
Performance Evaluation of Information Retrieval Systems

Many slides in this section are adapted from Prof. Joydeep Ghosh (UT ECE) who in turn adapted them from Prof. Dik Lee (Univ. of Science and Tech, Hong Kong)

1

1

Why System Evaluation?

- There are many retrieval models/ algorithms/ 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 some/all relevant documents?

2

2

Difficulties in Evaluating IR Systems

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

3

3

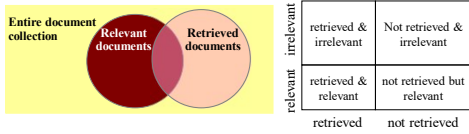
Human Labeled Corpora (Gold Standard)

- Start with a corpus of documents.
- Collect a set of queries for this corpus.
- Have one or more human experts exhaustively label the relevant documents for each query.
- Typically assumes binary relevance judgments.
- Requires considerable human effort for large document/query corpora.

4

4

Precision and Recall



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

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

5

5

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.

6

6

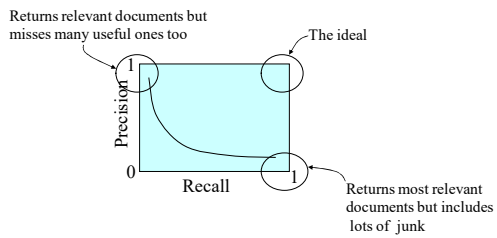
Determining Recall is Difficult

- 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.

7

7

Trade-off between Recall and Precision



8

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.

9

9

Computing Recall/Precision Points: Example 1

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

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

Missing one relevant document.
Never reach 100% recall

10

10

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

11

11

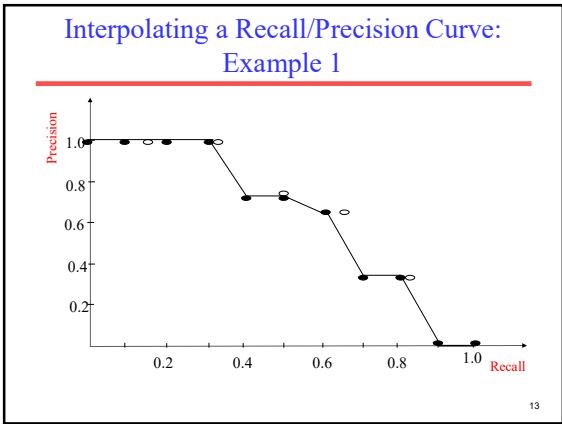
Interpolating a Recall/Precision Curve

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

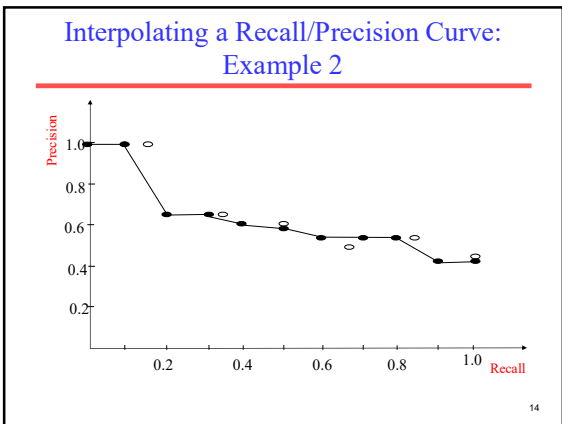
$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$

12

12



13



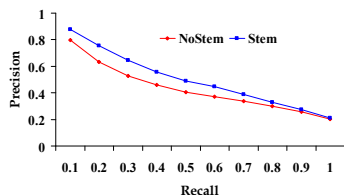
14

- ### Average Recall/Precision 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.

15

Compare Two or More Systems

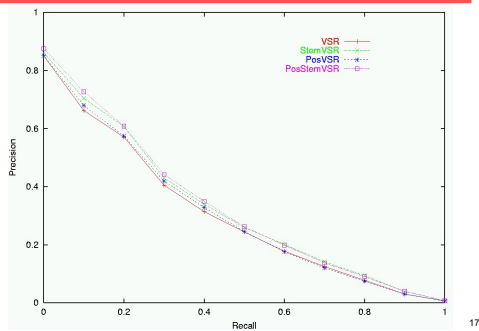
- The curve closest to the upper right-hand corner of the graph indicates the best performance



16

16

Sample RP Curve for CF Corpus



17

17

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

18

18

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.

19

19

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.

20

20

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: $(1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429)/6 = 0.625$
- **Mean Average Precision:** Average of the average precision value for a set of queries.

21

21

Non-Binary Relevance

- Documents are rarely entirely relevant or non-relevant to a query
- Many sources of *graded relevance judgments*
 - Relevance judgments on a 5-point scale
 - Multiple judges
 - Click distribution and deviation from expected levels (but click-through != relevance judgments)

22

22

Cumulative Gain

- With graded relevance judgments, we can compute the *gain* at each rank.
- **Cumulative Gain** at rank n:

$$CG_n = \sum_{i=1}^n rel_i$$

(Where rel_i is the graded relevance of the document at position i)

n	doc #	relevance	
		(gain)	CG _n
1	588	1.0	1.0
2	589	0.6	1.6
3	576	0.0	1.6
4	590	0.8	2.4
5	986	0.0	2.4
6	592	1.0	3.4
7	984	0.0	3.4
8	988	0.0	3.4
9	578	0.0	3.4
10	985	0.0	3.4
11	103	0.0	3.4
12	591	0.0	3.4
13	772	0.2	3.6
14	990	0.0	3.6

23

23

Discounting Based on Position

- Users care more about high-ranked documents, so we **discount** results by $1/\log_2(rank)$

- **Discounted Cumulative Gain:**

$$DCG_n = rel_1 + \sum_{i=2}^n \frac{rel_i}{\log_2 i}$$

n	doc #	rel		log ₂	DCG _n
		(gain)	CG _n		
1	588	1.0	1.0	-	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

24

24

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.

28

28

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).

29

29

Experimental Setup for Benchmarking

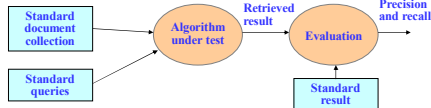
- *Analytical* performance evaluation is difficult for document retrieval systems because many characteristics such as relevance, distribution of words, etc., are difficult to describe with mathematical precision.
- Performance is measured by *benchmarking*. That is, the retrieval effectiveness of a 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.

30

30

Benchmarks

- A benchmark collection contains:
 - A set of standard documents and queries/topics.
 - A list of relevant documents for each query.
- Standard collections for traditional IR:
 - Smart collection: <ftp://ftp.cs.cornell.edu/pub/smart>
 - TREC: <http://trec.nist.gov/>



31

31

Benchmarking – The Problems

- Performance data is valid only for a particular benchmark.
- Building a benchmark corpus is a difficult task.
- Representative corpora for web search are hard to make public due to privacy concerns.
- Benchmark foreign-language corpora are limited.

32

32

The TREC Benchmark

- TREC: Text REtrieval Conference (<http://trec.nist.gov/>)
Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.
- Participants submit the P/R values for the final document and query corpus and present their results at the conference.

33

33

Characteristics of the TREC Collection

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).

- Test documents consist of:

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

34

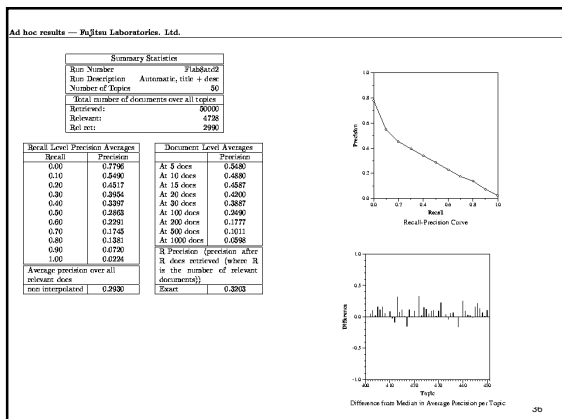
34

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.

35

35



36

Cystic Fibrosis (CF) Collection

- 1,239 abstracts of medical journal articles on CF.
- 100 information requests (queries) in the form of complete English questions.
- Relevant documents determined and rated by 4 separate medical experts on 0-2 scale:
 - 0: Not relevant.
 - 1: Marginally relevant.
 - 2: Highly relevant.

37

37

CF Document Fields

- MEDLINE access number
- Author
- Title
- Source
- Major subjects
- Minor subjects
- Abstract (or extract)
- References to other documents
- Citations to this document

38

38

Sample CF Document

AN 74154352
AU Burnell-R-H. Robertson-E-F.
TI Cystic fibrosis in a patient with Kartagener syndrome.
SO Am-J-Dis-Child. 1974 May. 127(5). P 746-7.
MJ CYSTIC-FIBROSIS: co. KARTAGENER-TRIAD: co.
MN CASE-REPORT. CHLORIDES: an. HUMAN. INFANT. LUNG: ra. MALE.
SITUS-INVERSUS: co. ra. SODIUM: an. SWEAT: an.
AB A patient exhibited the features of both Kartagener syndrome and cystic fibrosis. At most, to the authors' knowledge, this represents the third such report of the combination. Cystic fibrosis should be excluded before a diagnosis of Kartagener syndrome is made.
RF 001 KARTAGENER M BEITR KLIN TUBERK 83 489 933
002 SCHWARZ V ARCH DIS CHILD 43 695 968
003 MACE JW CLIN PEDIATR 10 285 971
...
CT 1 BOCHKOVA DN GENETIKA (SOVIET GENETICS) 11 154 975
2 WOOD RE AM REV RESPIR DIS 113 833 976
3 MOSSBERG B MT SINAI J MED 44 837 977
...

39

39

Sample CF Queries

QN 00002
QU Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respiratory tract in CF?
NR 00007
RD 169 1000 434 1001 454 0100 498 1000 499 1000 592 0002 875 1011

QN 00004
QU What is the lipid composition of CF respiratory secretions?
NR 00009
RD 503 0001 538 0100 539 0100 540 0100 553 0001 604 2222 669 1010 711 2122 876 2222

NR: Number of Relevant documents
RD: Relevant Documents

Ratings code: Four 0-2 ratings, one from each expert

40

40

Preprocessing for VSR Experiments

- Separate file for each document with just:
 - Author
 - Title
 - Major and Minor Topics
 - Abstract (Extract)
- Relevance judgment made binary by assuming that *all* documents rated 1 or 2 by *any* expert were relevant.

41

41
