Probabilistic Retrieval

Probabilistic Retrieval

- Suppose that for any query and document we are able to calculate the probability that the document is relevant to the query.
- thus, for a given query, we can order the documents according to these probabilities, giving us our desired ranking.

Relevant Document

- Any given document in the collection has two possibilities in relation to the Query:
  - The document is relevant OR
  - The document is NOT relevant.
- Let the probability of document D_i is relevant be Pr(rel| D_i) and
- The probability of document D_i is NOT relevant be Pr(nrel| D_i).

Estimating Pr(rel| D_i)

- Use Bayes' Theorem:

$$\Pr(rel \mid D_i) = \frac{\Pr(D_i \mid rel)\Pr(rel)}{P(D_i)}$$

- The same estimation can be done for Pr(nrel|D_i).

$$\Pr(nrel \mid D_i) = \frac{\Pr(D_i \mid nrel)\Pr(nrel)}{P(D_i)}$$

Ranking Function

- Document D_j consists of *n* terms => $D_j=(t_1, t_2, t_3,\ldots, t_n)$
- Assume that each term is independent then $Pr(D_j|rel)=Pr(t_1|rel) * Pr(t_2|rel) * Pr(t_3|rel) * ... * Pr(t_n|rel)$
- Let $x_{ij}=Pr(t_i=1|rel)$
  - probability of term t_i exists in the document and document D_j is relevant.
- Let $y_{ij}=Pr(t_i=1|nrel)$
  - probability of term t_i exists in the document but document D_j is NOT relevant.

Ranking Function(2)

$$g(D_j)=\sum_{i=1}^{n} tf_{ij} \ln \frac{x_{ij}(1-y_{ij})}{y_{ij}(1-x_{ij})}$$

- $x_{ij}=Pr(t_{ij}=1|rel)$
- $y_{ij}=Pr(t_{ij}=1|nrel)$
- We can estimate:
  - x_{ij} as *tf.idf* weighting
  - y_{ij} as *df* weighting

Estimation

$$x_{ij}=0.5+(0.5\times ntf_{ij}\times nidf_j)$$

$$y_{ij}=0.5\times\frac{f_j}{N}$$

$$ntf_{ij}=\frac{tf_{ij}}{\max(tf_i)} \qquad nidf_j=\frac{\ln\left(\frac{N}{f_j}\right)}{\ln N}$$

Example

- Consider the same five document collection:
  D1 = “Dogs eat the same things that cats eat”
  D2 = “no dog is a mouse”
  D3 = “mice eat little things”
  D4 = “Cats often play with rats and mice”
  D5 = “cats often play, but not with other cats”

  with dictionary (cat,dog,eat,mouse,play,rat)
- the query: What do cats play with, a vector (1,0,0,0,1,0) be formed.

Example- D1

$$ntf_{cat} = \frac{1}{2} = 0.5 \qquad x_{cat} = 0.5 + (0.5 \times 0.5 \times 0.32) = 0.58$$

$$nidf_{cat} = \frac{\ln\left(\frac{5}{3}\right)}{\ln(5)} = 0.32 \qquad y_{cat} = 0.5 \times \left(\frac{3}{5}\right) = 0.3$$

$$g(D_1) = 1 \times \ln \frac{0.58(1-0.3)}{0.3(1-0.58)} = 1.2$$

Example D4

$$ntf_{play} = \frac{1}{1} = 1 \qquad x_{play} = 0.5 + (0.5 \times 1 \times 0.57) = 0.79$$

$$nidf_{play} = \frac{\ln\left(\frac{5}{2}\right)}{\ln(5)} = 0.57 \qquad y_{play} = 0.5 \times \frac{2}{5} = 0.2$$

$$g(D_4) = 1 \times \ln \frac{0.58(1-0.3)}{0.3(1-0.58)} + 1 \times \ln \frac{0.79(1-0.2)}{0.2(1-0.79)} = 1.2 + 2.7 = 3.9$$

Example D5

$$ntf_{cat} = \frac{2}{2} = 1 \qquad x_{cat} = 0.5 + (0.5 \times 1 \times 0.32) = 0.66$$

$$nidf_{cat} = \frac{\log\left(\frac{5}{3}\right)}{\log(5)} = 0.32 \qquad y_{cat} = 0.5 \times \frac{3}{5} = 0.3$$

$$g(D_5) = 2 \times \ln \frac{0.66(1-0.3)}{0.3(1-0.66)} + 1 \times \ln \frac{0.79(1-0.2)}{0.2(1-.79)}$$

$$= 3.0 + 2.7 = 5.7$$

Performance Measurement

Traditional Measures

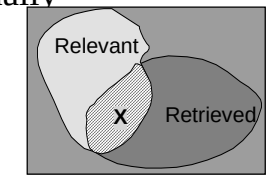
- A perfect retrieval system will show exactly the documents in which a user is interested, and show them in the appropriate order.
- The effectiveness of a text retrieval system can be measures in terms of recall and precision.

Measures

- Recall = No. of relevant documents retrieved / no. of relevant documents
- Precision = No. of relevant documents retrieved / No of retrieved documents

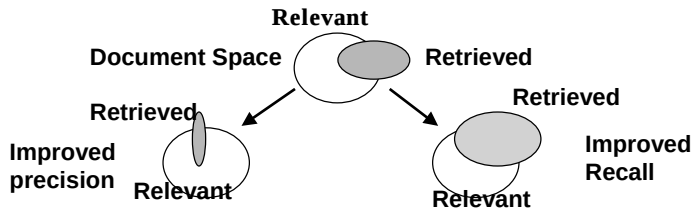
Performance Measure

- Precision: $\frac{X}{Retrieved}$
  - proportion of retrieved items actually relevant
- Recall: $\frac{X}{Relevant}$
  - proportion of relevant information actually retrieved



Measures

- Notice that the more documents that are retrieved the easier it is to obtain better recall, and the fewer documents retrieved the easier it is obtain better precision. the diagram illustrates this. The ideal is to obtain 1.0 for both!



Example

- A document database consists of 20 documents.
- A query submitted to a system resulted in the ranked output: D1,D2,D3,…,D20.
- Assume the relevant documents are D1,D3,D7,D10.

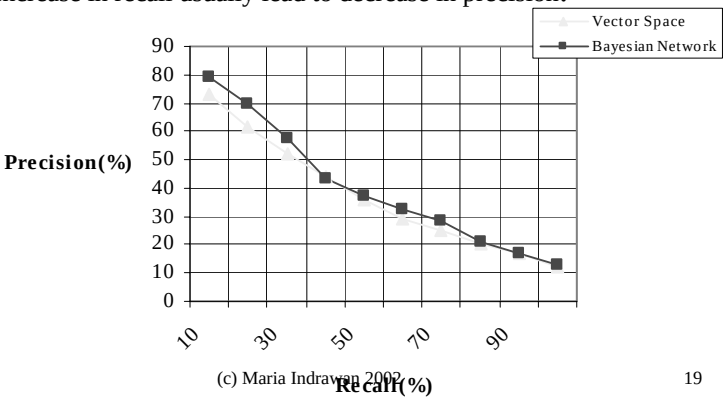
No of documents viewed	Recall	Precision
1	¼	1
2	¼	½
3	2/4	2/3
…	…	…
10	4/4	4/10

Recall Level

- Sometime it is easier to see the relation between recall and precision when precision is measured at a given recall level, eg precision at 10%, 20% recall.

Recall and Precision Relation

- Inverse relation.
  - Increase in recall usually lead to decrease in precision.



Recall and Precision

- The designers of an IR system need to be aware of the user requirements.
  - Is precision the most important consideration?
    - Eg. Search engine
  - Is recall the most important consideration?
    - Eg. Patent office, Law Firm searches for similar cases.

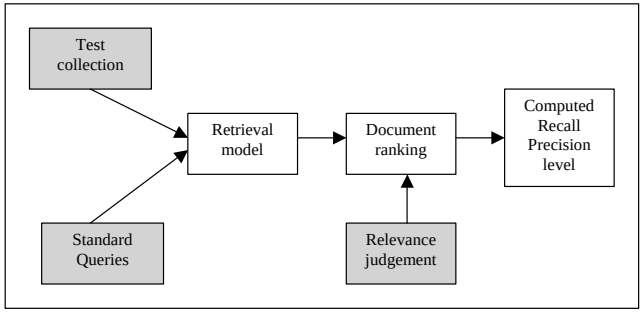
Is 100% Recall and Precision Possible?

- An ideal IR system should produce 100% recall and precision.
- Impossible:
  - Recall and precision relations.
  - Imprecise formulation of information need.
  - Ambiguity of language.

Which System Is Better?

- Suppose two systems produce different ranking of documents. How do we decide which one is better?
- We need to prepare the test environment and run some experiments.
- Test environment:
  - a test database,
  - a set of queries,
  - and the relevance judgments for these queries.

Test Environment



Test Collection(1)

- Traditional collections: ADI, MEDLINE, CACM.

	ADI	MEDLINE	CACM
Information domain	Computing	Medical	Computing
No. documents	82	1033	3204
No. Index terms	2086	52,831	74,391
Ave. no.index term/doc	25.451	51.145	23.218
St.dev of no.index term/doc	8.282	22.547	19.903
No. queries	35	30	64
Ave.no.query terms	9.967	9.967	10.577
Size in kilobytes	2,188	1,091	37,158

Test Collection(2)

- Text Retrieval Conference (TREC) - http://trec.nist.gov/
- Large collection – up to 5 Gb
- Relevant judgement is created using polling methods.
  - Each participants sends a ranked output to a given query.
  - A relevant judgment is created on the merging of the output.

Measure the Test

- We then run both systems to produce, say, two different rankings.
- A simple way of comparing is to provide average precision at various levels of recall for both systems and examine the numbers. The system with the highest number wins!

Other Measures

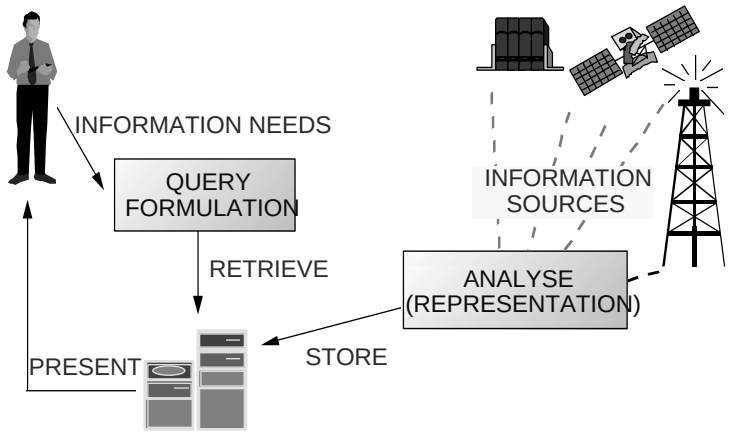
- Other important measures include:
  - time taken to index documents
  - speed with which documents are matched against the query and retrieved
  - storage overhead required for the indexing system.

Some Commercial Systems

Company	Product Name
Oracle	ConTEXT®
ZyLab	ZyImage
Personal Library System	PLS
Open Text Corp	

Relevance Feedback

IR Model - Revisited



Search Process

- Most of IR systems allow user to refine their query. Why?
  - Some new knowledge may be gained by the user after reading some of the suggested documents.
  - A clearer understanding of the information need may be gained by reading the suggested documents.
- The refining process is called Relevance Feedback.

Relevance Feedback

- Relevance Feedback can be achieved by means of:
  - Re-weight the terms in the query.
  - Use the documents as an example, eg find documents which discussed similar topic to D2,D6,D8 but NOT like D1,D5.