Performance Evaluation of Information Retrieval Systems

Dan Smith

UEA
University of East Anglia

Overview

- Document summaries
  - presenting results to the user
- Evaluation methods
- Metrics
- TREC conferences

3

PRESENTING RESULTS

4

Presenting results

- Having ranked the documents matching a query, we need to present the results
  - Most commonly, the document title plus a short summary
- The title is typically automatically extracted from document metadata
- What about the summaries?

5

Summaries

- A static summary of a document is always the same, regardless of the query
- Dynamic summaries are *query-dependent* attempt to explain why the document was retrieved for the query at hand

6

Static summaries

- In typical systems, the static summary is a subset of the document
- Simplest heuristic: the first 50 (or so) words of the document
  - Summary cached at indexing time
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
- Most sophisticated: NLP used to synthesize a summary
  - Seldom used in IR; cf. text summarization work

7

Dynamic summaries

- Present one or more passages from the document containing the query terms
  - "KWIC" snippets: Keyword in Context presentation
- Generated in conjunction with scoring
  - If query found as a phrase, passages containing the phrase
  - If not, passages that contain multiple query terms

8

Creating dynamic summaries

- If we have only a positional index, we cannot (easily) reconstruct context surrounding hits
- If we *cache the documents* at index time, can run a sliding window through it, cueing to hits found in the positional index
  - e.g., positional index says "the query is a phrase in position 4378" so we go to this position in the cached document and stream out the content
- Most often, cache a fixed-size prefix of the doc
  - which can be outdated…

9

Summary

- Producing good dynamic summaries is a tricky optimization problem
  - The screen area for the summary is normally small and fixed
  - Want short item, so show as many KWIC matches as possible, and perhaps other things like title
  - Want snippets to be long enough to be useful
  - Want linguistically well-formed snippets: users prefer snippets that contain complete phrases
  - Want snippets that are maximally informative
- Users really like snippets, even if they complicate IR system design

10

SYSTEM EVALUATION

11

Why system evaluation?

- There are many retrieval models, algorithms and systems, which one is the best?
- What is the best component for:
  - Ranking function (dot-product, cosine, …)
  - Term selection (stopword removal, stemming, …)
  - Term weighting (TF, TF-IDF, …)
- How far down the ranked list will a user need to look to find relevant documents?

12

Difficulties in evaluating IR systems

- Effectiveness is related to the ***relevance*** of retrieved items.
- Relevance is not typically binary but continuous.
- Even if relevance is binary, it can be a difficult judgment to make.
- Relevance, from a human standpoint, is:
  - Subjective: Depends on a specific user's judgment
  - Situational: Relates to user's current needs
  - Cognitive: Depends on human perception, behaviour
  - Dynamic: Changes over time

Human labeled corpora (gold standard)

13

- Start with a corpus of documents
- Collect a set of queries for this corpus
- Have one or more human experts label the relevant documents for each query
- Typically assumes binary relevance judgments
- Requires considerable human effort for large document/query corpora
- Human judges are not all consistent

Accuracy

14

- Given a query an engine classifies each doc as "Relevant" or "Irrelevant"
- Accuracy of an engine: the fraction of these classifications that is correct

15

What's wrong with accuracy?

- How to build a 99.9999% accurate search engine on a low budget….
- Never return any results!

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} = \frac{0 + HUGE}{0 + HUGE + 0 + 0} = 1.0$$

- People doing information retrieval *want to find something* and have a certain tolerance for junk

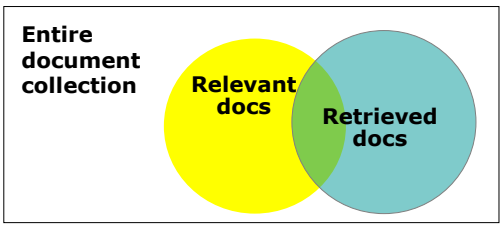
16

Precision and recall

- Precision
  - The ability to retrieve top-ranked documents that are mostly relevant
- Recall
  - The ability of the search to find ***all*** the relevant items in the corpus

17

Precision and Recall



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved & relevant

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$precision = \frac{Number\ \ of\ relevant\ documents\ \ retrieved}{Total\ number\ of\ \ documents\ \ retrieved}$$

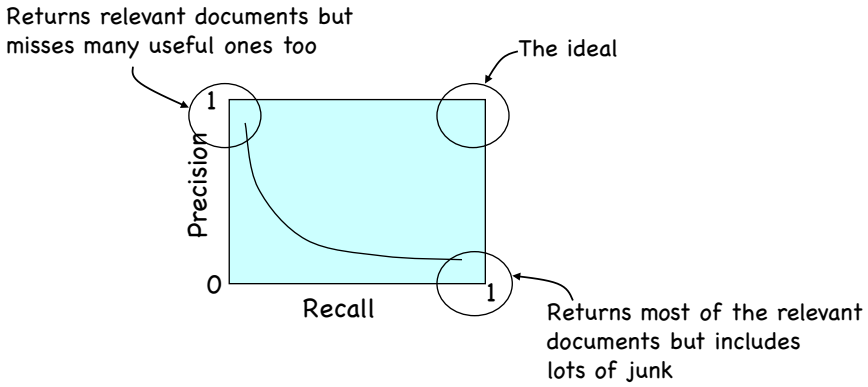
18

Determining recall is difficult

- Total number of relevant items is sometimes not available:
  - Sample across the database and perform relevance judgment on these items.
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.

19

Trade-off between recall and precision



20

Computing precision/recall 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 precision/recall measures
- Mark each document in the ranked list that is relevant according to the gold standard
- Compute a precision/recall pair for each position in the ranked list that contains a relevant document

21

Computing precision/recall points

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	

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=2/2=1

R=3/6=0.5; P=3/4=0.75

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

Missing one relevant document. Never reach 100% recall

R=5/6=0.833; p=5/13=0.38

22

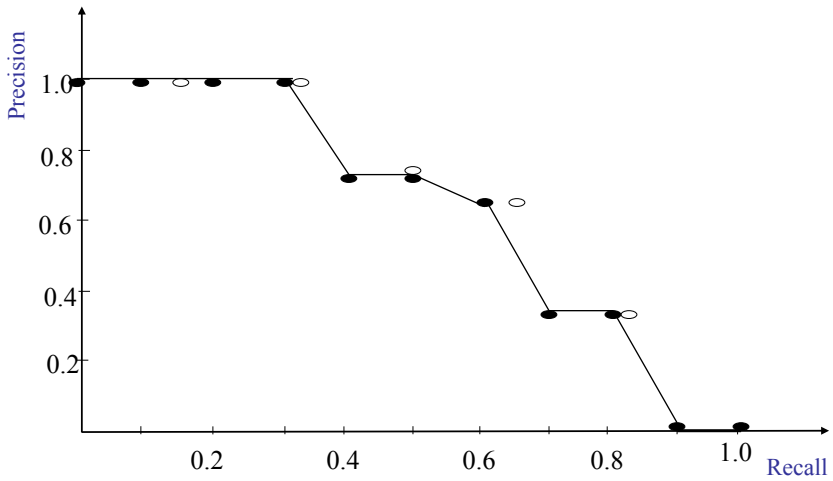
Interpolating a precision/recall curve

- Interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- The interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level between the *j*-th and (*j* + 1)-th level:

$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

23

Precision/recall curve example



24

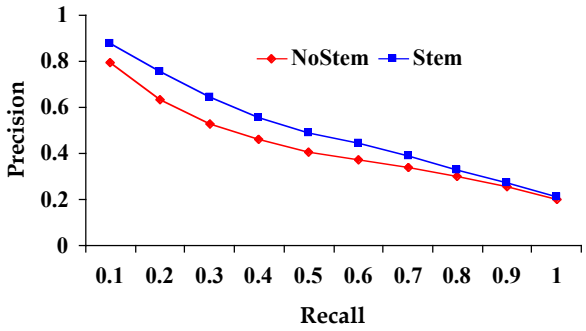
Average precision/recall curve

- 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

25

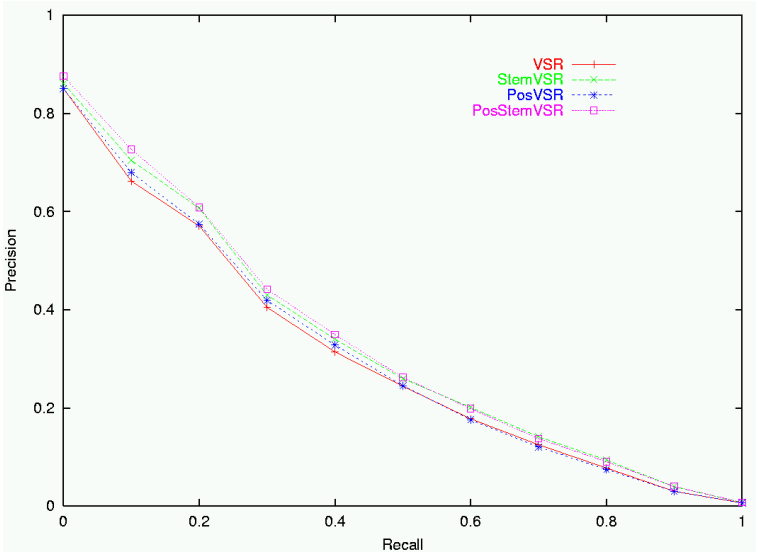
Compare two or more systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



26

Sample p/r curve for CF corpus



27

Mean Average Precision

- Mean average precision (MAP)
  - Average of the precision value obtained for the top *k* documents, each time a relevant doc is retrieved
  - Avoids interpolation, use of fixed recall levels
  - MAP for query collection is arithmetic average
    - Macro-averaging: each query counts equally

28

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

29

F-measure

- A 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

30

Evaluation for classification

- Accuracy
  - Proportion of items correctly classified
- In a two-class problem
  - Proportion of items wrongly classified as irrelevant (Type I error)
  - Proportion of items wrongly accepted as relevant (Type II error)
- **Always important to understand why items are misclassified**
  - Misclassification patterns indicate areas for improvement

31

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

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

32

SUBJECTIVE MEASURES OF RELEVANCE

33

Subjective relevance measure

- *Novelty Ratio*: The proportion of items retrieved and judged relevant by the user and of which they were previously unaware.
  - Ability to find ***new*** information on a topic
- *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 2015/16

34

Other factors

- *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

35

TEST COLLECTIONS AND EVALUATIONS

36

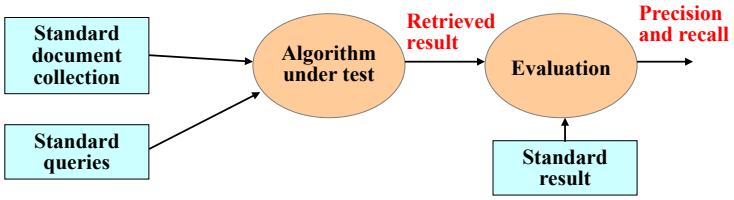
Experimental setup for benchmarking

- ***Analytical*** performance evaluation is difficult for document retrieval systems
  - relevance is difficult to describe with mathematical precision
- Performance is measured by ***benchmarking***
  - 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

37

Benchmarks

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



38

Benchmarking – the problems

- Performance data is valid only for a particular benchmark
- Building a benchmark corpus is difficult
  - size and suitability
  - availability of data
  - judgments
- Benchmark corpora are not as well developed in languages other than English
  - but the situation has greatly improved

39

Early test collections

- Early experiments were based on the SMART collection - which is 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

- Different researchers used different test collections and evaluation techniques

40

TREC

- TREC: Text REtrieval Conference
  - organised by NIST (http://trec.nist.gov/)
- Annual conference from 1992
- Participants are given parts of a standard set of documents and topics for training and testing
- Participants submit the P/R values for the final document and query corpus
- Results presented at the conference.

41

TREC objectives

- Provide a common ground for comparing different IR techniques
- Sharing of resources and experiences in developing benchmark collections and tasks
- Development of evaluation techniques, particularly for new applications

42

TREC Tasks

- Organised as a set of tracks
- Tracks are decided by consensus
  - data sent out in March
  - results submitted July/August
  - evaluation feedback late September
  - conference mid-November
- Research groups participate in 1 or more tracks

43

Evolution of TREC tracks

2011

- Chemical IR
- Crowdsourcing
- Entity
- Legal
- Medical Records
- Microblog
- Session
- Web

2016

- Clinical Decision Support
- Contextual Suggestion
- Dynamic Domain
- Live QA
- OpenSearch
- Real-Time Summarization
- Tasks
- Total Recall

44

TREC 2016 track descriptions

- Clinical Decision Support
  - techniques for linking medical cases to biomedical literature relevant for patient care
- Contextual Suggestion
  - search techniques for complex information needs that are highly dependent on context and user interests
- Dynamic Domain
  - domain-specific search algorithms that adapt to the dynamic information needs of professional users as they explore in complex domains
- Live QA
  - generate answers to real questions originating from real users via a live question stream, in real time

45

TREC 2016 track descriptions

- OpenSearch
  - an evaluation paradigm for IR that involves real users of operational search engines
- Real-Time Summarization
  - techniques for constructing real-time update summaries from social media streams in response to users' information needs
- Tasks
  - can systems induce the possible tasks users might be trying to accomplish given a query
- Total Recall
  - methods to achieve very high recall, including methods that include a human assessor in the loop

46

TREC collections

- Early collections included

WSJ	Wall Street Journal articles (1986-1992)	550 MB
AP	Associated Press Newswire (1989)	514 MB
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 MB
FR	Federal Register	469 MB
DOE	Abstracts from Department of Energy reports	190 MB

- A small TREC collection is about 150GB
- 2011 Web track is about 25TB uncompressed…

47

Sample TREC 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>

48

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>

Evaluation

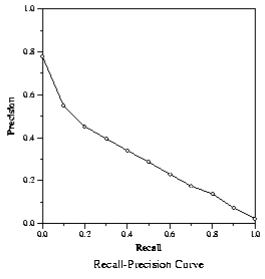
- Summary table statistics: Number of topics, number of documents retrieved, number of relevant documents.
- Recall-precision average: Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- Document level average: Average precision when 5, 10, .., 100, ... 1000 documents are retrieved.
- Average precision histogram: Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

Ad hoc results — Fujitsu Laboratorics. Ltd.

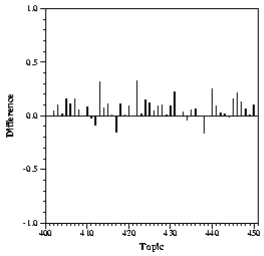
Summary Statistics	
Run Number	Flab8atd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

Recall Level Precision Averages	
Recall	Precision
0.00	0.7796
0.10	0.5490
0.20	0.4517
0.30	0.3954
0.40	0.3397
0.50	0.2863
0.60	0.2291
0.70	0.1745
0.80	0.1381
0.90	0.0720
1.00	0.0224
Average precision over all relevant docs	
non interpolated	0.2930

Document Level Averages	
	Precision
At 5 docs	0.5480
At 10 docs	0.4880
At 15 docs	0.4587
At 20 docs	0.4200
At 30 docs	0.3887
At 100 docs	0.2490
At 200 docs	0.1777
At 500 docs	0.1011
At 1000 docs	0.0598
R Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3203



Recall-Precision Curve



Difference from Median in Average Precision per Topic

Resources

- TREC http://trec.nist.gov/
- CLEF http://www.iei.pi.cnr.it/DELOS/CLEF/
- TRECVid http://trecvid.nist.gov/

- *Intro to IR*, Ch 8
- R. Belew, *Finding Out About*, Ch. 4
- S. Mizzaro, "How many relevances in information retrieval?", *Interacting with Computers,* 10(3): 303-320 (1998)