CIS 833 – Information Retrieval and Text Mining

Lecture 9

Evaluation in IR

September 22, 2015

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze.

Assignments

- The *warmup* WordCount MapReduce programming assignment due September 23rd
- HW2 due September 25th

Required Reading

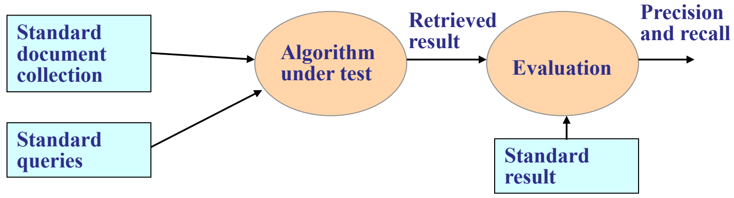
- "Information Retrieval" textbook
  - Chapter 8: Evaluation in IR

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

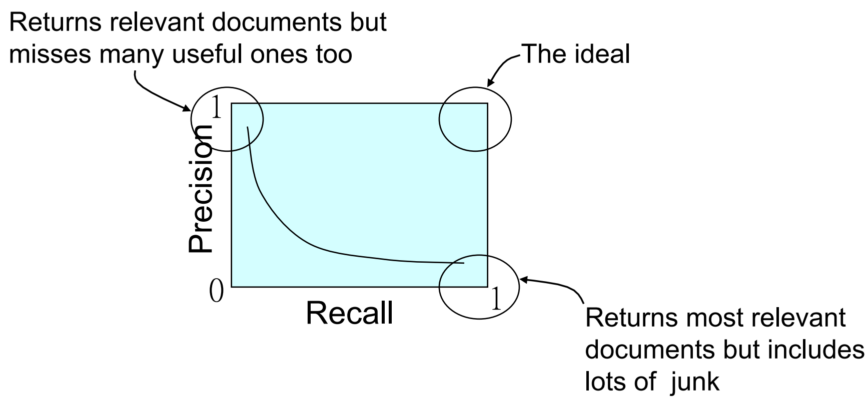
- A benchmark collection contains:
  - A set of standard documents and queries.
  - A list of relevant documents for each query.
- Standard collections for traditional IR:
  - TREC: http://trec.nist.gov/



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.

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: Example 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

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

Computing Recall/Precision Points: Example 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

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=2/3=0.667

R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

R=6/6=1.0; p=6/14=0.429

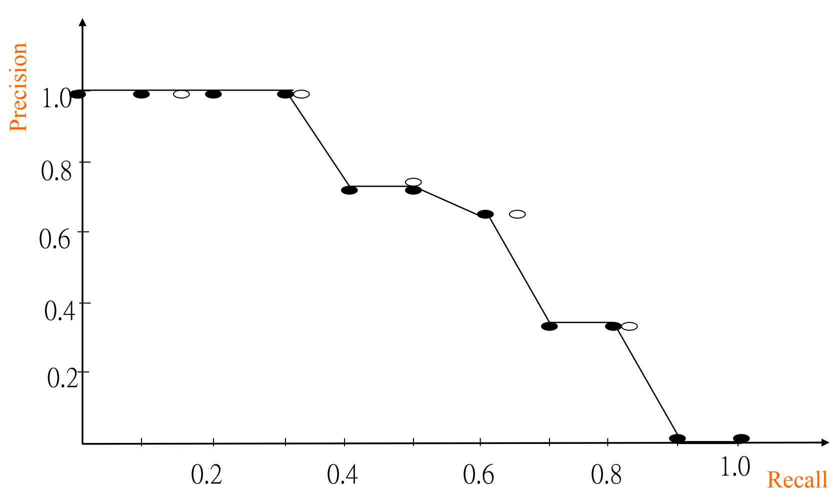
Interpolating a Recall/Precision Curve

- Interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0$, $r_1 = 0.1$, …, $r_{10}=1.0$

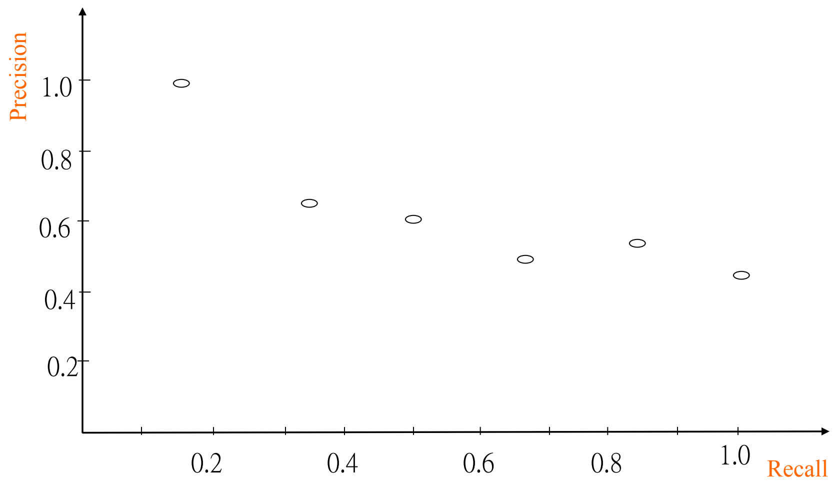
- The interpolated precision at a certain recall level is defined as the highest precision found for any recall level $r' \geq r$:

$$P(r) = \max_{r' \geq r} P(r')$$

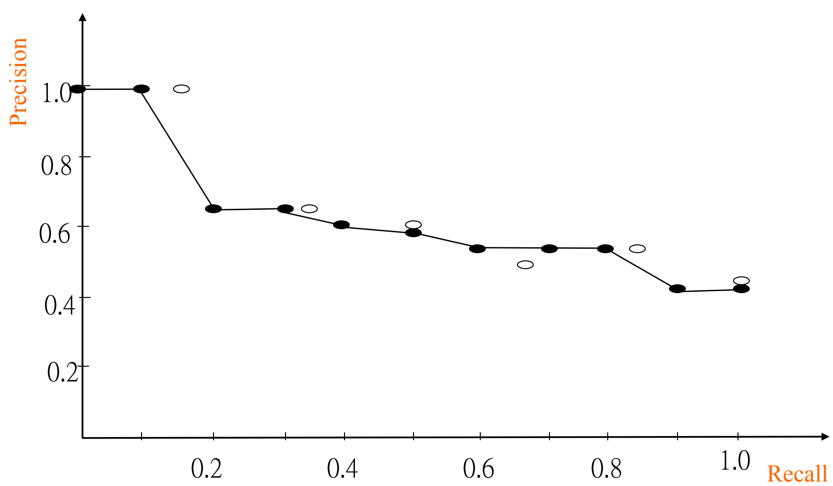
Interpolating a Recall/Precision Curve: Example 1



Interpolating a Recall/Precision Curve: Example 2



Interpolating a Recall/Precision Curve: Example 2

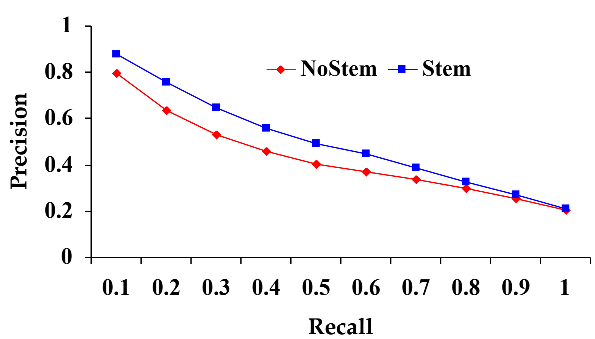


Average Recall/Precision Curve

- Typically, we calculate 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



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:
  - $\beta = 1$: Equally weight precision and recall (E=F).
  - $\beta > 1$: Weight recall more.
  - $\beta < 1$: Weight precision more.

Mean Average Precision (MAP)

- **Average Precision**: Average of the precision values at the points at which each relevant document is retrieved.
  - Ex1: (1 + 1 + 0.75 + 0.667 + 0.38 + 0)/6 = 0.633
  - Ex2: (1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429) = 0.625
- **Mean Average Precision**: Average of the average precision values for a set of queries.

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.

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

Issues with Relevance

- ***Marginal Relevance:*** Do later documents in the ranking add new information beyond what is already given in higher documents.
  - Choice of retrieved set should encourage **diversity** and **novelty.**
- ***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.

Benchmarking - The Problems

- Performance data is valid only for a particular benchmark
- Building a benchmark corpus is a difficult task
- Benchmark web corpora even harder
- Benchmark foreign-language corpora less developed

A/B Testing in a Deployed System

- Can exploit an existing user base to provide useful feedback.
- Randomly send a small fraction (1−10%) of incoming users to a variant of the system that includes a single change.
- Judge effectiveness by measuring change in ***clickthrough:*** The percentage of users that click on the top result (or any result on the first page).

Classes of Retrieval Models

- Boolean models (set theoretic)
  - Extended Boolean
- Vector space models (algebraic)
  - Generalized VS
  - Latent Semantic Indexing
- Probabilistic models
  - Inference Networks
  - Belief Networks

Exact match

Ranking - "Best" match

Required Reading

- Textbook - Chapter 18 (latent semantic indexing)
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, "Indexing by latent semantic analysis". *Journal of the American Society for Information Science, Volume 41, Issue 6, 1990*

Telcordia Technologies

Telcordia™ Technologies
Performance from Experience

Telcordia Latent Semantic Indexing (LSI) Demo Machine

Latent Semantic Indexing (LSI) is a novel, patented information retrieval method developed at Telcordia Technologies, Inc. By using statistical algorithms, LSI can retrieve relevant documents even when they do not share any words with a query. LSI uses these statistically derived "concepts" to improve search performance by up to 30%.

Available on this site are the following:

- LSI Executive Summary
- LSI Demos
- References to Papers on LSI

For more information about LSI, please contact us at: lsi@research.telcordia.com.

Deficiencies with Conventional Automatic Indexing

Synonymy: Various words and phrases refer to the same concept (lowers recall).

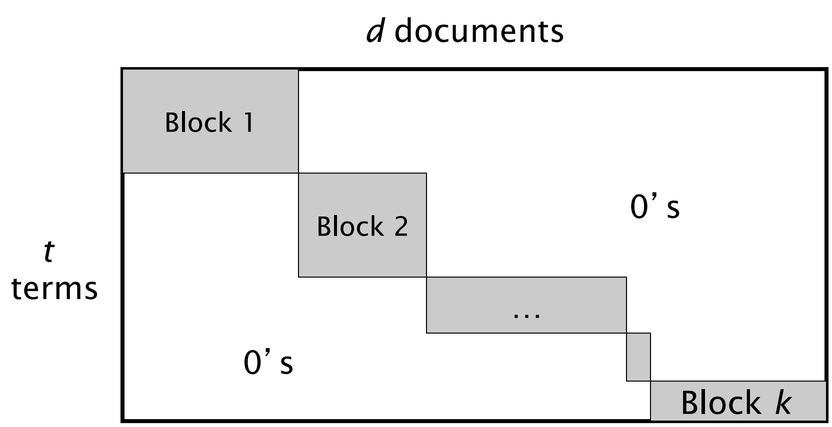
Polysemy: Individual words have more than one meaning (lowers precision)

Independence: No significance is given to two terms that frequently appear together

Latent semantic indexing addresses successfully the first of these (synonymy), and the third (dependence)

- and to a degree the second one (polysemy) - less successfully

Intuition from block matrices



Vocabulary partitioned into *k* topics (clusters); each doc discusses only one topic.

Latent Semantic Indexing

Variant of the vector space model

Objective

Replace indexes that use **sets of terms** by indexes that use **concepts**

Approach

Map the term vector space into a lower dimensional space, using singular value decomposition.

https://en.wikipedia.org/wiki/Singular_value_decomposition

Each dimension in the new space corresponds to a latent concept in the original data - uncorrelated, significant basis vectors

Replace original words with a subset of the new concepts (say 100, but the number may vary) in both documents and queries

Compute similarities in this new space

Computationally expensive, uncertain effectiveness

[Deerwester et al., 1990]

Example

Query: "IDF in computer-based information look-up"

Index terms for a document: access, document, retrieval, indexing

How can we recognize that information look-up is related to retrieval and indexing?

Conversely, if information has many different contexts in the set of documents, how can we discover that it is an unhelpful term for retrieval?

Technical Memo Example: Titles

c1 *Human* machine *interface* for Lab ABC *computer* applications

c2 A *survey* of *user* opinion of *computer system response time*

c3 The *EPS user interface* management *system*

c4 *System* and *human system* engineering testing of *EPS*

c5 Relation of *user*-perceived *response time* to error measurement

m1 The generation of random, binary, unordered *trees*

m2 The intersection *graph* of paths in *trees*

m3 *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m4 *Graph minors*: A *survey*

Technical Memo Example: Terms and Documents

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Technical Memo Example: Query

Query:

Find documents relevant to "human computer interaction"

Simple Term Matching?

Technical Memo Example: Query

Query:

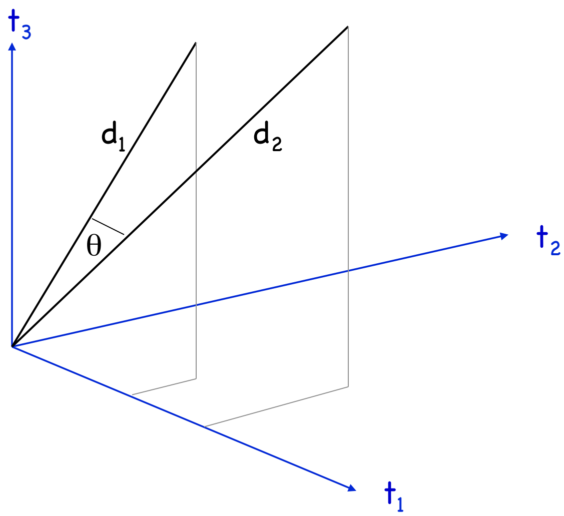
Find documents relevant to "human computer interaction"

Simple Term Matching:

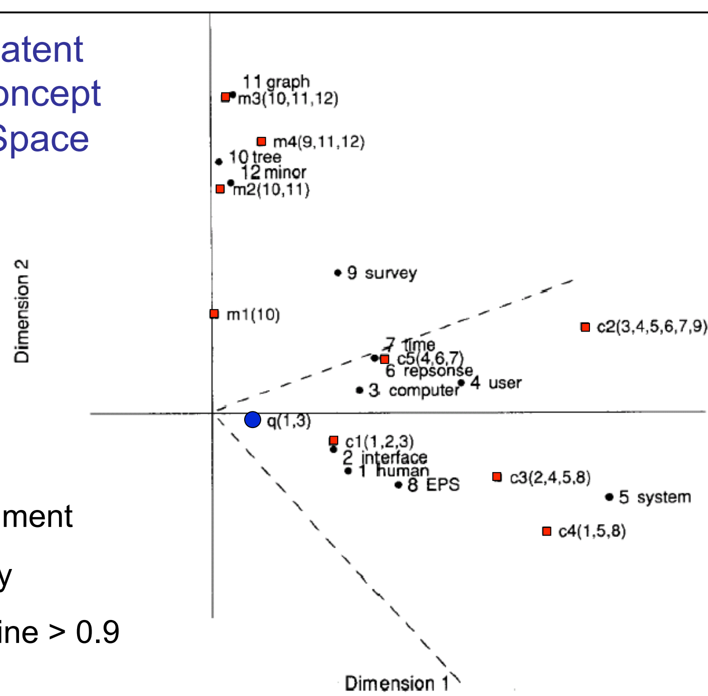
Matches c1, c2, and c4
Misses c3 and c5

Term Vector Space

The space has as many dimensions as there are terms in the vocabulary.



Latent Concept Space



- term
- document
- query

--- cosine > 0.9