance Evaluation

2.1 Measures of Effectiveness – Precision and Recall



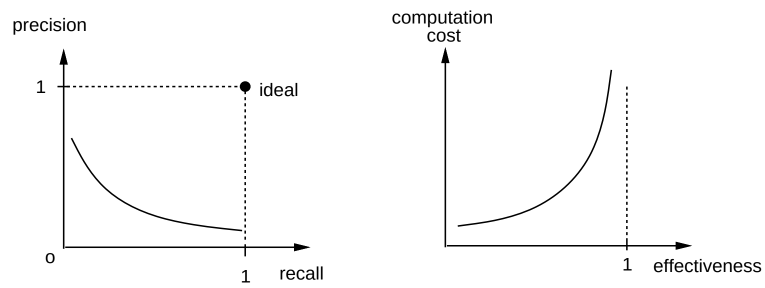
$$\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{total Number of relevant documents}}$$

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

2.2 Fallout Rate

- Problems with precision and recall:
  - recall is undefined when there is no relevant document in the collection
  - precision is undefined when no document is retrieved
  - number of irrelevant documents in the collection is not taken into account
- Fallout $= \frac{\text{number of nonrelevant items retrieved}}{\text{total number of nonrelevant items in the collection}}$
- A good system should have high recall and low fallout.

2.3 Tradeoffs between Cost and Effectiveness

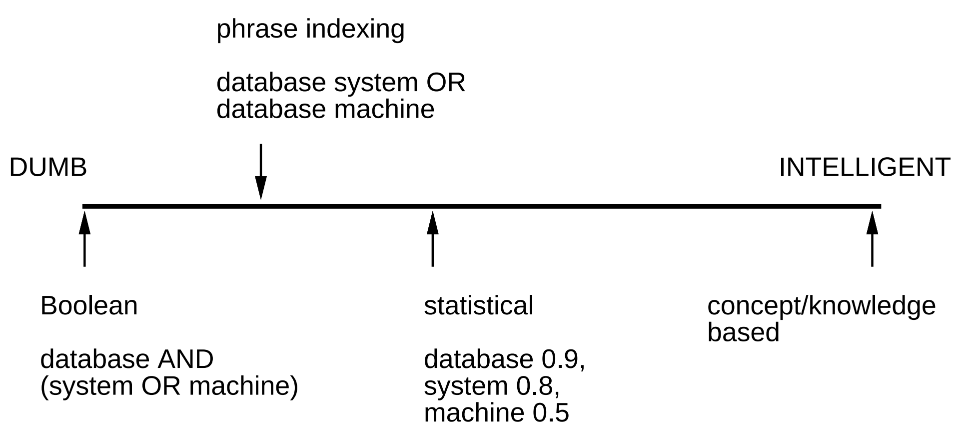


Instead of showing the precision/recall graph, we can

- give the average precision value
- give the precision values at 0.2, 0.5 and 0.8 recall points
- give a single value combining both precision and recall:

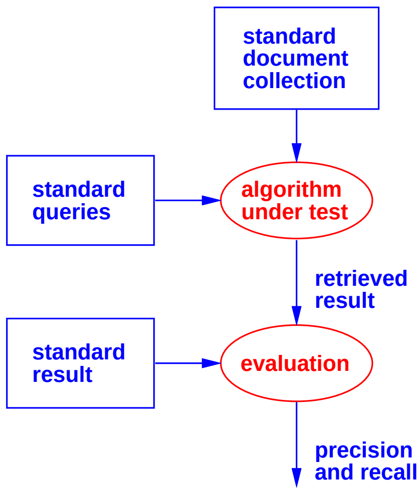
$$E = 1 - \frac{(1 - b^2)PR}{b^2 P + R}$$

2.4 Spectrum of Indexing Methods



2.5 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 relevant judgement. This is analogous to benchmarking of computing systems (e.g, SPECMARKS).
- Performance data is valid only for the environment under which the system is evaluated.



2.6 Problems with Previous Test Collections

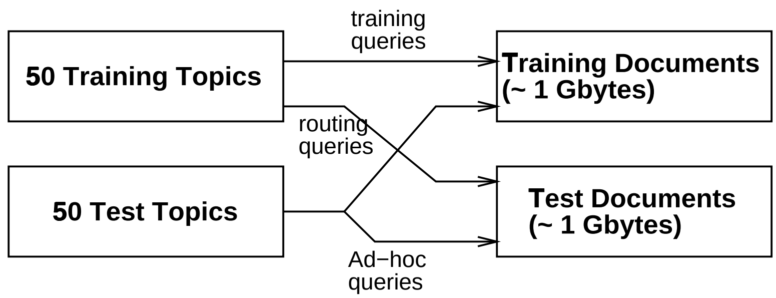
- Previous experiments were based on small collections.

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.

2.7 From Tipster to TREC

- TREC (Text Retrieval Conference) originated from the Darpa sponsored Tipster project in 1990, which involved four defense contractors.
- TREC has been sponsored by both Darpa (Arpa) and NIST starting from 92-93.
- TREC evaluates both Ad hoc and routing queries and provides both training and test collections:
  1. 50 training topics + 1 Gbytes of training documents + relevance judgement
  2. 50 training topics + 1 Gbytes of test documents
  3. 50 test topics + training and test documents.



2.7.1 Characteristics of the TREC collection

- 2 Gbytes of documents (TREC-1)
- 100 topics
- both long and short documents (from a few hundred to over one thousand unique terms in a document)
- test documents consist of:

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

- Documents are marked up with SGML (Standard General Markup Language):

  ⟨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⟩

- A query is markup in SGML with various fields:

  ⟨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
3. software applications
⟨fac⟩ Factor(s):
⟨nat⟩ Nationality: U.S.
⟨/fac⟩
⟨def⟩ Definitions(s):
⟨/top⟩

2.7.2 Relevance Judgement

- exhaustive evaluation:
  100 topics $\times$ 742611 documents = over 74 million judgements

- sampling:
  with average 200 and maximum 900 relevant documents per topic, the sample size is still too large

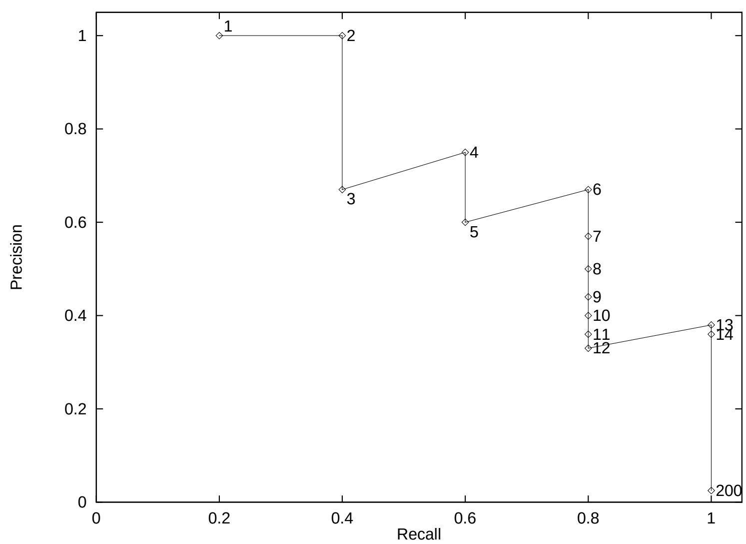
- polling (combine the retrieved documents from each system under test):
  33 runs of 200 top documents: 2398 documents per topic
  22 runs of 100 top documents: 1932 documents per topics.

3 Experimental Methods for Effectiveness Evaluation

3.1 Calculation of Recall and Precision Values

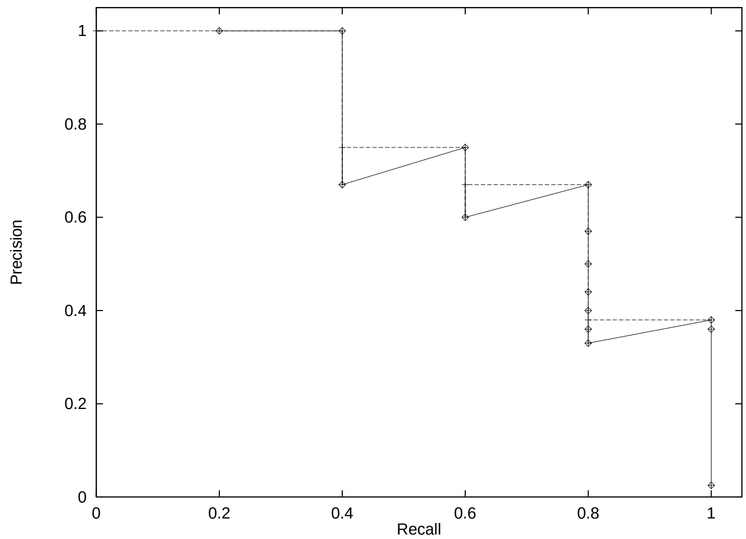
Recall-precision after retrieval of n documents			
n	Doc ID	Recall	Precision
1	588	0.2	1.0
2	589	0.4	1.0
3	576	0.4	0.67
4	590	0.6	0.75
5	986	0.6	0.60
6	592	0.8	0.67
7	984	0.8	0.57
8	988	0.8	0.50
9	578	0.8	0.44
10	985	0.8	0.40
11	103	0.8	0.36
12	591	0.8	0.33
13	772	1.0	0.38
200	...	1.0	0.025

3.2 The Precision-Recall Graph



- Note the sawtooth shape of the graph.
- Values are not defined at every point (e.g., when Recall=0.5).
- Represent performance of one query on one document collection.

3.3 Precision-Recall Graph After Interpolation

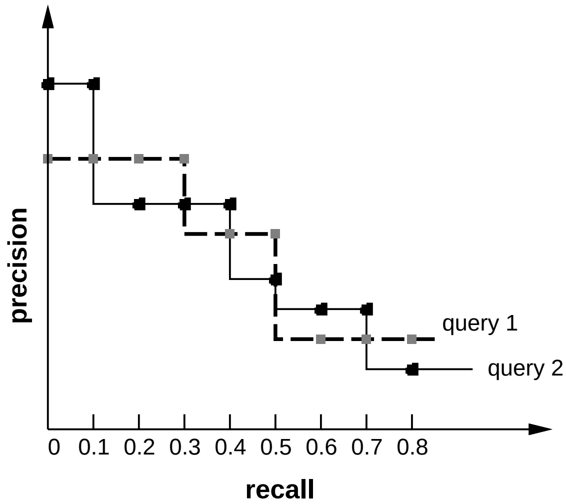


- The interpolation method represents the best result the user can expect.
- Typically values are interpolated at increments of 0.1 or 0.05, resulting in 11 points and 21 points, respectively.

3.4 Averaging Performance Over a Set of Queries

- User-oriented recall-level average:
  - Obtain the precision-recall values for each query and then average over all queries.
- System-oriented document-level average:
  - Accumulate the total numbers of relevant documents, relevant documents retrieved and document retrieved over all queries and then compute the precision and recall values.
- User-oriented recall-level average is more commonly used, because it reflects the performance from a user point of view.

3.5 User-oriented recall-level average



- Average at each recall level after interpolation.