Information Retrieval

Week 4
Evaluation

Qianni Zhang (qianni.zhang@qmul.ac.uk)

Roadmap of this lecture

- What to evaluate
- Relevance
- Test collections
- Precision and recall
- Single value measurements
- Discounted cumulative gain

Introduction

What can we evaluate in IR

A range of aspects affect the quality of a search system:

- **coverage** of the collection: extent to which the system includes relevant material
  - this is (was) important in web retrieval since it is known that individual search engine covers maybe up to 16% of the web space[1]
- **efficiency** in terms of speed, memory usage, etc.
- **time lag (efficiency)**: average interval between the time a request is made and the answer is given
- **presentation** of the output, which has to do with interface and visualisation issues, user interaction
- **effort** involved by user in obtaining answers to a request
- **effectiveness** of the system (retrieval quality)

[1] This was the case with search engines such as Altavista, Lycos, etc.

Introduction

Relevance

- A document is relevant if it has a significant and demonstrable bearing on the matter at hand (the search task/query).
- There are common assumptions about the nature of relevance in system-centred evaluation:
  - **Objectivity**: everybody agrees on whether a document is relevant or not to a query
  - **Topicality**: relevance is about whether the document is about the topic expressed in the query
  - **Binary nature**: a document is either relevant or not
  - **Independence**: the fact that a document is relevant to a query has no effect on the relevance of another document for that same query

Relevance

- Relevance: difficult to define satisfactorily
- A relevant document is judged useful in the context of a query
  - Who judges? What is useful?
  - Humans not very consistent
  - Judgements depend on more than the document and query
- With real collections, we never know full set of relevant documents
- It has become clear that there is much more than 'system' relevance (i.e. the match of a document to the query)
- There has been a huge amount of research on studying relevance and its different types

Introduction

Types of Relevance: Saracevic

There are various different types of relevance defined by researchers. A very popular ’classification’ is given by Saracevic with the following 5 types:

1. **System/algorithmic relevance**: Extent by which the query is matched by the document. System/algorithm dependent.
2. **Topical/aboutness/subject relevance**: Extent by which the subject or topic expressed in the query is also expressed in a document. Depends on user expectations and on intent behind the query.
3. **Pertinence/cognitive relevance**: Extent by which the state of knowledge and the cognitive information need of the user is covered by the document. Highly personal and subjective, related to information need.

Types of Relevance: Saracevic (II)

4. **Situational relevance**: Extent by which the situation, task, or problem at hand of the user is covered by the document. Highly personal and subjective, related to work task.
5. **Motivational/affective relevance**: Extent by which the intents, goals and motivations of the user are covered by the document. Highly personal, subjective, or even emotional.

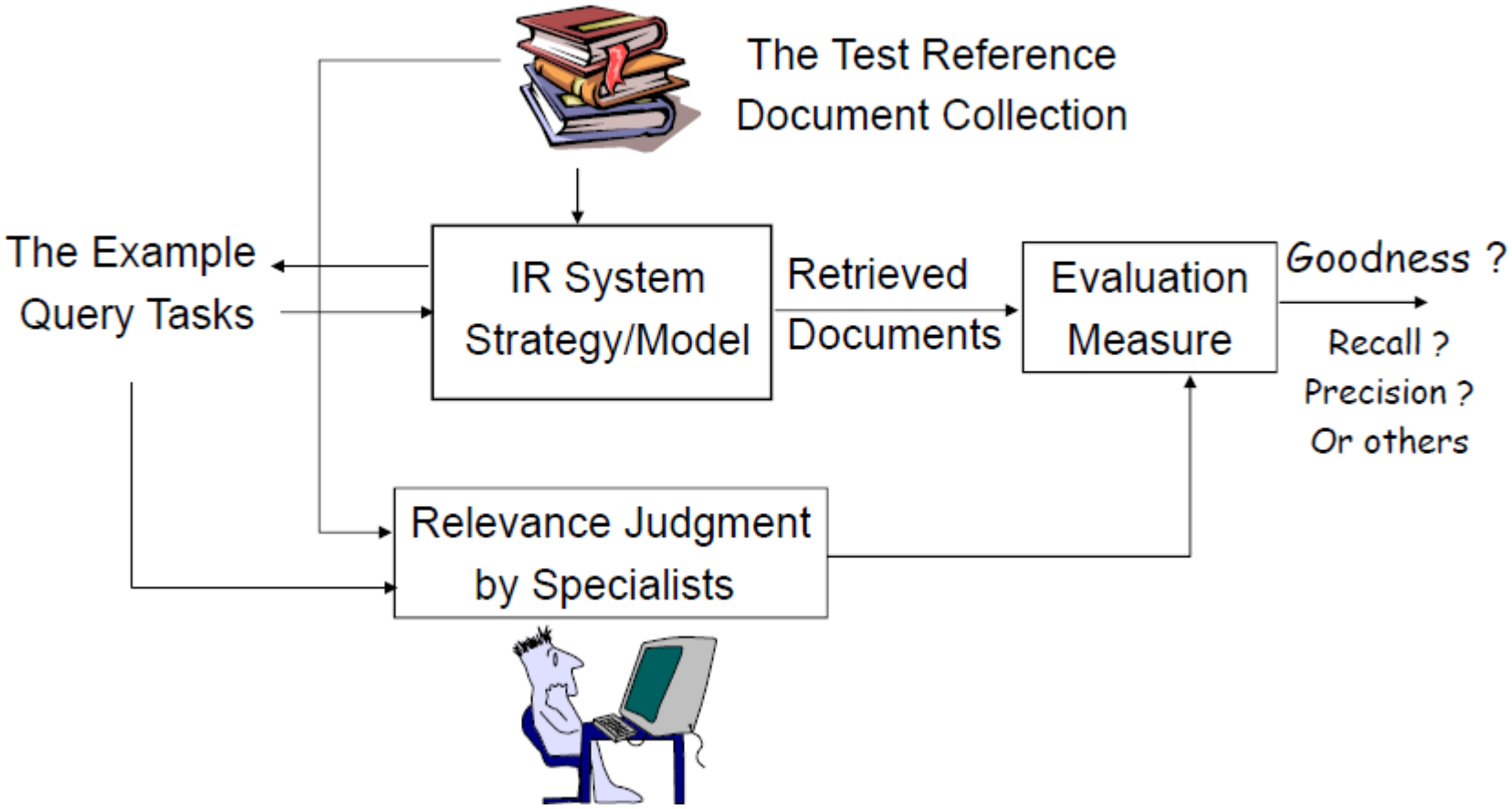
Introduction

Batch and Interactive Mode

- Batch mode (laboratory experiments)
  - The user submits a query and receives an answer back
  - Measure: the quality of the generated answer set
  - Still the dominant evaluation (Discussed here !)
    - Main reasons: repeatability and scalability
- Interactive mode (real life situations)
  - The user specifies his information need through a series of interactive steps with the system
  - Measure: user effort, interface design, system's guidance, session duration, or the context in which the query is posed
  - Get a lot more attention since 1990s

Introduction

A pictorial representation



Introduction

System-oriented evaluation

- Test collection
  - Benchmark (data set) upon which effectiveness is measured and compared
  - Data that tell us for a given query what are the relevant documents.
- Measuring effectiveness has been the most predominant in IR evaluation
  - **recall** of the system: proportion of **relevant** documents retrieved
  - **precision** of the system: proportion of the retrieved documents that are actually **relevant**
- Looking at these two aspects is part of what is called **system-oriented evaluation**

Test Collections

Components and why do we need them?

- Typical components:
  - Document collection, i.e. the document themselves. This depends on the task, e.g. evaluating web retrieval requires a collection of HTML documents
  - Queries / requests, which simulate real-user information needs
  - Relevance judgements, i.e. stating for a query the relevant documents

Test Collections

Components and why do we need them?

- To compare the performance of different retrieval techniques
  - each technique used to evaluate test queries
  - results (set or ranked list) compared using some performance measure
  - most common measures -precision and recall
- Usually use multiple measures to get different views of performance
- Usually test with multiple collections as performance is collection dependent

Test Collections

Finding the relevant documents

- Question: did the system find all relevant material?
- To answer accurately, collection needs complete judgements
  - i.e., yes, no or some score for every query-document pair
- For small test collections we can review all documents for all queries
- Not practical for large collections
  - TREC collections have millions of documents
- Pooling method

Test Collections

Relevance judgements creation

- Manual method:
  - Every document in the collection is judged against every query by one of several judges (human assessors)
  - This is feasible for small document collection.
- Pooling method (used for large document collection):
  - The queries are run against several IR systems first
  - The top, for example 100, documents retrieved by each system are pooled together.
  - The pool is then judged for relevance (by human assessors)

Test Collections

TREC (Text REtrieval Conference)

- Text REtrieval Conference
- Established in 1992 to evaluate large-scale IR
  - Retrieving documents from a gigabyte collection
- Has run continuously since then
- Run by NIST (National Institute of Standards and Technology) Information Access Division
- Started with 25 participating organisations; now many groups from many countries
- Proceedings online http://trec.nist.gov

Test Collections

TREC Conference

- Participants
  - Industrial, commercial and academic
  - Must submit results of retrieval tasks to the TREC conference each November
- Tracks
  - Ad-hoc retrieval (standard pull scenario)
  - Routing (filtering, push scenario)
  - There are other tracks, changing over the year: interactive, cross-lingual, web, speech, short queries, video,
  - question-answering (factoids), blog, genomic, . . .

Test Collections

Example (excerpt) of a TREC document

```
<doc>
<docno> WSJ880406-0090 </docno>
<hl> AT&T Unveils Services to Upgrade Phone Networks
Under Global Plan </hl>
<author> Janet Guyon (WSJ Sta) </author>
<dateline> New York </dateline>
<text>
American Telephone & Telegraph Co. introduced the
   rest of a new generation of phone services with
   broad ...
</text>
</doc>
```

Test Collections

Example (excerpt) of a TREC topic

```
<top>
<num> Number: 168 </docno>
<title> Topic: Financing AMTRAK
<desc> Description
A document will address the role of the Federal Government
    in financing the operation of the National Railroad
    Transportation Corporation (AMTRAK)
<nar> Narrative:
A relevant document must provide information on the
    government's responsibility to make AMTRAK an
    economically viable entity. It could also discuss the
    privatisation of AMTRAK. Documents comparing government
    subsidies given to air and bus transportation with those
    provided to AMTRAK would also be relevant.
</top>
```

Precision and Recall

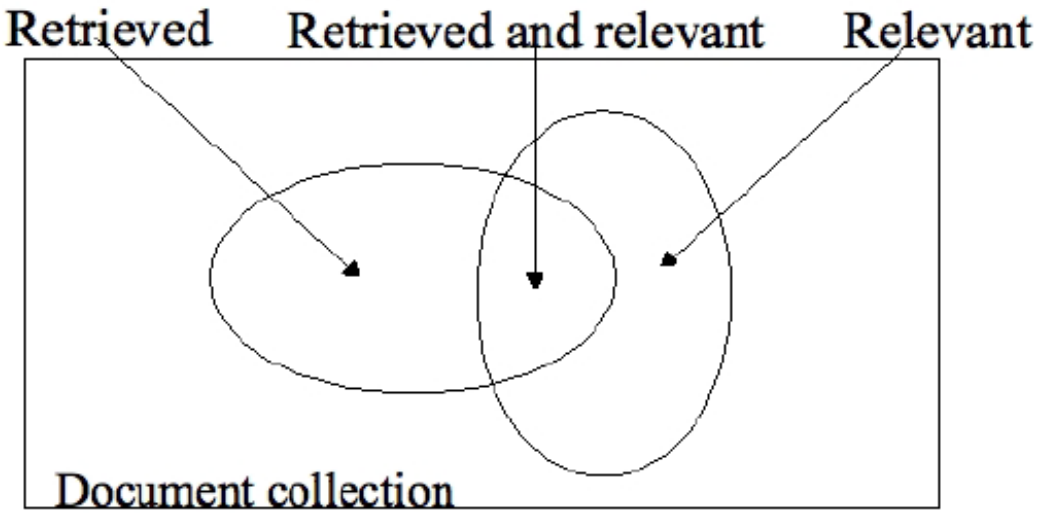
Effectiveness

- We recall that the goal of an IR system is to retrieve as many relevant documents as possible and as few non-relevant documents as possible.
- Evaluating the above consists of a comparative evaluation of technical performance of IR system(s):
  - In traditional IR, technical performance means the effectiveness of the IR system: the ability of the IR system to retrieve relevant documents and suppress non-relevant documents
  - Effectiveness is measured by the combination of **recall** and **precision**

Precision and Recall

Recall / Precision

For a given query, the document collection can be divided into three sets: the set of retrieved documents, the set of relevant documents, and the rest of the documents.



Note: knowing which documents are relevant comes from the test collection

Precision and Recall

Recall / Precision

In the ideal case, the set of retrieved documents is equal to the set of relevant documents. However, in most cases, the two sets will be different. This difference is formally measured with precision and recall.

$$Precision = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$$

$$Recall = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}$$

Precision and Recall

Recall / Precision

We can say that Precision corresponds to **correctness** and Recall to **completeness**.

Users in different scenarios may have different preference between the two measures. For example:

- Precision-oriented searches dominate e.g. web search. In general, in web searches users want some relevant document high in the ranked list of results (but this depends on the search task, etc.)
- Recall-oriented searches are important in domains where completeness of relevant documents is important, e.g. in the legal domain, in cases of considering patents for filing, etc. - in cases where it is bad to overlook some relevant items.

Precision and Recall

Recall / Precision

$$Precision = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$$

$$Recall = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}$$

The above two measures do not take into account where the relevant documents are retrieved, i.e. at which rank.

This is very important because an effective IR system should not only retrieve as many relevant documents as possible and as few non-relevant documents as possible, but also it should retrieve relevant documents **before** the non-relevant ones.

Precision and Recall

Recall / Precision

- Let us assume that for a given query, the following documents are relevant (10 relevant documents)

  {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

- Now suppose that the following documents are retrieved for that query:

rank	doc	precision	recall	rank	doc	precision	recall
1	d123	1/1	1/10	8	d129		
2	d84			9	d187		
3	d56	2/3	2/10	10	d25	4/10	4/10
4	d6			11	d48		
5	d8			12	d250		
6	d9	3/6	3/10	13	d113		
7	d511			14	d3	5/14	5/10

Precision and Recall

Recall / Precision

- For each relevant document (in red bold), we calculate the precision value and the recall value. For example, for d56, we have 3 retrieved documents, and 2 among them are relevant, so the precision is 2/3. We have 2 of the relevant documents so far retrieved (the total number of relevant documents being 10), so recall is 2/10.

rank	doc	precision	recall	rank	doc	precision	recall
1	**d123**	1/1	1/10	8	d129		
2	d84			9	d187		
3	**d56**	2/3	2/10	10	**d25**	4/10	4/10
4	d6			11	d48		
5	d8			12	d250		
6	**d9**	3/6	3/10	13	d113		
7	d511			14	**d3**	5/14	5/10

Precision and Recall

Recall / Precision

- For each query, we obtain pairs of recall and precision values
  - In our example, we would obtain (1/10, 1/1) (2/10, 2/3) (3/10, 3/6) (4/10, 4/10) (5/10, 5/14) . . . which are usually expressed in % (10%, 100%) (20%, 66.66%) (30%, 50%) (40%, 40%) (50%, 35.71%) . . .
  - This can be read for instance: at 20% recall, we have 66.66% precision; at 50% recall, we have 35.71% precision

rank	doc	precision	recall	rank	doc	precision	recall
1	d123	1/1	1/10	8	d129		
2	d84			9	d187		
3	d56	2/3	2/10	10	d25	4/10	4/10
4	d6			11	d48		
5	d8			12	d250		
6	d9	3/6	3/10	13	d113		
7	d511			14	d3	5/14	5/10

Precision and Recall

Averaging Recall/Precision values

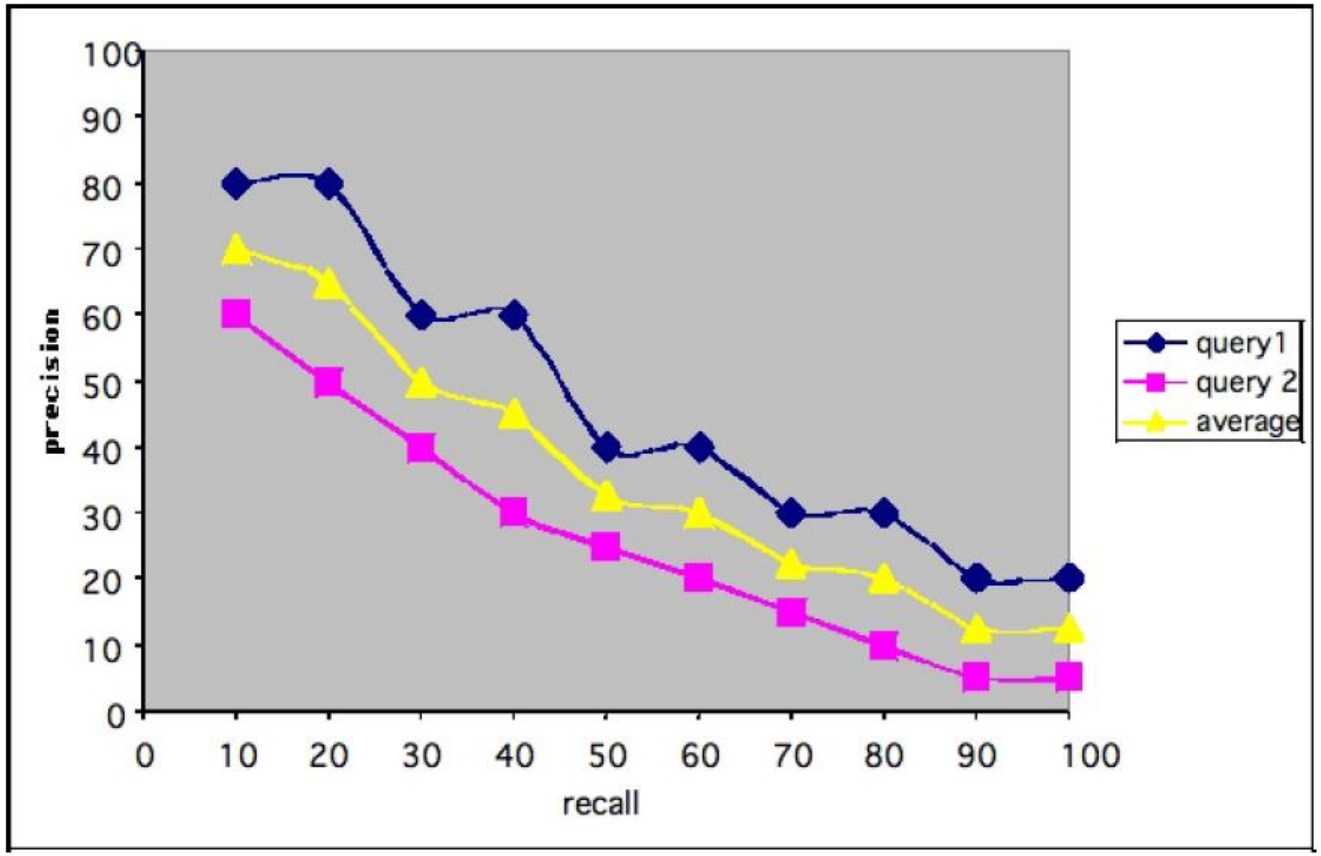
- Hard to compare precision and recall graphs or tables for individual queries (too much data)

Recall in %	Precision in %		
	Query 1	Query 2	Average
10	80	60	70
20	80	50	65
30	60	40	50
40	60	30	45
50	40	25	32.5
60	40	20	30
70	30	15	22.5
80	30	10	20
90	20	5	11.5
100	20	5	11.5

Precision and Recall

Averaging II

- Interpolated precision-recall curves



Precision and Recall

Recall-Precision Curve

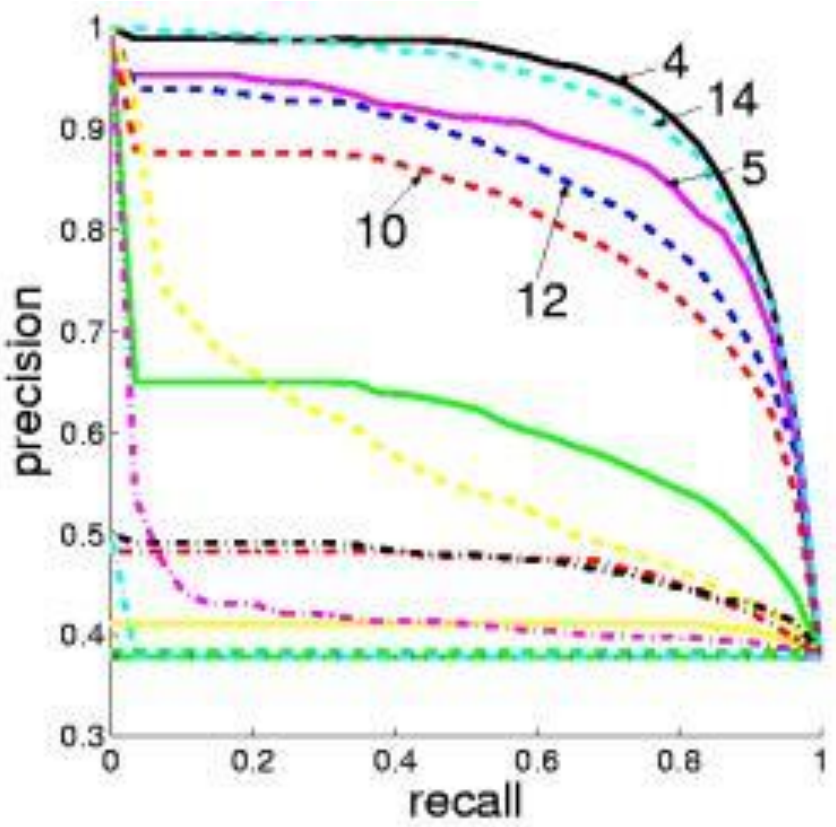
- Advantages
  - Simple, intuitive, and combined in single curve
  - Provide quantitative evaluation of the answer set and comparison among retrieval algorithms
  - A standard evaluation strategy for IR systems

- Disadvantages
  - Can't know true recall value except in small document collections
  - Assume a strict document rank ordering

Precision and Recall

Comparison of systems

- We can now compare IR systems (or versions)



Precision and Recall

The complete methodology

To evaluate an IR system (or version):

1. For each query in the test collection
    1) Run the query against the system to obtain a ranked list of retrieved documents
    2) Use the ranking and relevance judgements to calculate recall/precision pairs
2. Average recall / precision values across all queries to obtain an overall measure of the effectiveness

Precision and Recall

Problem with the test collection methodology

- Building larger test collections along with complete relevance judgement is difficult or impossible
  - require assessor time - very expensive
  - Require many diverse retrieval runs.
- Recall is difficult if not impossible to get correctly as there is no way we can find all the relevant documents for each query
- Issues:
  - Non-judged documents are assumed non-relevant
  - Can we reuse the test collection later on?

Single Value Summaries

- Interpolated recall-precision curve
  - Compare the performance of retrieval algorithms over a set of example queries
    - Might disguise the important anomalies
  - How is the performance for each individual query ?

- A single precision value (for each query) is used instead
  - Interpreted as a summary of the corresponding precision versus recall curve
    - Just evaluate the precision based on the top 1 relevant document ?
    - Or averaged over all relevant documents

Single Value Summaries

Precision at document cut-off level

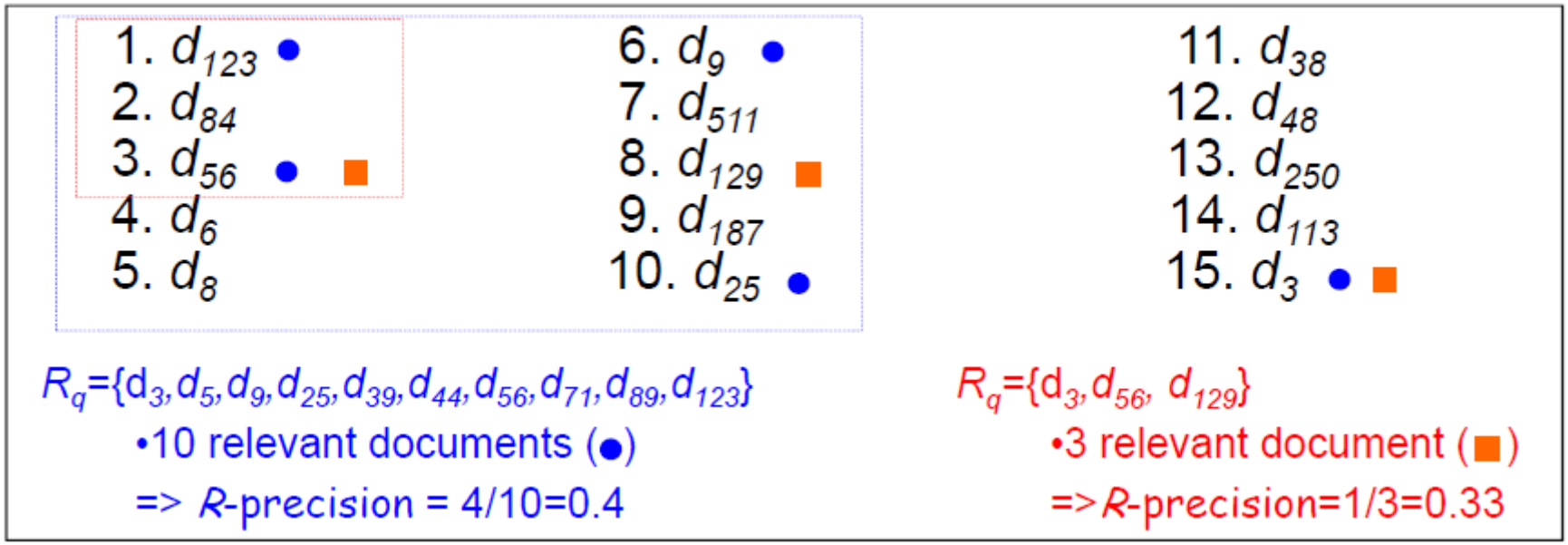
Document Level Average	
Recall	Precision
At 5 documents	0.42
At 10 documents	0.40
...	...
At 1000 documents	0.05

- Actual performance as a user might see it
- often used in web retrieval, usually written P@i; e.g. P@10

Single Value Summaries

R-Precision

- Generate a single value summary of ranking by computing the precision at the R-th position in the ranking
  - Where R is the total number of relevant docs for the current query



Single Value Summaries

Mean Average Precision (mAP)

- Averaged at relevant docs and across queries
  - E.g. relevant docs ranked at 1, 5, 10, precisions are 1/1, 2/5, 3/10,
    - non-interpolated average precision
  - Mean Average Precision (denoted as mAP or MAP)

$$\frac{1}{|Q|}\sum_{q=1}^{|Q|}(\text{non}-\text{interpoloated average precision})_q$$

  - Widely used in IR performance evaluation

Alternative Measures

The Harmonic Mean (F Measure)

- The harmonic mean F of recall and precision
  - r(j): the recall for the j-th document in the ranking
  - P(j): the precision for the j-th document in the ranking

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} = \frac{2 \cdot P(j) \cdot r(j)}{P(j) + r(j)}$$

- Characteristics
  - F = 0: no relevant documents were retrieved
  - F = 1: all ranked documents are relevant
  - A high F achieved only when both recall and precision are high
  - Determination of the maximal F can be interpreted as an attempt to find the best possible compromise between recall and precision

Alternative Measures

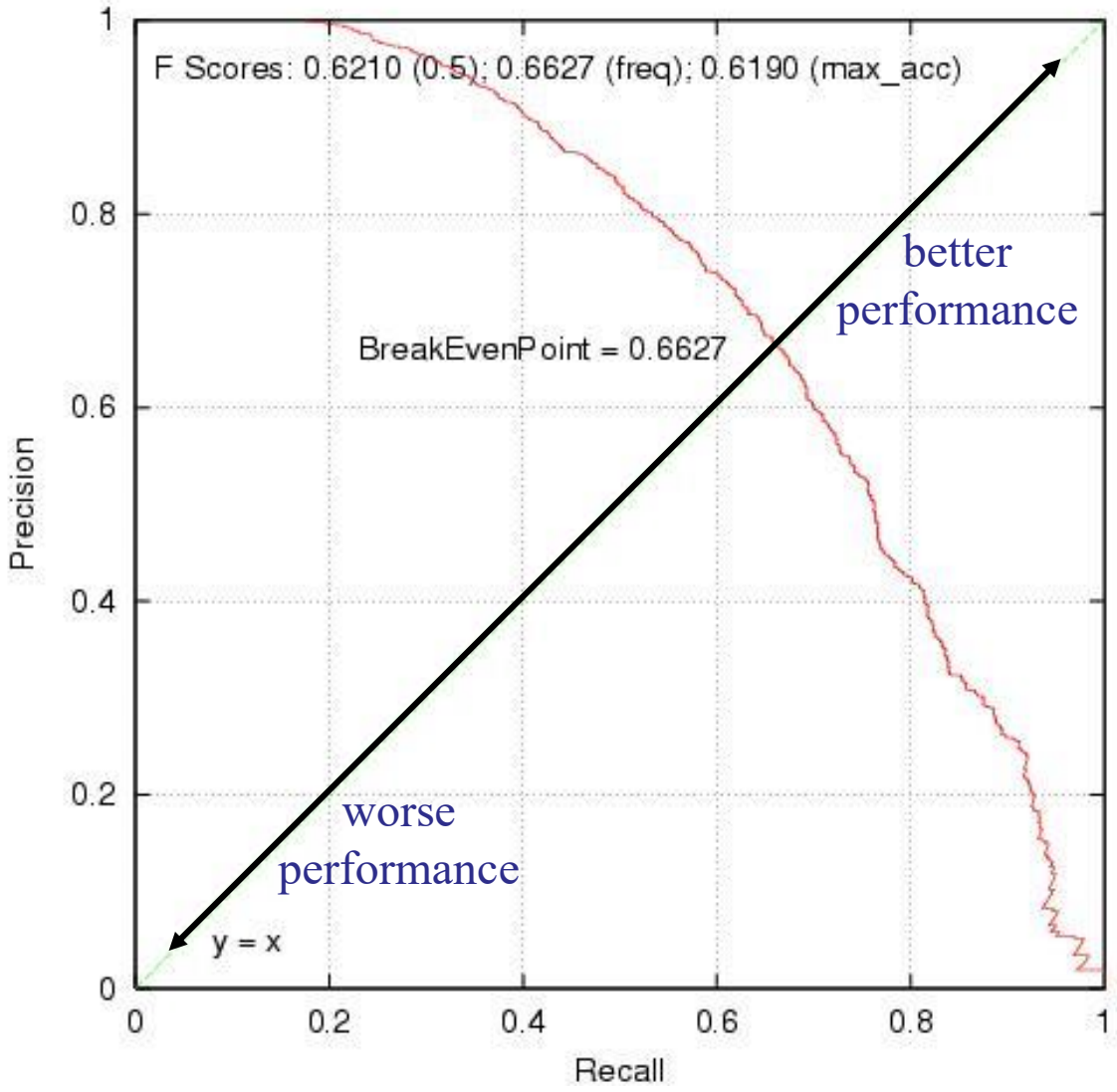
Break-even Point

$$F = \frac{2 * (PRECISION \times RECALL)}{(PRECISION + RECALL)}$$

harmonic average of precision and recall

$$BreakEvenPoint = PRECISION = RECALL$$

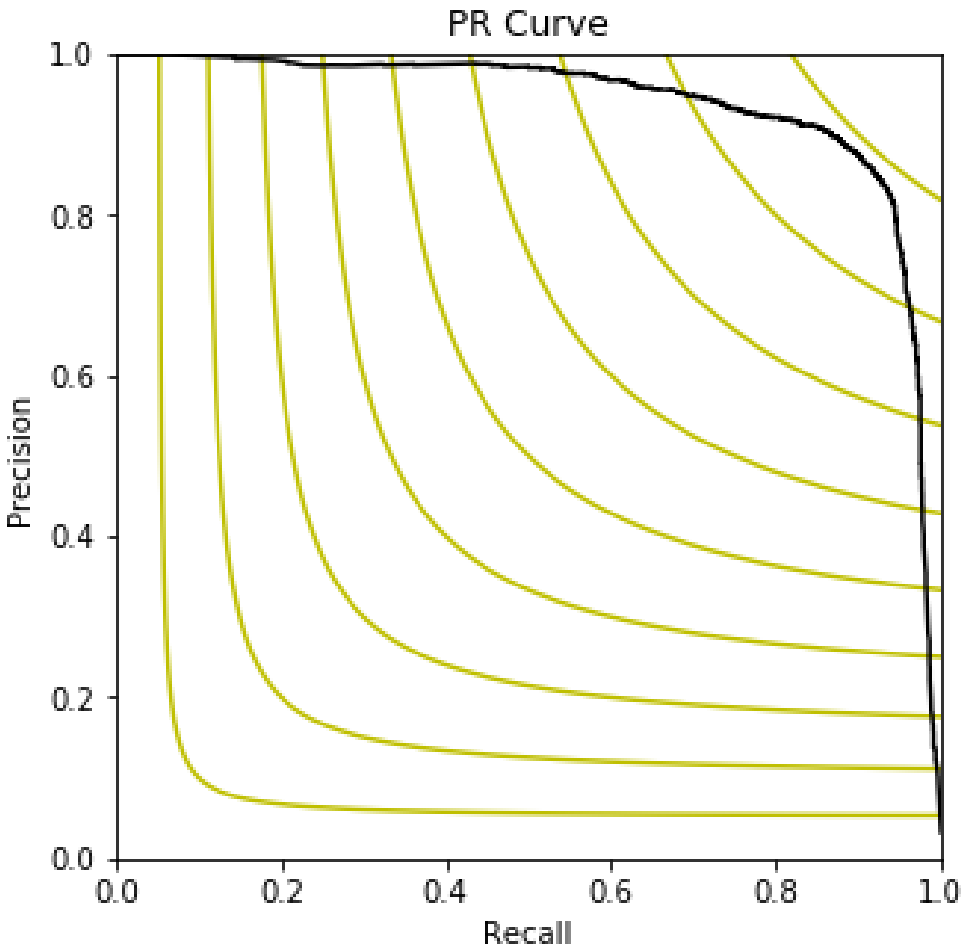
Alternative Measures



Use the best F score to represent this P-R curve:

the break-even point

Alternative Measures



You can calculate F1-score for each point on precision-recall plane and for each value of F1-score you will have multiple satisfying points. Therefore, you can plot level curves of F1-score - in the following example yellow lines represent level curves of F1-score from 0.1 to 0.9

Alternative Measures

The E Measure

- Another measure which combines recall and precision
- Allow the user to specify whether he/she is more interested in recall or precision

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}} = 1 - \frac{(1 + b^2) \cdot P(j) \cdot r(j)}{b^2 \cdot P(j) + r(j)}$$

- Characteristics
  - b = 1: act as the complement of F Measure
  - b > 1: more interested in recall
  - b < 1: more interested in precision

Alternative Measures

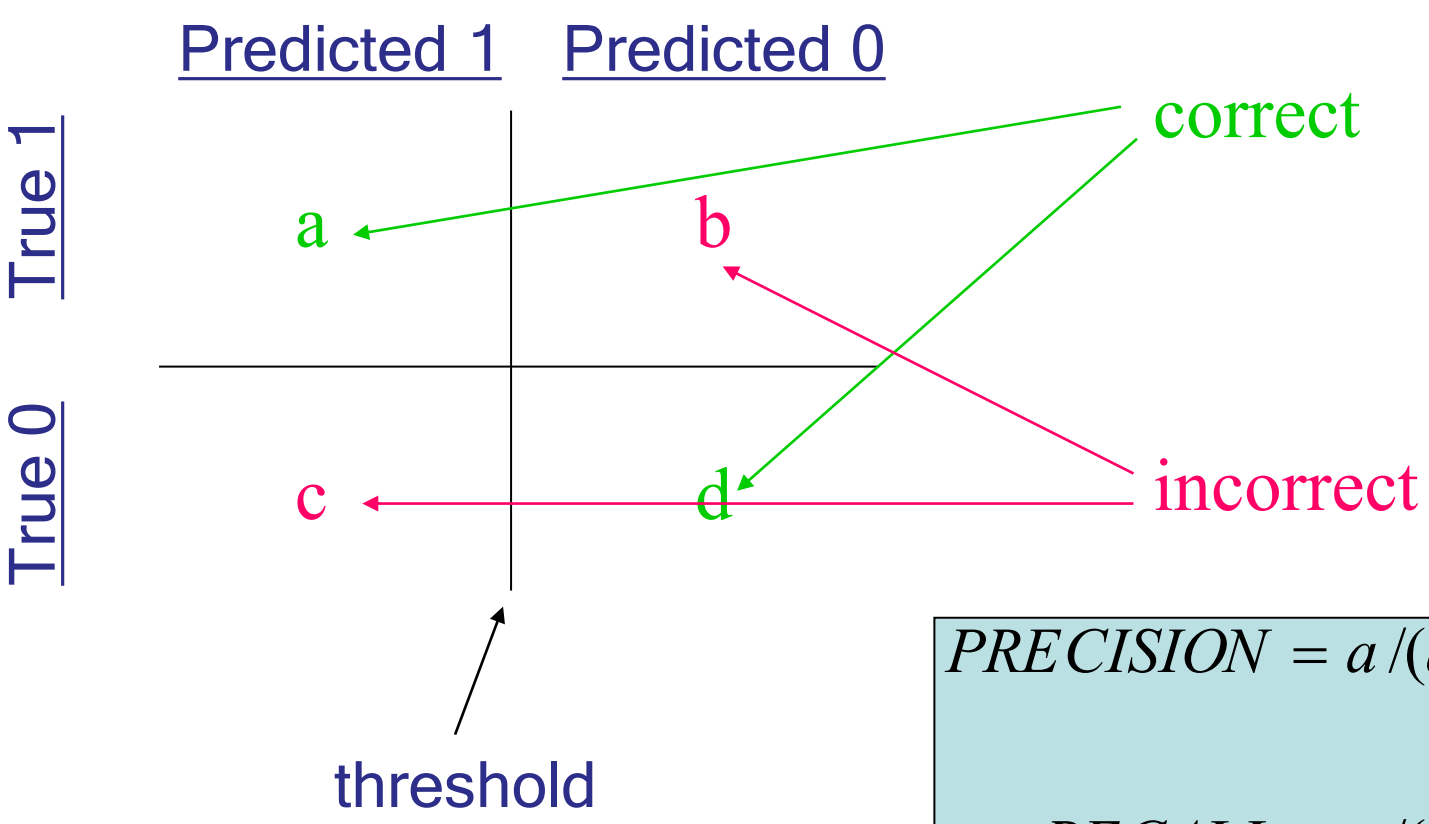
Receiver Operating Characteristics

ROC Plot and ROC Area

- Developed in WWII to statistically model false positive and false negative detections of radar operators
- Better statistical foundations than most other measures
- Standard measure in medicine and biology
- Becoming more popular in ML

Alternative Measures

Confusion matrix



a: TP
b: FN
c: FP
d: TN

$$PRECISION = a/(a+c)$$

$$RECALL = a/(a+b)$$

Alternative Measures

ROC Plot

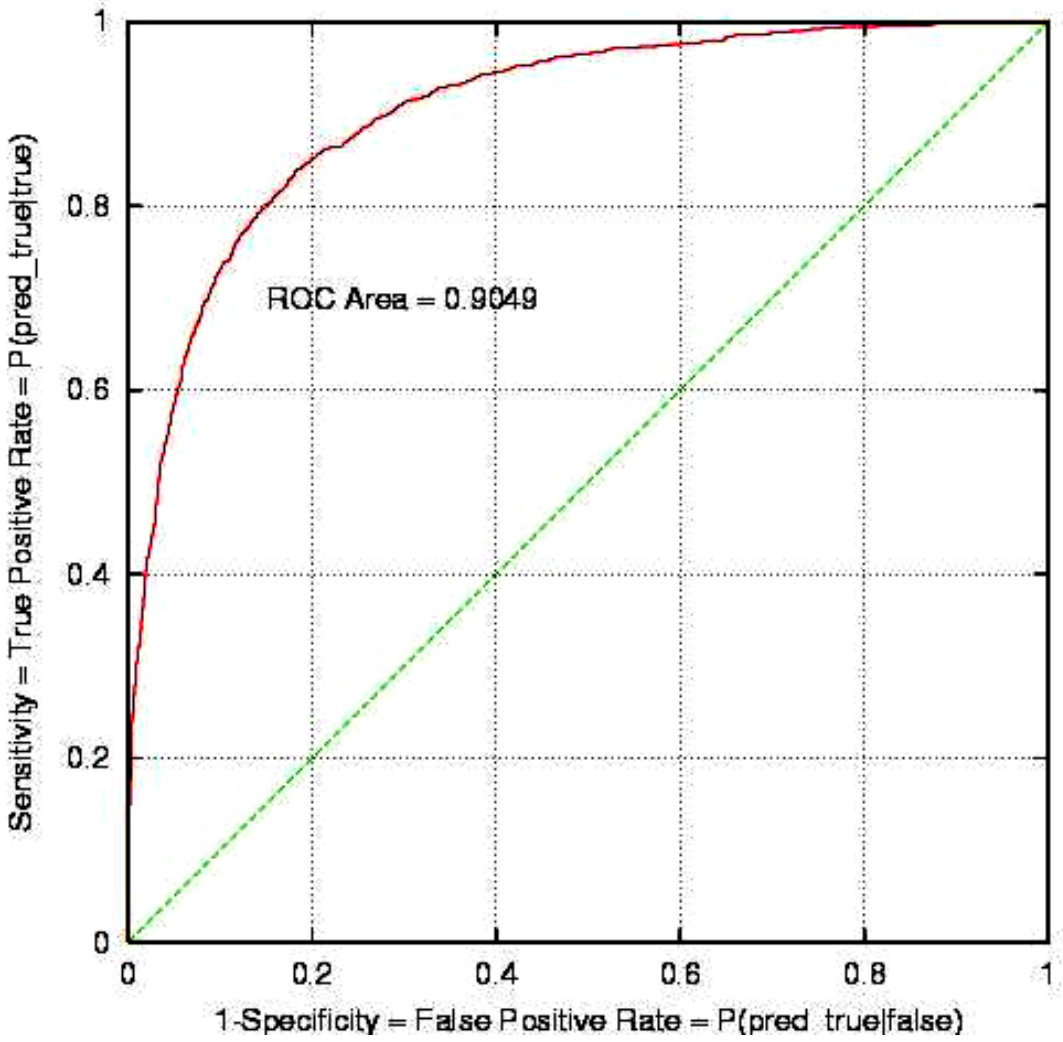
- Plot
  - TP rate vs. FP rate
  - P(true|true) vs. P(true|false)
- Sensitivity = a/(a+b) = Recall
- 1 - Specificity = 1 - d/(c+d)=1-TN/(FP+TN)
- Sensitivity vs. 1-Specificity

Alternative Measures

Properties of ROC

ROC Area:

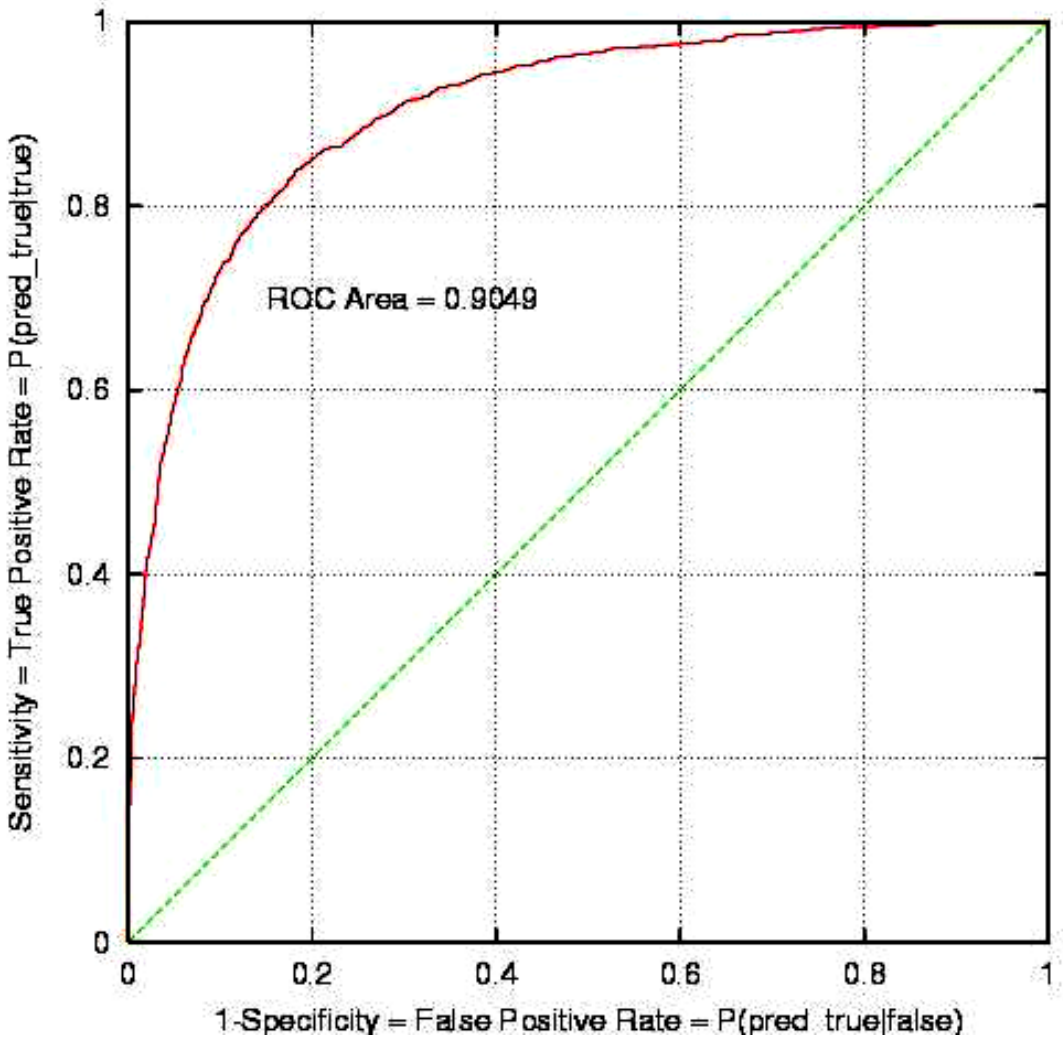
- 1.0: perfect prediction
- 0.9: excellent prediction
- 0.8: good prediction
- 0.7: mediocre prediction
- 0.6: poor prediction
- 0.5: random prediction
- <0.5: something wrong!



Alternative Measures

Properties of ROC

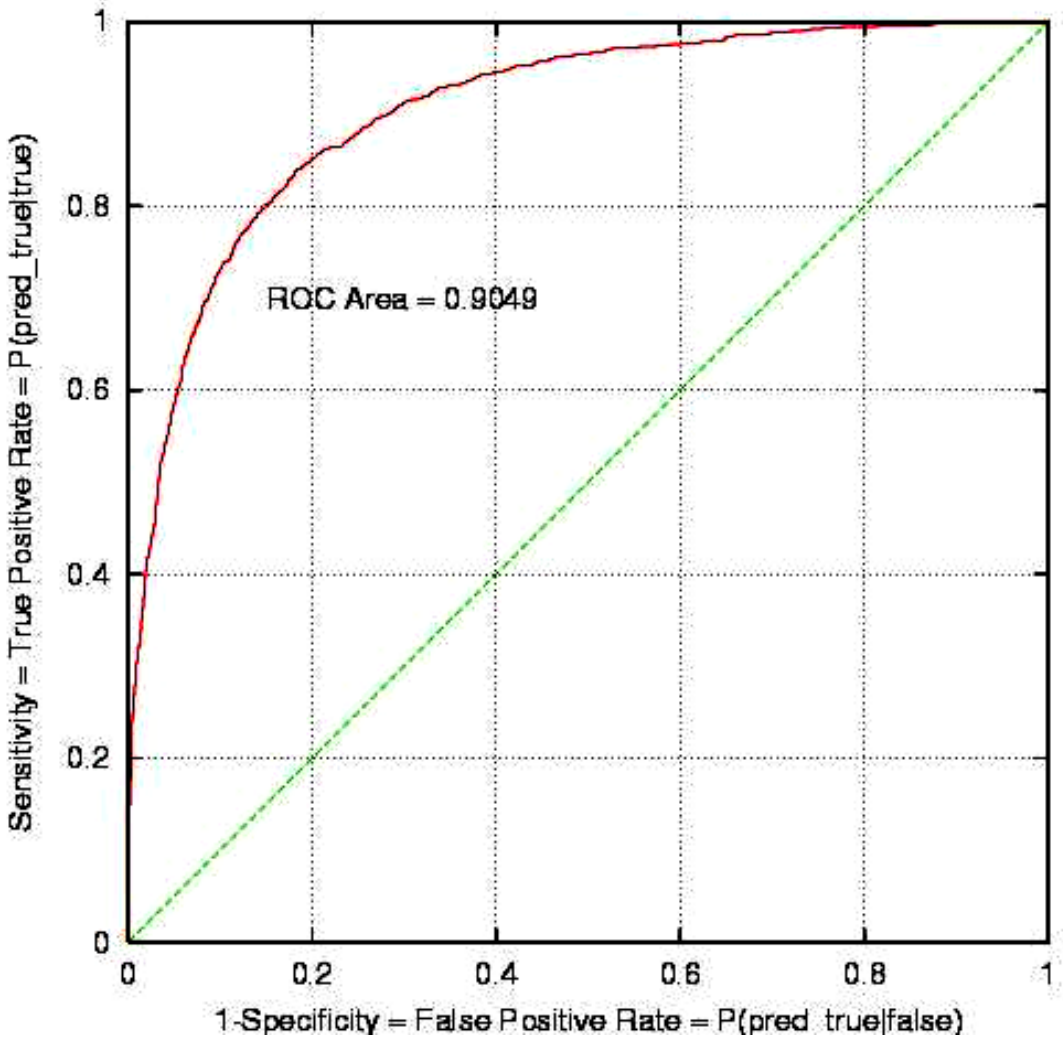
- Slope is non-increasing
- Each point on ROC represents different tradeoff (cost ratio) between false positives and false negatives
- Slope of line tangent to curve defines the cost ratio
- ROC Area represents performance averaged over all possible cost ratios



Alternative Measures

Properties of ROC

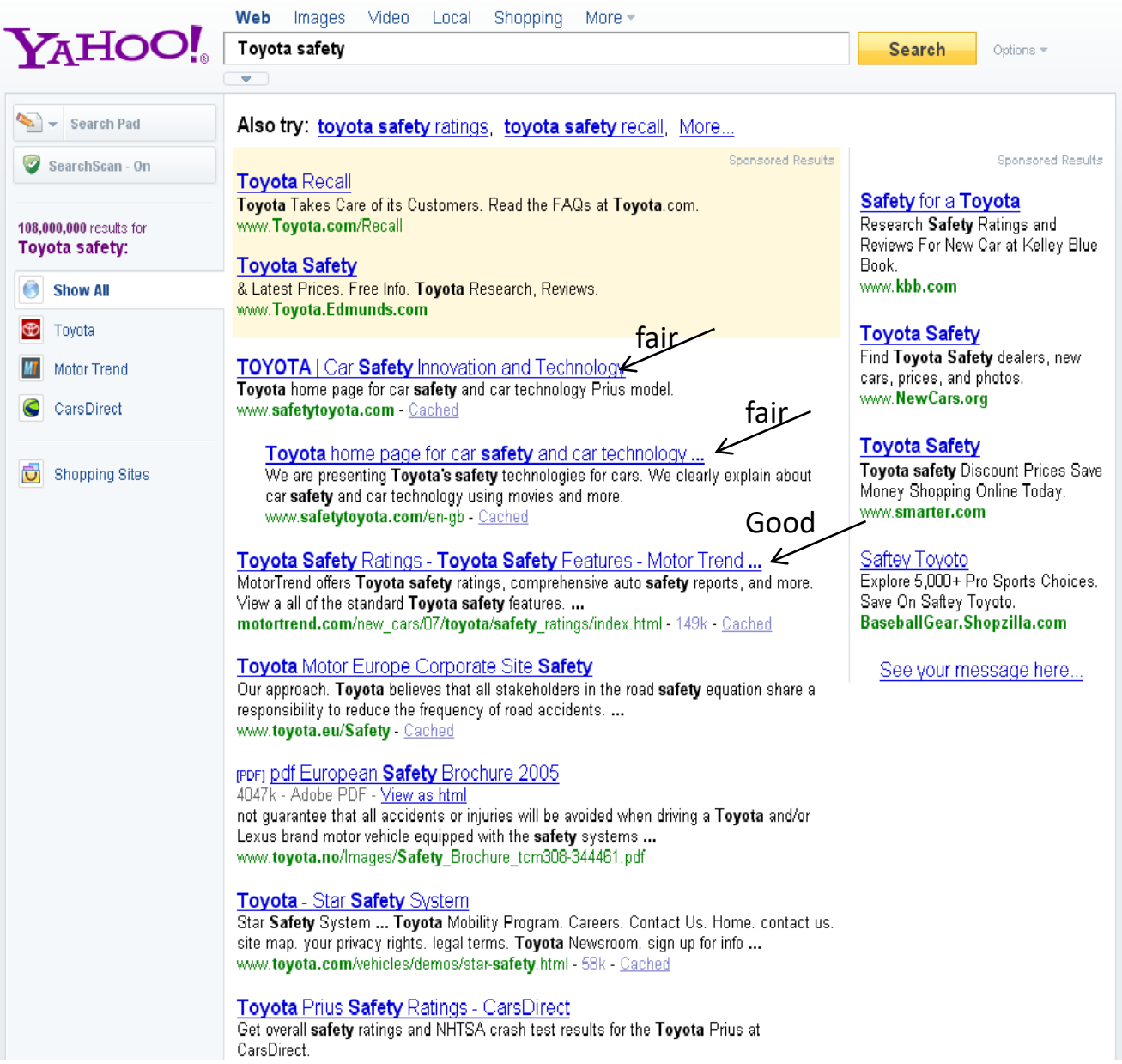
- If two ROC curves do not intersect, one method dominates the other
- If two ROC curves intersect, one method is better for some cost ratios, and other method is better for other cost ratios



Beyond Binary Relevance

- Precision and recall allow only binary relevance assessments
- As a result, there is no distinction between highly relevant docs and mildly relevant docs
- These limitations can be overcome by adopting graded relevance assessments and metrics that combine them

Beyond Binary Relevance



Discounted Cumulative Gain

- The discounted cumulated gain (DCG) is a metric that combines graded relevance assessments effectively
- Popular measure for evaluating web search and related tasks

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/log *(rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

Summarize a Ranking: DCG

- What if relevance judgments are in a scale of [0,r]?

- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)
  - $CG_n = r_1+r_2+\ldots r_n$

- Discounted Cumulative Gain (DCG) at rank n
  - $DCG_n = r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$
    - We may use any base for the logarithm

Discounted Cumulative Gain

- *DCG* is the total gain accumulated at a particular rank *p*:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

Discounted Cumulative Gain

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- DG= $r_i / \log_2 i$ *(i>=2)*

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- DCG= $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$

  = 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Discounted Cumulative Gain

Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank *n*
  - Normalize DCG at rank *n* by the DCG value at rank *n* of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Normalization useful for contrasting queries with varying numbers of relevant results
- NDCG is now quite popular in evaluating Web search

Discounted Cumulative Gain

Summarize a Ranking: NDCG

- Perfect ranking:
  - 3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- ideal DCG values:
  - 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88
- Actual DCG for the actual ranking (3, 2, 3, 0, 0, 1, 2, 2, 3, 0):
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- NDCG values (divide actual by ideal):

1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

- NDCG $\leq$ 1 at any rank position

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Doc. Order	r_i	Doc. Order	r_i	Doc. Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309 \qquad DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309 \qquad MaxDCG = DCG_{GT} = 4.6309$$

Discounted Cumulative Gain

Discussion on DCG Metrics

- CG and DCG metrics aim at taking into account multiple level relevance assessments
- This has the advantage of distinguishing highly relevant documents from mildly relevant ones
- The inherent disadvantages are that multiple level relevance assessments are harder and more time consuming to generate

Discounted Cumulative Gain

Discussion on DCG Metrics

- Despite these inherent difficulties, the CG and DCG metrics present benefits:
  - They allow systematically combining document ranks and relevance scores
  - Cumulated gain (CG) provides a single metric of retrieval performance at any position in the ranking
  - It also stresses the gain produced by relevant docs up to a position in the ranking, which makes the metrics more immune to outliers
  - Further, discounted cumulated gain (DCG) allows down weighting the impact of relevant documents found late in the ranking

Trends and Research Issues

- A major trend today is research on interactive user interfaces and their evaluation
  - Which evaluation measures are most appropriate?
- Another important trend is **crowdsourcing**
  - Use the population of Web users to conduct well defined evaluation tasks in exchange for small sums of money
- Further, the proposal, the study, and the characterization of alternative measures to precision and recall continue to be of interest
  - Some specific scenarios are not well covered by precision-recall figures