Information Retrieval Methods

Helena Ahonen-Myka
Spring 2007, part 2
Relevance. Evaluation.
Translation from Finnish: Greger Lindén

In this part …

- About the concept of relevance
- About evaluation of information retrieval

Relevance

- **relevance** is an important concept in information retrieval (IR), but it is hard to define
- The goal of IR is to find relevant information for the person who needs it
- But:
  - What is relevance?
  - What kind of information or document is relevant?
  - Who evaluates the relevance of a text or a document?
  - On what criteria?

Relevance

- Retrieval results, indexing, etc., are evaluated with methods that are based on the concept of relevance
- There is no single agreement on the definition of relevance
  - relatedness
  - topicality
  - beneficiality
  - utility

Topicality vs. user relevance

- There are two main directions in relevance definitions:
  - Topical relevance: relevance to a subject (topic), topicality, system relevance
    - In its most simple form, matching words in documents and queries
  - User relevance: user oriented view of relevance
    - Based on the user's evaluation of the usefulness of the documents

Topicality vs. user relevance

- Basic assumption about topicality: index words (or phrases) can describe the semantics of a document and a retrieval task sufficiently
  - It is commonly believed that a better matching of keywords leads to a better result
  - For example, the system may try to infer the meaning of a text with advanced linguistic methods
  - But no system has been shown to be perfect

Topicality vs. user relevance

- Topical relevance is useful because it is easy to define and to measure, but it does not contain everything related to relevance
- The main focus in research is now towards user relevance

7

A more specific classification

- Algorithmic relevance
  - Similarity between query and document depending on the matching method
- Topicality
  - Correspondence between topic and text as an interpretation by a human being
- Cognitive relevance
  - The relevance of a document according to the knowledge state of the user

8

A more specific classification, cont.

- Situational relevance
  - The relevance of the document according to the situation, task or problem of the user
- Motivational/emotional relevance
  - The relevance of the document according to the objectives or motives of the user, e.g., the entertainment value

9

Evaluation of IR

- IR research is usually only able to evaluate systems (or methods) in relation to other systems (or methods)
- Assume we want to compare a set of systems S
  - Or one system with different methods or parameter settings
- Assume we have
  - A (large) set of documents D
  - A set of retrieval tasks T (= information needs)
  - Relevance assessments for documents in D

10

Relevance assessments

- Two common models
  - Classical evaluation model (aka Cranfield evaluation framework)
    - For each retrieval task t in T, all the documents in D have been relevance-judged (by human judges)
    - In practice, it is impossible to judge each document for each task
  - TREC (Text REtrieval Conference) framework
    - The set of relevance judgements for each retrieval task is not complete.
    - Retrieval pool: e.g. the top-100 documents returned by each system are collected
    - Only these documents are relevance-judged

11

Relevance assessments

- Relevance values are usually binary
  - A document is either relevant or non-relevant (not relevant) for a task
  - Multi-graded relevance values could be used (e.g. significant/useful/marginal/irrelevant)
- Developers of the systems/methods are not allowed to participate
  - in defining retrieval tasks
  - in relevance assessments

12

Evaluation process

- For each system and each retrieval task, formulate a query
- Let each system match each query against the documents in the database
- Let's define:
  - a search request = processing one retrieval task by one system
  - includes formulating a query, matching the query against documents, and returning a result
- Result of a search request: a set of documents (often in some order)
- The results are evaluated based on some evaluation criteria

13

Evaluation criteria

- The most common evaluation criteria
  - Recall (saanti; åtkomst)
  - Precision (tarkkuus; precision)

14

Recall and precision

- The result divides the documents in the database into two sets
  - The retrieved documents
  - The documents that were not retrieved
- In principle, all documents in the database should be evaluated for relevance; then we could divide the database into
  - Relevant documents for the task
  - Not relevant documents for the task

15

Definition of recall and precision

	Re	levan	ce
Answer set	Relevant	Non-relevant	Total
Retrieved	a Matches/true positives	b false positives	a + b retrieved
Rejected	c false negatives	d true negatives	c + d rejected
Total	a + c relevant	b + d non-relevant	a + b + c + d database

16

Recall and precision

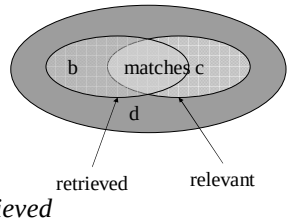
- **recall**
  - The fraction of relevant documents that were retrieved: a / (a + c)
  - The number of correct responses divided by the number of possibly correct responses
- **precision**
  - The fraction of retrieved documents that is relevant: a / (a + b)
  - The number of correct responses divided by the total number of actual responses
- Both are represented by decimal numbers [0,1] or by percentage numbers 0...100%

17

Recall and precision

- precision

$$\pi = \frac{relevant\_and\_retrieved}{retrieved}$$



- recall

$$\rho = \frac{relevant\_and\_retrieved}{relevant}$$

18

Recall and precision

- Together recall and precision are two concrete measurements for how well the retrieval succeeded
  - The recall denotes how much information the user received (in relation to how much there would have been)
  - The precision measure denotes how much work the user must do in order to find the relevant documents in the answer set
- We can often influence recall and precision by our design decisions

19

Relation between recall and precision

- The relation between recall and precision is inverse
  - Better recall usually means worse precision and vice versa
  - 100% recall is always possible by returning all documents → precision might then be close to zero
- E.g. if we add keywords to a query, the recall will increase but the precision will decrease
  - New keywords find other documents that use different words to describe the same topic but these keywords might also refer to other topics

20

Computing recall and precision

- The line below denotes the result of a search request:
  - 20 documents were returned in the result set: the documents are numbered in the order they were returned
  - For each document, we denote if it is relevant (+) or not (-)

d# 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

\- - - + - + - - - - - - + - - + - - - + -

21

Computing recall and precision

- Let us assume that there are 10 relevant documents in the document collection (i.e. relevant for this retrieval task)
- Exact matching (e.g. the query is a Boolean expression)
  - The result is a set where the documents are not ordered
    - Usually a subset of the whole document collection
    - It is possible that some relevant documents are not found
  - precision: 5/20 = 25%
  - recall: 5/10 = 50%

22

Computing recall and precision

- Partial matching (the query is a set of terms)
  - The result is a list of documents ordered according to the relevance of the document
    - Relevance is defined by the search system according to the similarity between the query and the documents
  - In principle, the whole document collection is the result, ordered according to relevance probability
    - All relevant documents will be found at some stage
    - It is not reasonable to calculate just one recall and precision
  - The result can be evaluated at separate stages

23

Computing recall and precision

Document #	Recall %	Precision %
4	10	25
6	20	33
12	30	25
15	40	27
19	50	26

24

d# 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17...30...45

\- + - - - - - - - + - - - - - - + + +

	r%	p%		r%	p%
1:	0	0	17:	60	18
2:	20	50	...		
3:	20	33	30	80	13
4:	20	25	...		
5:	20	20	45	100	11
...					
9:	20	11			
10:	40	20			

25

Recall-precision curve

- We compute precision values for different recall values
  - recall 20%, precision 50%
  - recall 40%, precision 20%
  - recall 60%, precision 18%
  - recall 80%, precision 13%
  - recall 100%, precision 11%
- We draw the points in the coordinate system and interpolate a curve between the points
- Usually 10% steps are used

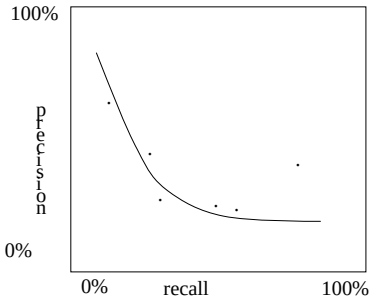
26

Recall and precision

- Usually we study a large set of results and are interested in the average recall and precision values
- We can, for example, gather the precision values for each search request (of a system) when recall is 10%, 20%,…, 100%, and compute the average precision at each stage (over the search requests)
- Average values can also be presented in a recall-precision graph

27

Recall-precision graph



28

The DCV curve

- The user may be interested mainly only in the first retrieved documents
- We can focus on recall and precision at stages that correspond to a certain size of the answer set
  - After 5 documents, after 10 documents
  - → DCV (Document Cut-off Value) curve

29

d# 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17...30...45

\- + - - - - - - - + - - - - - - + + +

	r%	p%	
2:	20	50	1st relevant document
5:	20	20	
10:	40	20	2nd relevant
15:	40	13	
20:	60	15	3rd relevant
25:	60	12	
30:	80	13	4th relevant
35:	80	11	
40:	80	10	
45:	100	11	5th relevant

30

Problems with recall and precision

- We do not know (in practice) the number of relevant documents in the document collection
  - An approximate value is used
- It can happen that we see a document in the result set that is not relevance-judged
  - In TREC framework, non-judged documents are assumed to be irrelevant
  - Other evaluation methods which ignore non-judged documents exist

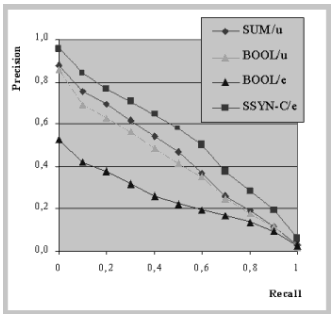
31

Comparing search methods

- We can compute from the results for each search request its successfulness, e.g. as a recall-precision curve
- If we compute the average precisions for a set of search requests of a system, we detect the performance of this retrieval system
- Usually we study average performances of several different methods

32

Result curves for some methods



33

Comparing search methods

- In the previous picture, the performance of four search methods is compared
  - Each method is represented by a recall-precision curve in a different colour
  - Each curve presents the average precision at different recall levels;
  - Each curve represents one search method using 30 retrieval tasks
- The average precision of the best method at 50% recall is almost 60% and only about 20% for the worst one → there seems to be differences in performance

34

Comparing search methods

- When developing retrieval methods, it is important to evaluate which differences are significant
- We often compute the average of the performance curve at 11 points
  - The average of the precision values at recall levels 0-100% (at each 10%, "standard recall levels")
  - E.g., the precision average of the best method over different recall levels is about 60%, the others' about 50%, 40% and 20%

35

Comparing search methods

- In practice, the meaning of the differences
  - difference over 15%: significant
  - difference 10-15%: important?
  - difference 5-10%: interesting
  - difference under 5%: marginal
- In addition we can compute the statistical significance
  - How probable is it that the difference could have emerged by chance?
  - Statistical tests, e.g. the t-test

36

Comparing search methods

- Interpreting the results from a recall-precision curve can be difficult, if recall bases for each task differ a lot (recall base = number of relevant documents in the database)
- If we know that the best method reaches 50% precision at recall level 60%, we still do not know how many documents the user will retrieve
- Varying sizes of the recall bases is also a problem in the DCV curve
  - If the recall base contains 5 documents, the precision at result size 50 documents cannot be very high

37

In this part

- Different views on how to define relevance
- Basic principles for evaluating IR methods and systems
  - Evaluation criteria recall and precision
  - Evaluation of the result of one search request
  - Evaluation of the performance of one system using a set of search requests
  - Comparing several systems

38