Evaluation in information retrieval

Agenda

- Measuring the effectiveness of IR systems
- The test collections
- Evaluating unranked retrieval results
- Evaluating ranked retrieval results
- User utility
- Other measures of the quality of a system

Information retrieval system evaluation

- To measure effectiveness we need test collection consisting of three things:
  - A document collection
  - A test suite of information needs, expressible as queries
  - A set of relevance judgments, standardly a binary assessment of either *relevant* or *nonrelevant* for each query-document pair.
- Information retrieval system evaluation revolves around the notion of *relevant* and *nonrelevant* documents
- A document in the test collection is given a binary classification as either relevant or nonrelevant. This decision is referred to as the *gold standard* or *ground truth* judgment of relevance.
- The test document collection and suite of information needs have to be of a reasonable size
- Relevance is assessed relative to an information need, *not* a query
- Eg. Here is a list of the most standard test collections and evaluation series
- This might be translated into a query such as:

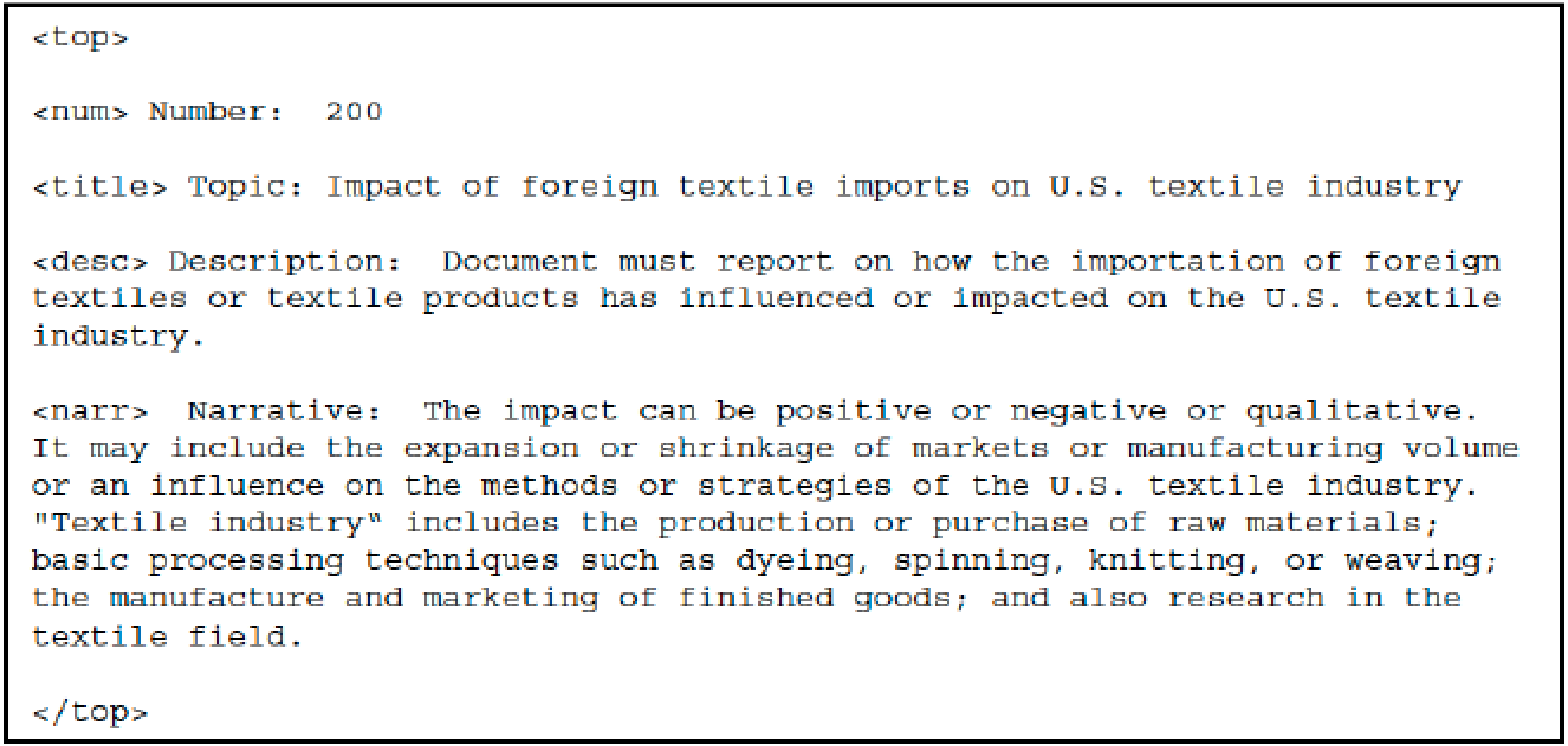
list AND most AND standard AND test AND collections AND evaluation AND series

- From a one word query, it is very difficult for a system to know what the information need is. Eg. Mouse
- Here we assume binary decision of relevance
- Many systems contain various weights (often known as parameters) that can be adjusted to tune system performance.
- It is wrong to report results on a test collection which were obtained by tuning these parameters to maximize performance on that collection.
- In such cases, the correct procedure is to have one or more *development test collections*, and to tune the parameters

Standard test collections

- The *Cranfield* collection: This was the pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness
- Nowadays too small for anything but the most elementary pilot experiments.
- Collected in the United Kingdom starting in the late 1950s, it contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgments of all (query, document)
- *Text Retrieval Conference (TREC)*. The U.S. National Institute of Standards and Technology (NIST) has run a large IR test bed evaluation series since 1992 pairs.
- In total, these test collections comprise 6 CDs containing 1.89 million documents and relevance judgments for 450 information needs, which are called *topics* and specified in detailed text passages.

Example TREC ad hoc topic

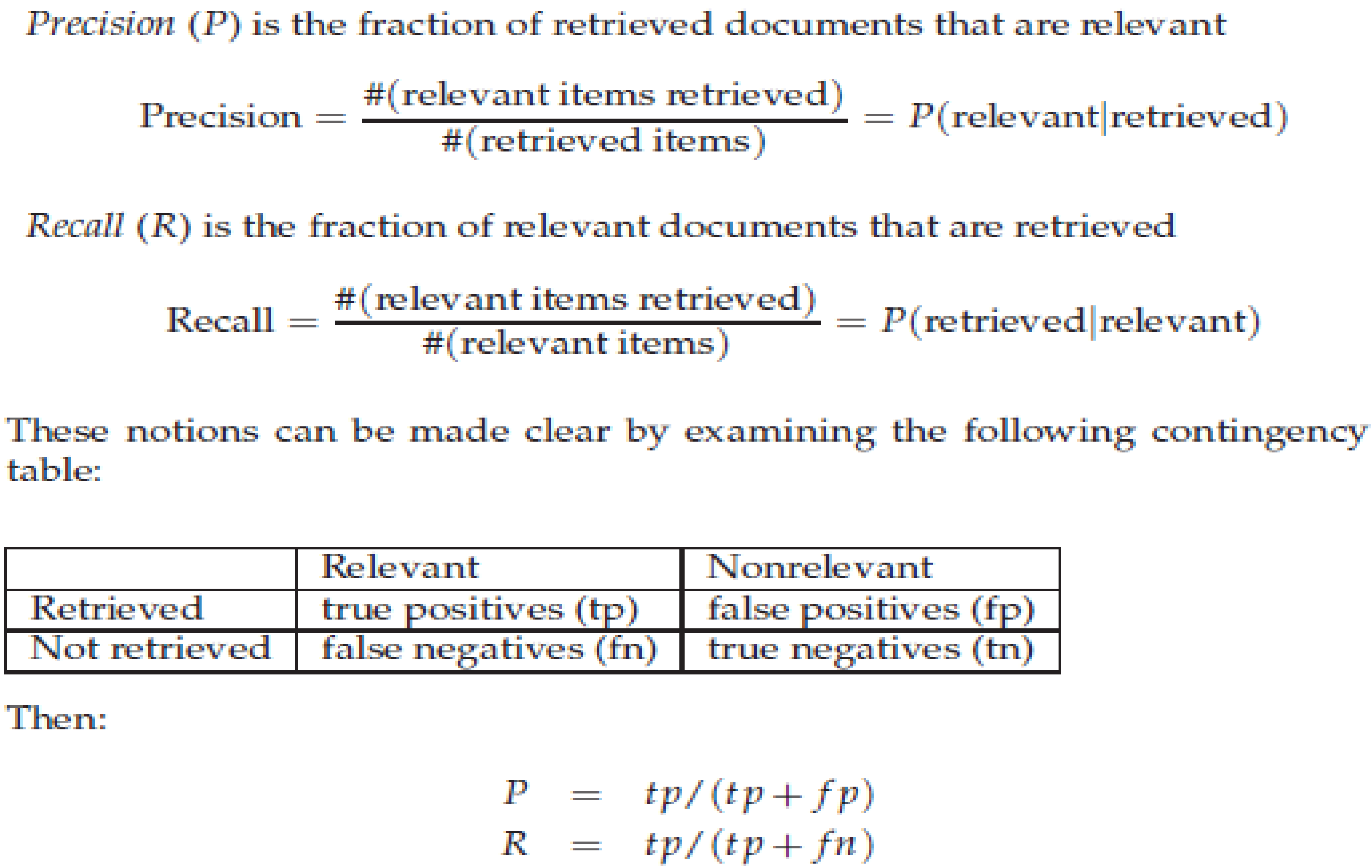


```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

- GOV2
    - Another TREC/NIST collection
    - 25 million web pages
    - Largest collection that is easily available
    - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
    - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
    - This evaluation series has concentrated on European languages and cross-language information retrieval.
- Many others

Evaluation of unranked retrieval sets

- The two most frequent and basic measures for information retrieval effectiveness are precision and recall.



Precision (P) is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Recall (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

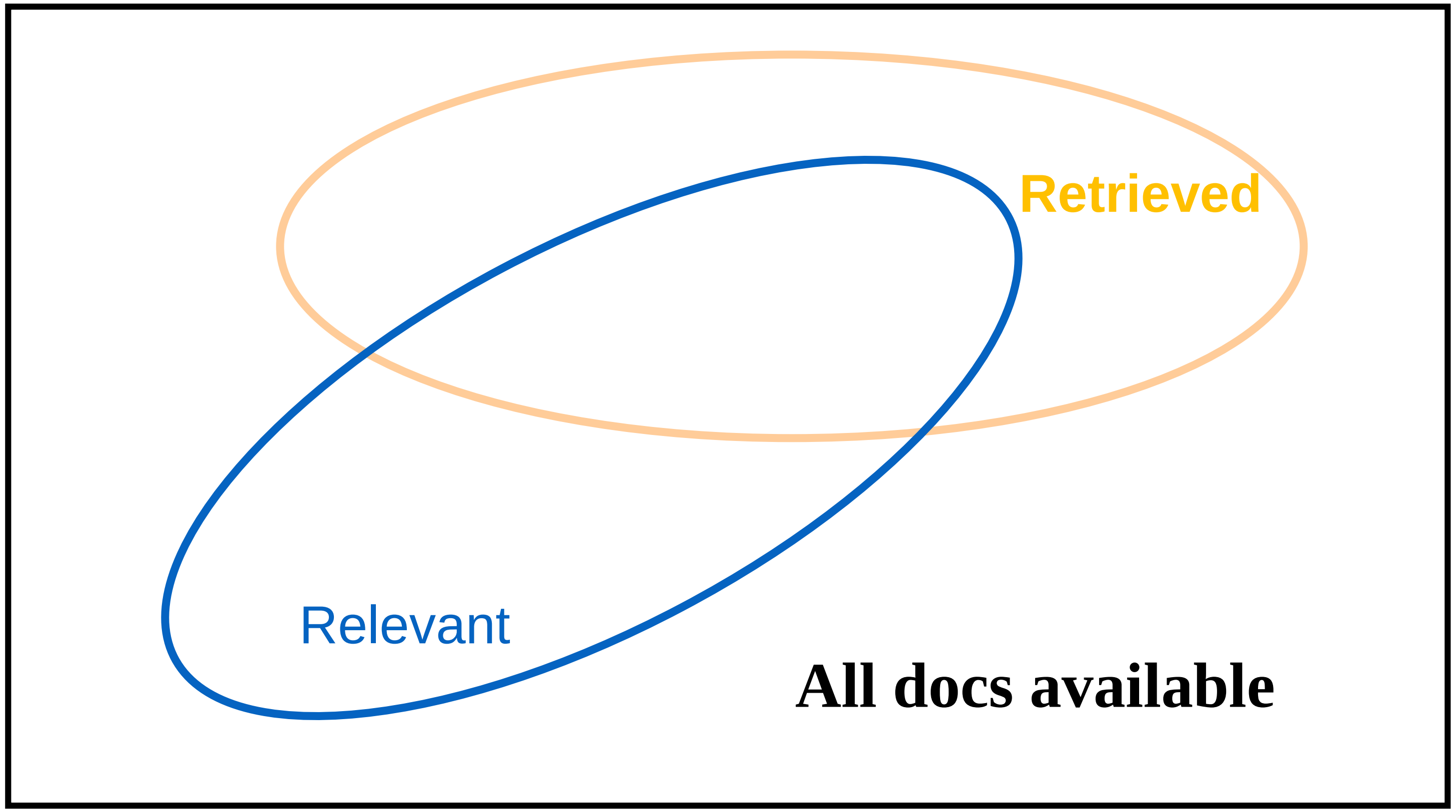
These notions can be made clear by examining the following contingency table:

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Then:

$$\begin{aligned} P &= tp/(tp+fp) \\ R &= tp/(tp+fn) \end{aligned}$$

Relevant vs. Retrieved Documents



Set theoretic approach

Example

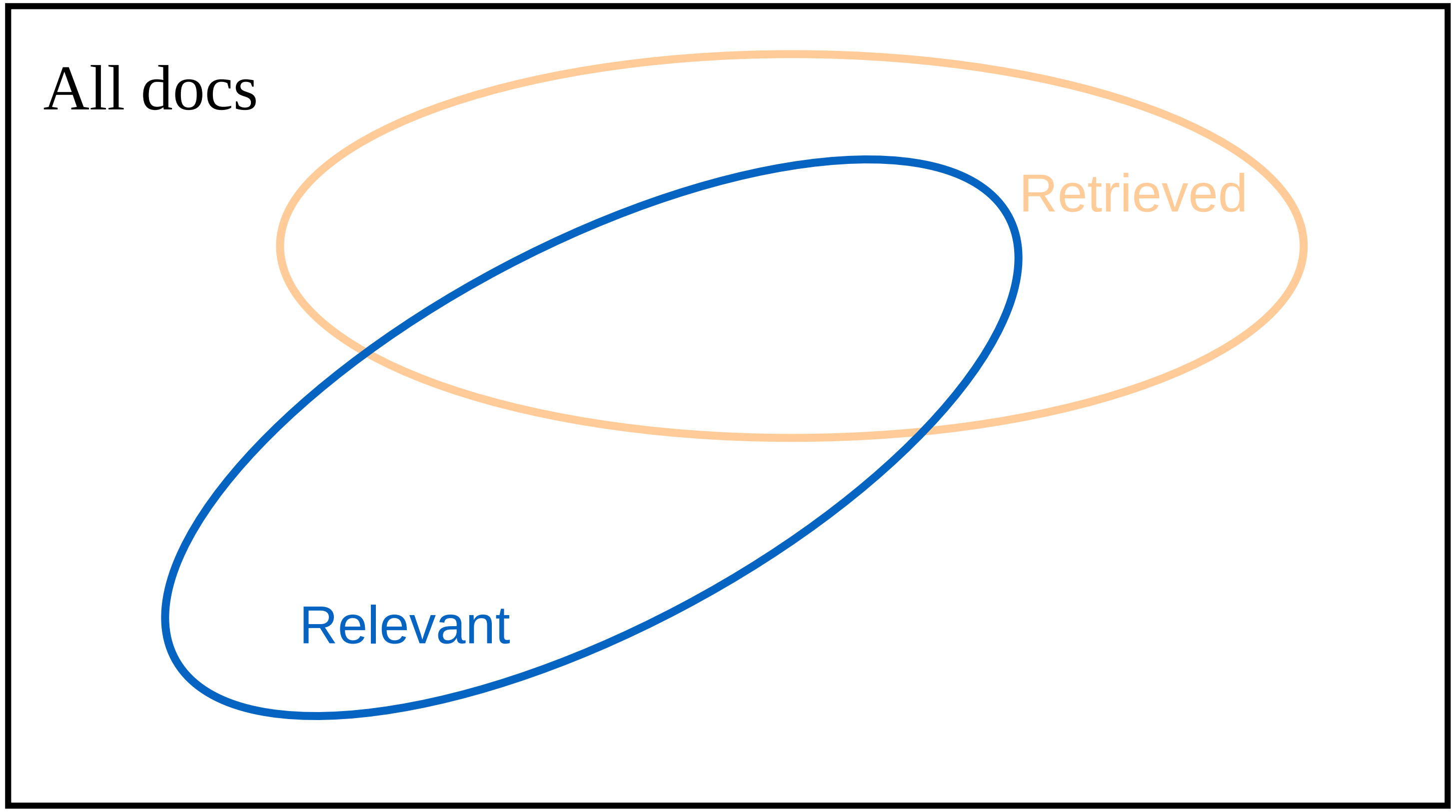
- Documents available: D1,D2,D3,D4,D5,D6,D7 ,D8,D9,D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

	relevant	not relevant
retrieved	D4,D5,D8	D2,D6,D9
not retrieved	D1,D10	D3,D7

Precision vs. Recall

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

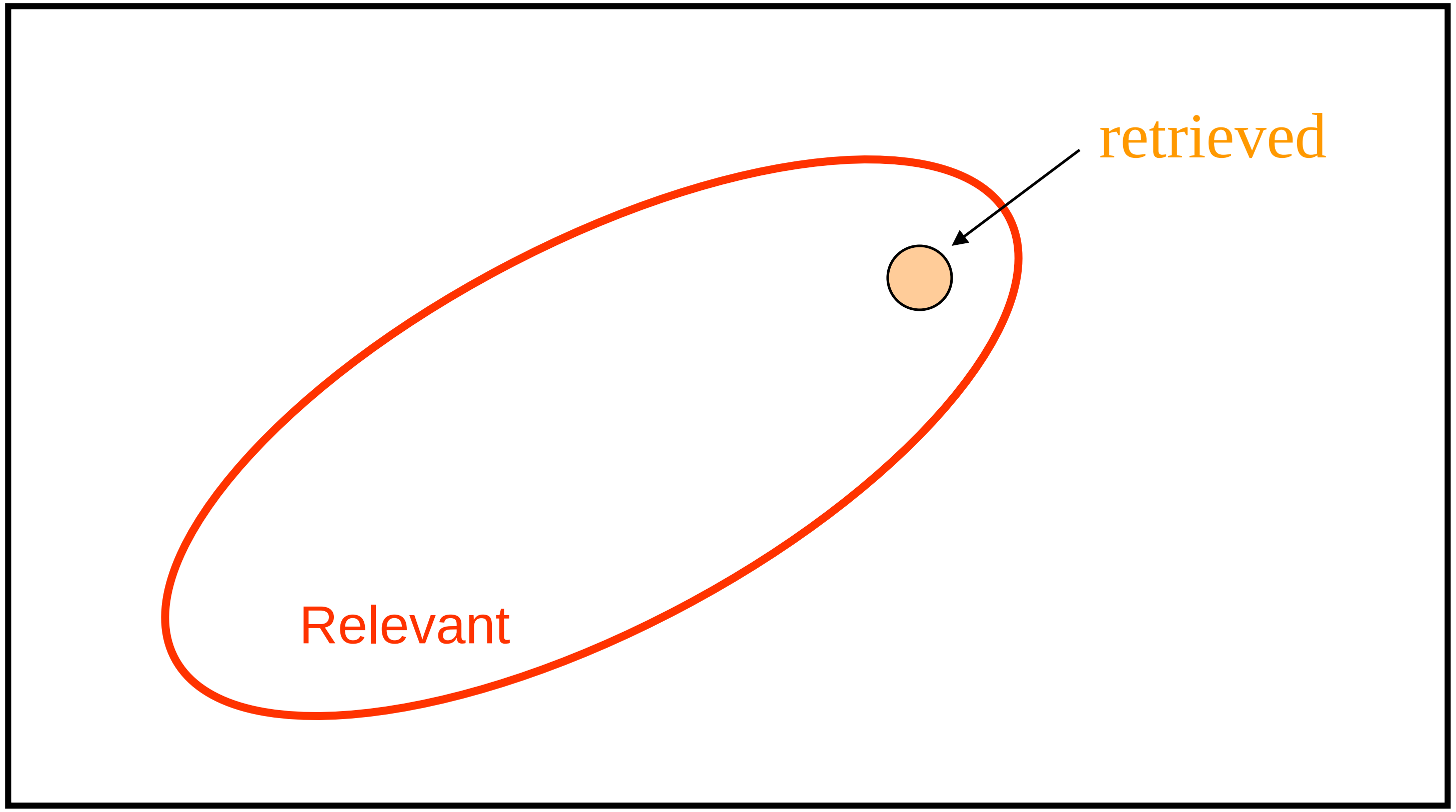


Precision can be seen as a measure of exactness.

Recall is a measure of completeness

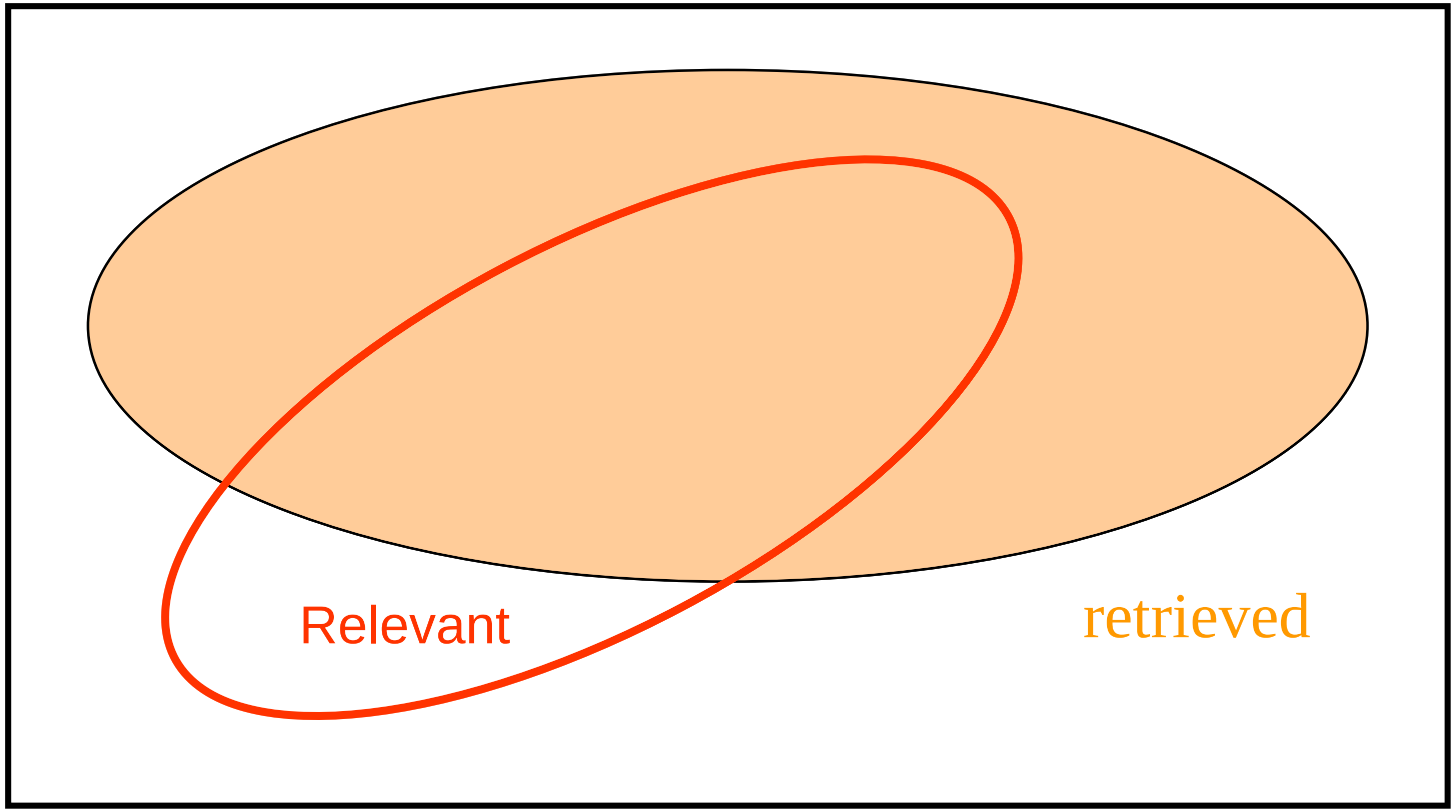
Retrieved vs. Relevant Documents

Very high precision, very low recall



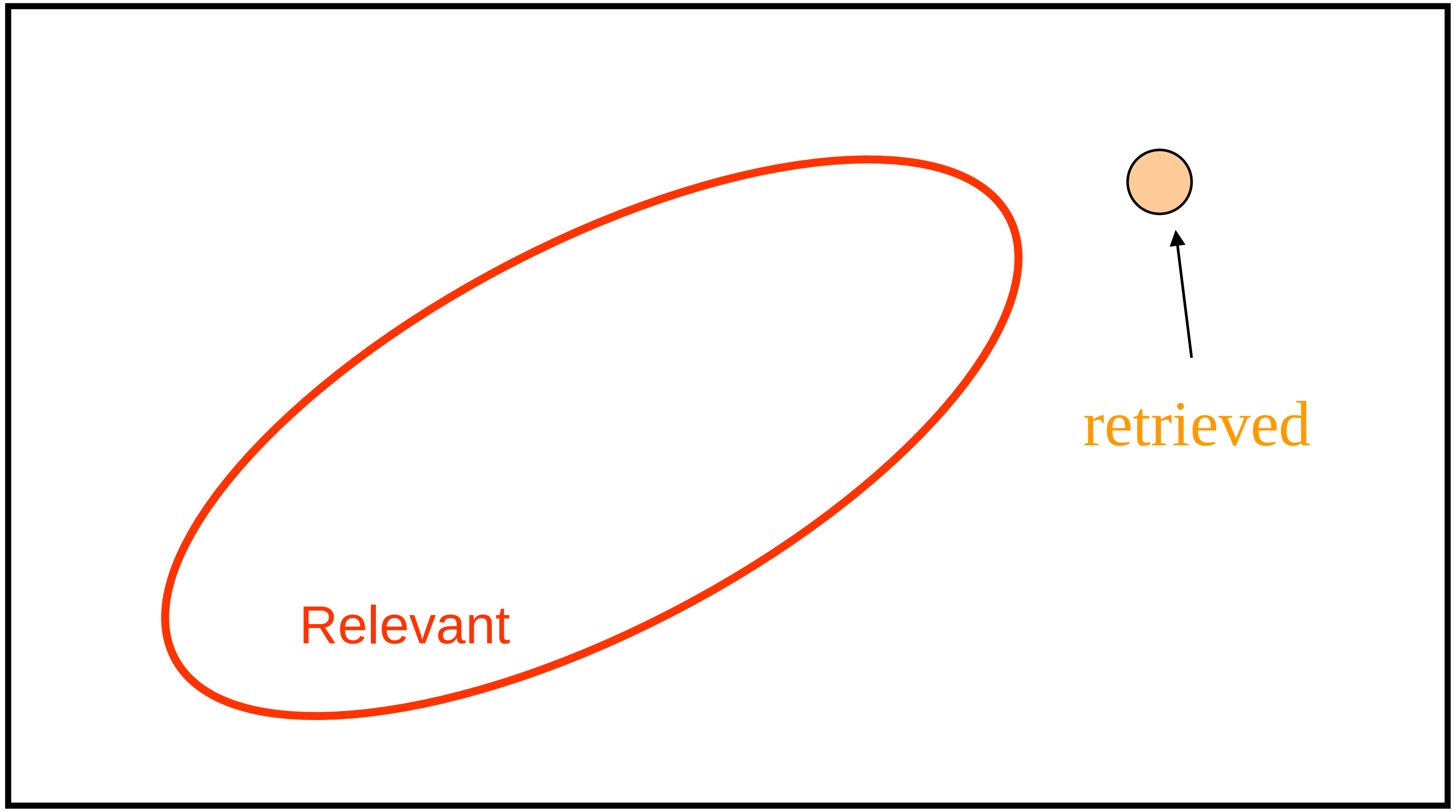
Retrieved vs. Relevant Documents

High recall, but low precision



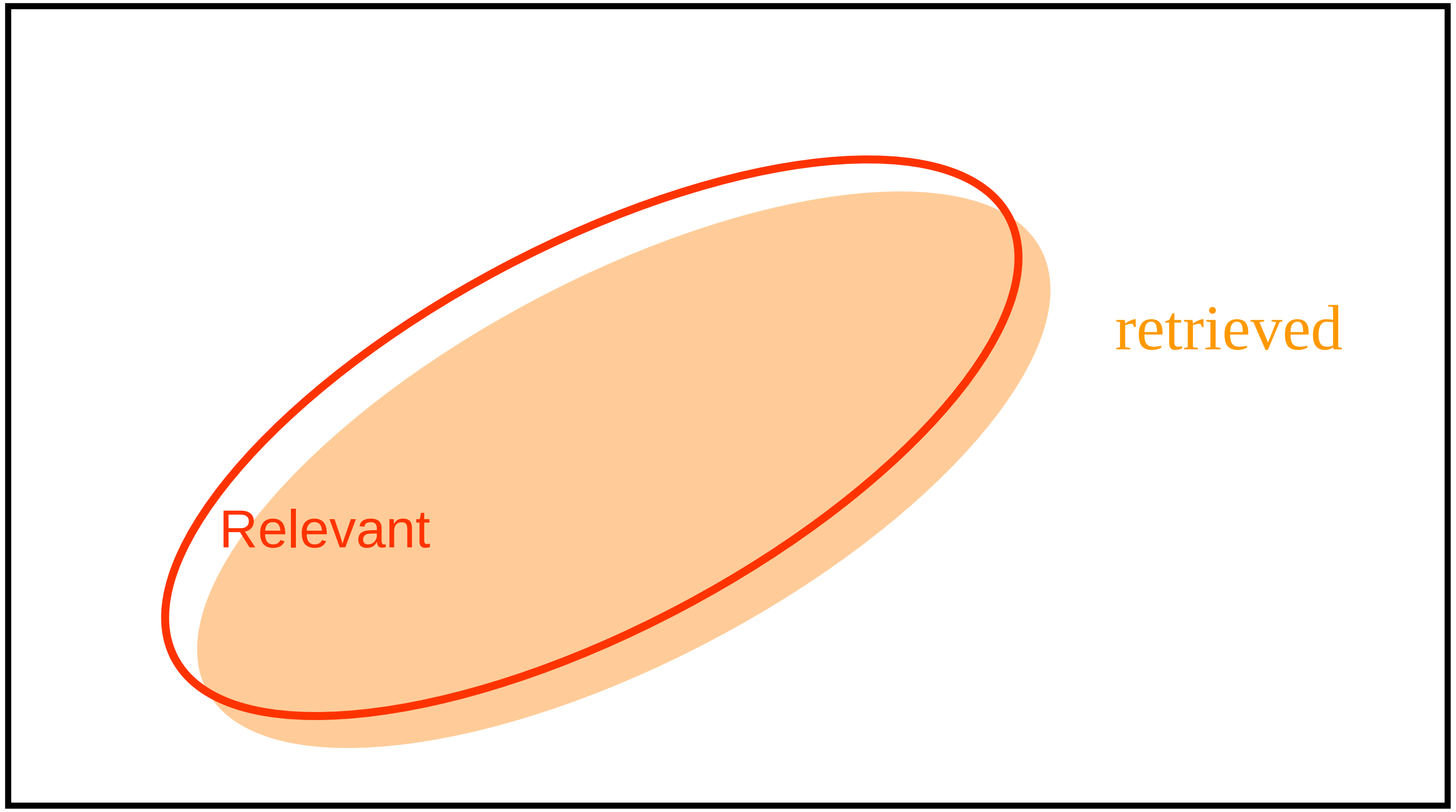
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 for both)



Retrieved vs. Relevant Documents

High precision, high recall (at last!)



Accuracy

- Given a query, an engine (**classifier**) classifies each doc as “Relevant” or “Nonrelevant”
  - *What is retrieved is classified by the engine as "relevant" and what is not retrieved is classified as "nonrelevant"*
- The **accuracy** of the engine: the fraction of these classifications that are correct
  - (tp + tn) / (tp + fp + fn + tn)
- **Accuracy** is a commonly used evaluation measure in **machine learning** classification work

Accuracy is not an appropriate measure for information retrieval problems.

- In almost all circumstances, the data is extremely skewed: normally over 99.9% of the documents are in the nonrelevant category.
- A system tuned to maximize accuracy can appear to perform well by simply deeming all documents nonrelevant to all queries.
- Labeling all documents as nonrelevant is completely unsatisfying to an information retrieval system user.
- Users are assumed to have a certain tolerance for seeing some false positives providing that they get some useful information
- The advantage of having the two numbers for precision and recall is that one is more important than the other in many circumstances.
- Typical web surfers would like every result on the first page to be relevant (high precision)
- In contrast, various professional searchers such as paralegals and intelligence analysts are very concerned with trying to get as high recall as possible, and will tolerate fairly low precision results in order to get it.

- In a good system, precision usually decreases as the number of documents retrieved is increased.
- In general we want to get some amount of recall while tolerating only a certain percentage of false positives.
- A single measure that trades off precision versus recall is the *F measure*, which is the weighted harmonic mean of precision and recall

AM, GM & HM

- ## Arithmetic Mean

  Arithmetic mean represents a number that is achieved by dividing the sum of the values of a set by the number of values in the set. If a_1, a_2, a_3,….,a_n, is a number of group of values or the Arithmetic Progression, then;

  $AM=(a_1+a_2+a_3+....,+a_n)/n$

- ## Geometric Mean

  The Geometric Mean for a given number of values containing n observations is the nth root of the product of the values.

  $GM = n\surd(a_1a_2a_3....a_n)$

  Or

  $GM = (a_1a_2a_3....a_n)^{1/n}$

- ## Harmonic Mean

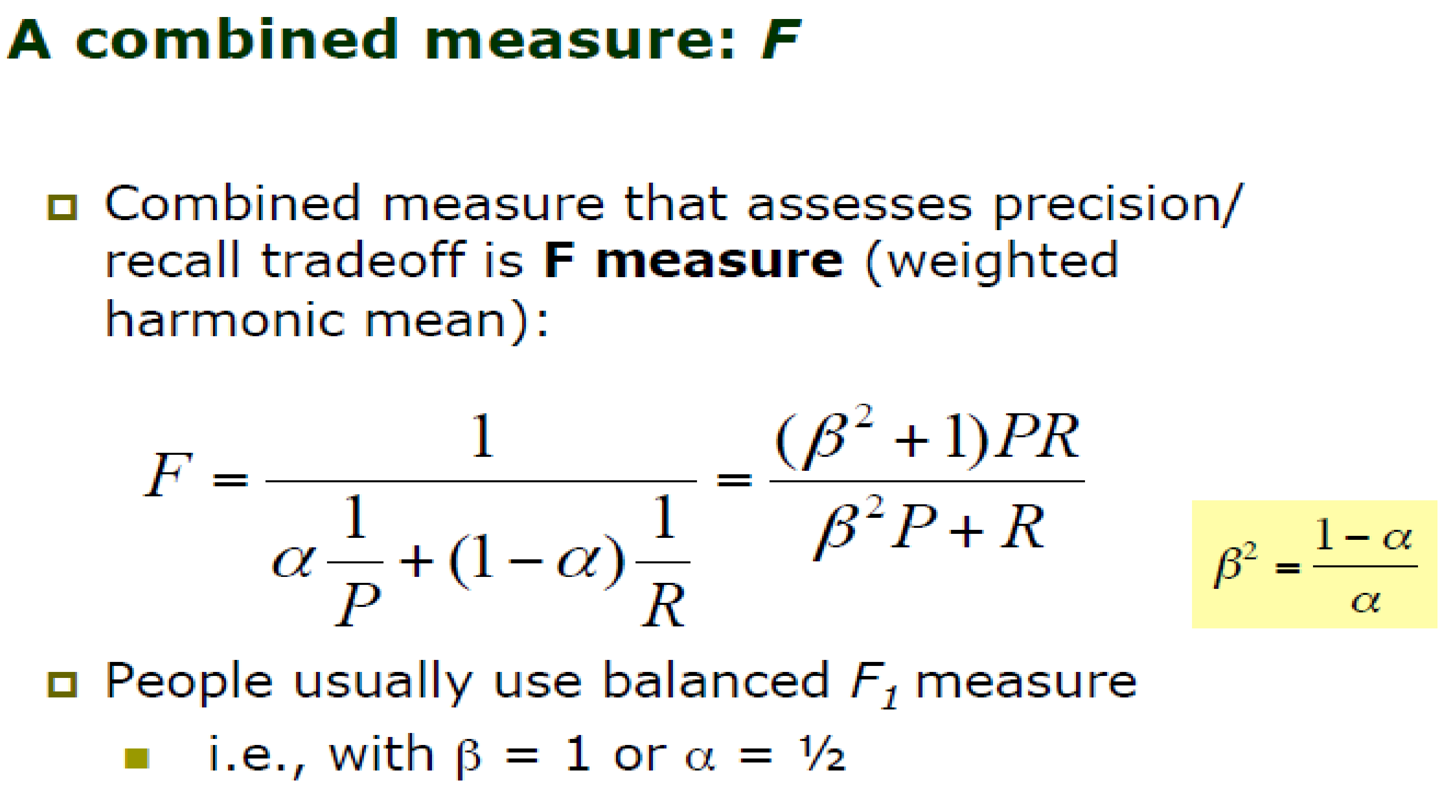
  HM is defined as the reciprocal of the arithmetic mean of the given data values. It is represented as:

  $HM = n/[(1/a_1) + (1/a_2) + (1/a_3) + ...+ (1/a_n)]$

F-measure

- A single measure that trades off precision versus recall is the *F measure*, which is the weighted harmonic mean of precision and recall.
- HM : less weightage to large values and high weightage to small values.
- we can always get 100% recall by just returning all documents, and therefore we can always get a 50% arithmetic mean by the same process.
- This strongly suggests that the arithmetic mean is an unsuitable measure to use.
- The harmonic mean is always less than or equal to the arithmetic mean and the geometric mean.
- When the values of two numbers differ greatly, the harmonic mean is closer to their minimum than to their arithmetic mean;

F- measure



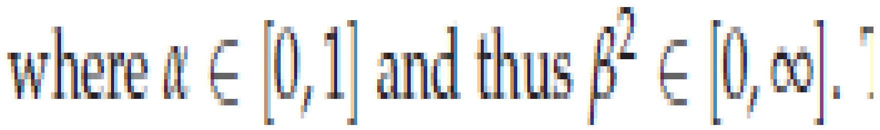
A combined measure: *F*

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F=\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

$$\beta^2=\frac{1-\alpha}{\alpha}$$

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ or $\alpha = ½$



where $\alpha\in[0,1]$ and thus $\beta^2\in[0,\infty]$.

F- measure

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$= \frac{1}{\frac{\alpha}{P} + \frac{\alpha\beta^2}{R}}$$

$$= \frac{1}{\frac{\alpha R + P\alpha\beta^2}{PR}}$$

$$= \frac{PR}{\alpha R + P\alpha\beta^2}$$

$$= \frac{PR}{\alpha(R + P\beta^2)}$$

$$\cancel{= \frac{PR}{(1+\beta^2)(}} = \frac{PR(1+\beta^2)}{(R+P\beta^2)}$$

$$\beta^2 = \frac{1-\alpha}{\alpha}$$

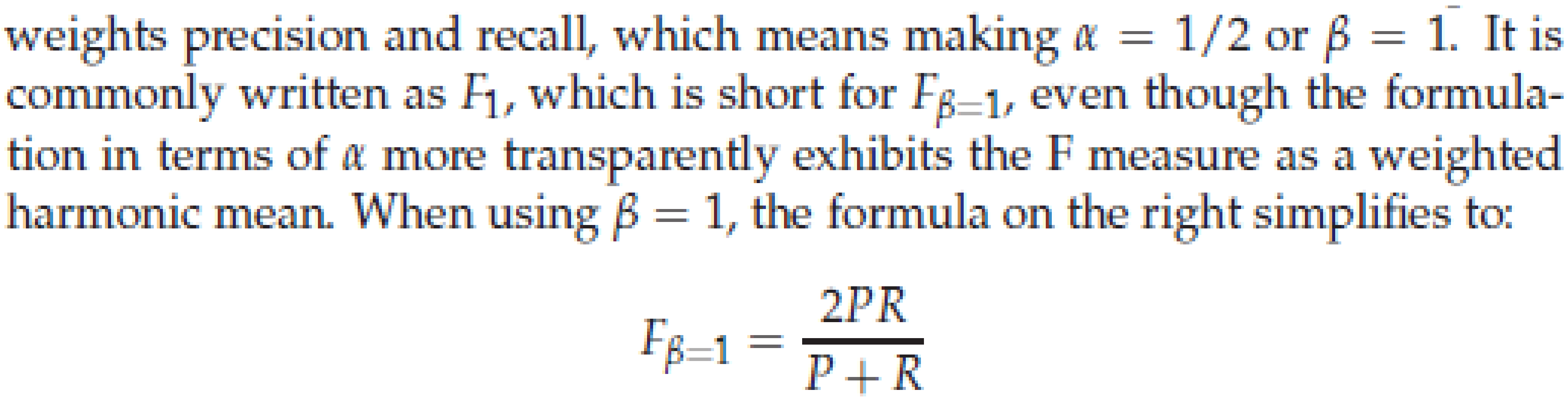
$$(1-\alpha) = \alpha\beta^2$$

$$\alpha\beta^2 + \alpha = 1$$

$$\alpha(1+\beta^2) = 1$$

$$\frac{1}{\alpha} = 1+\beta^2$$

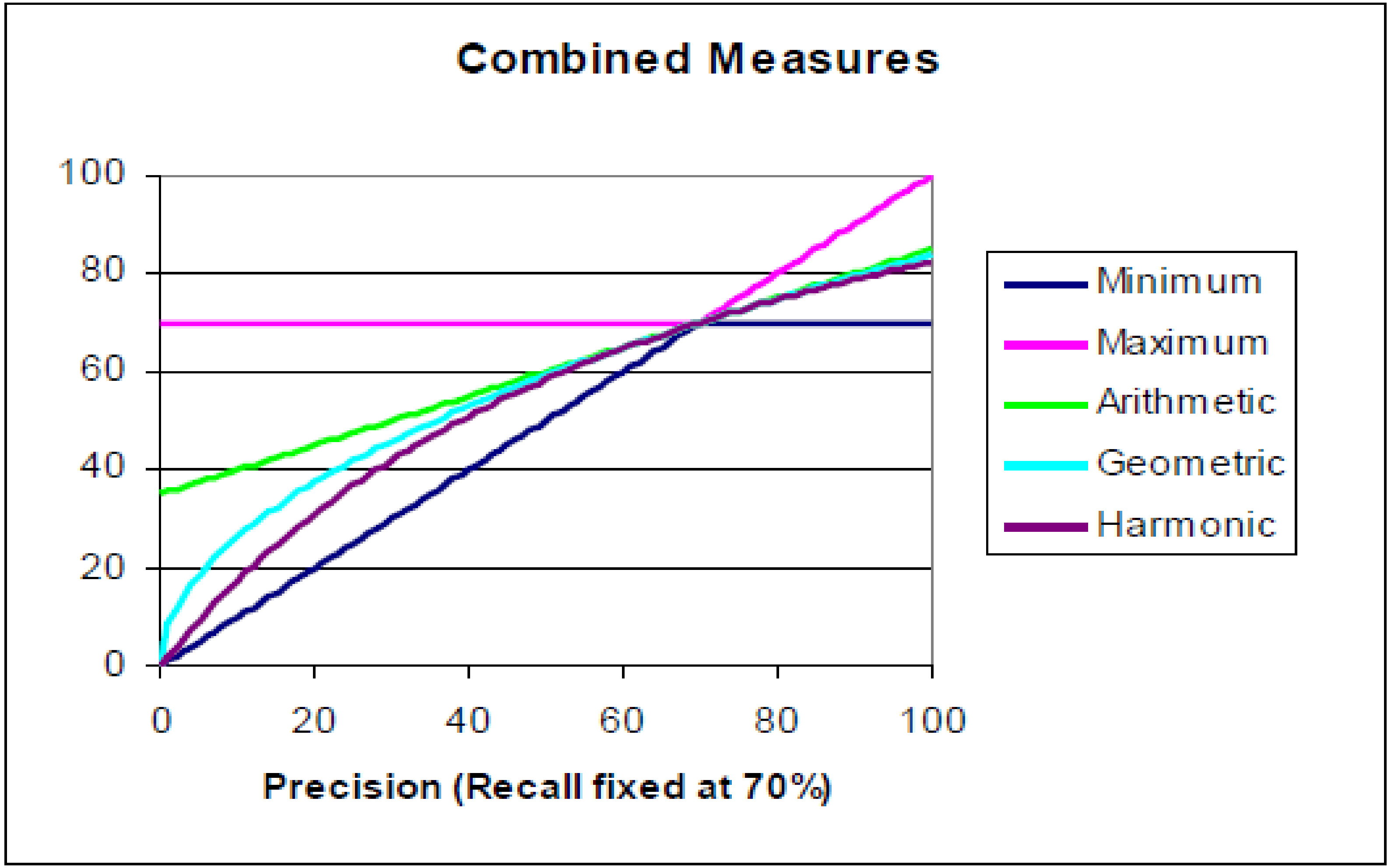
The default *balanced F measure* equally



weights precision and recall, which means making $\alpha = 1/2$ or $\beta = 1$. It is commonly written as F_1, which is short for $F_{\beta=1}$, even though the formulation in terms of α more transparently exhibits the F measure as a weighted harmonic mean. When using $\beta = 1$, the formula on the right simplifies to:

$$F_{\beta=1} = \frac{2PR}{P + R}$$

F_1 and other averages



Graph comparing the harmonic mean to other means.
Recall value is 70%. The harmonic mean is always less than either the arithmetic or geometric mean, and often quite close to the minimum of the two numbers.
When the precision is also 70%, all the measures coincide.

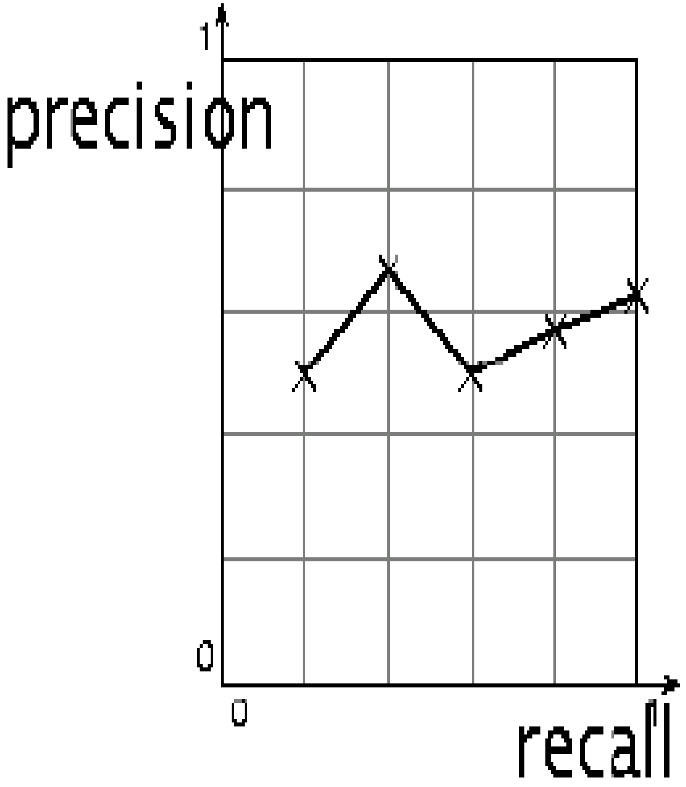
Geometric mean of *a* and *b* is $(a*b)^{1/2}$

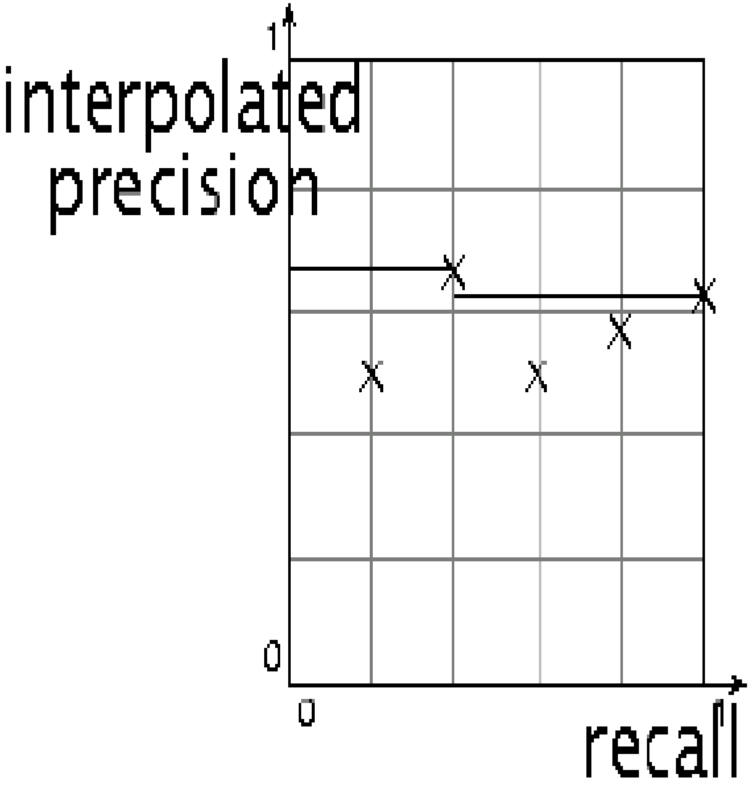
Evaluation of ranked retrieval results

- Precision, recall, and the F measure are set-based measures.
- We need to extend measures (Precision, Recall & F-measure) or define new measures if we are to evaluate the ranked retrieval results.
- *Precision-recall curve*
- *For each top-k documents*, precision and recall values can be plotted to give a *precision-recall curve*
- Precision-recall curves have a distinctive saw-tooth shape:
    - If the (*k* + 1)th document retrieved is nonrelevant then recall is the same as for the top *k* documents, but precision is dropped.
    - If it is relevant, then both precision and recall increase, and the curve jags up and to the right.
- The *interpolated precision pinterp* at a certain recall level *r* is defined as the highest precision found for any recall level *r*′ ≥ *r*:

$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

Interpolated precision

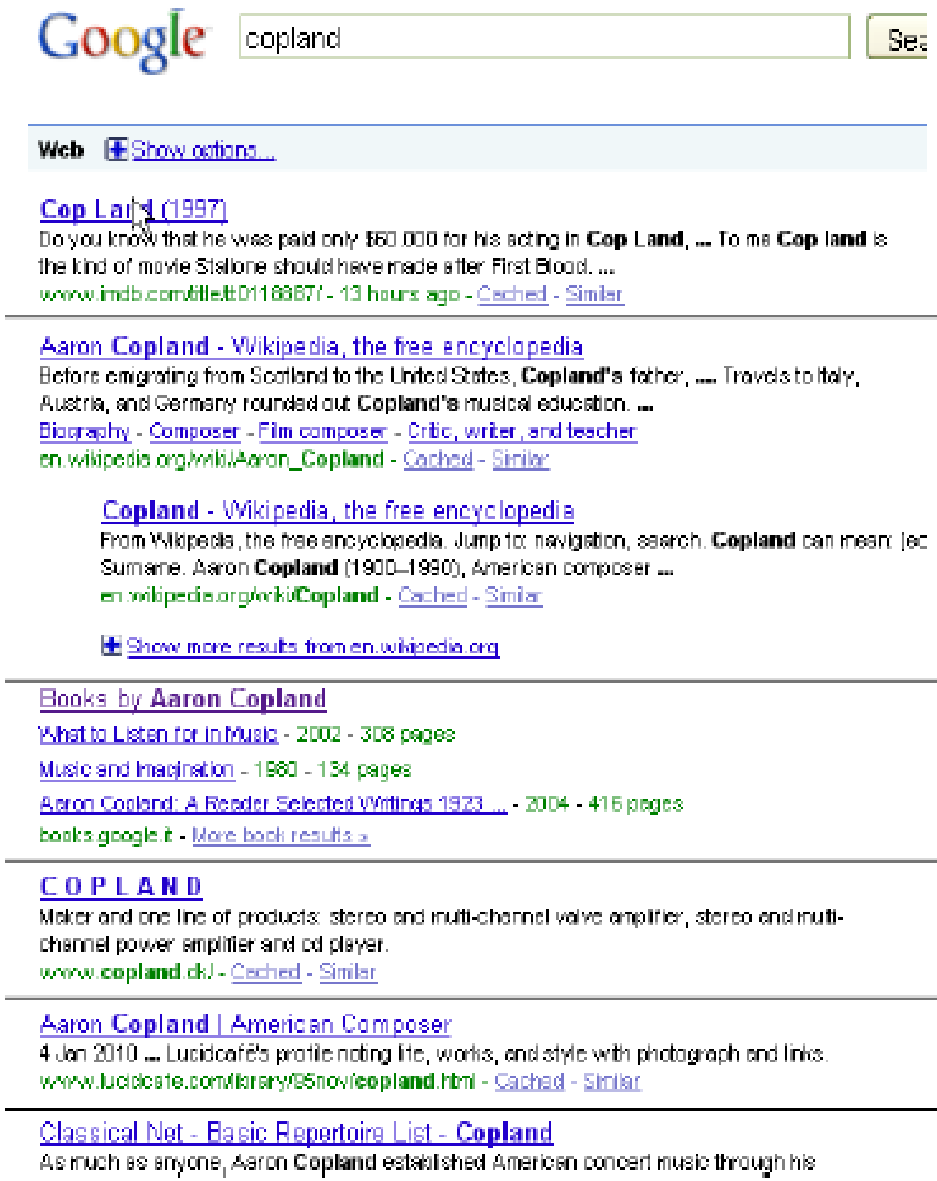




$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

Definition of interpolated precision

Precision-Recall



Cop Land (1997)
Do you know that he was paid only $60,000 for his acting in **Cop Land**, ... To me **Cop land** is the kind of movie Stallone should have made after First Blood. ...
www.imdb.com/title/tt0118887/ - 13 hours ago - Cached - Similar

P=0/1, R=0/1000

Aaron Copland - Wikipedia, the free encyclopedia
Before emigrating from Scotland to the United States, **Copland's** father, ... Travels to Italy, Austria, and Germany rounded out **Copland's** musical education. ...
Biography - Composer - Film composer - Critic, writer, and teacher
en.wikipedia.org/wiki/Aaron_**Copland** - Cached - Similar

P=1/2, R=1/1000

Copland - Wikipedia, the free encyclopedia
From Wikipedia, the free encyclopedia. Jump to: navigation, search. **Copland** can mean: [ec
Surname. Aaron **Copland** (1900–1990), American composer ...
en.wikipedia.org/wiki/**Copland** - Cached - Similar

Show more results from en.wikipedia.org

Books by **Aaron Copland**
What to Listen for in Music - 2002 - 308 pages
Music and Imagination - 1980 - 134 pages
Aaron Copland: A Reader Selected Writings 1923 ... - 2004 - 416 pages
books.google.it - More book results »

P=2/3, R=2/1000

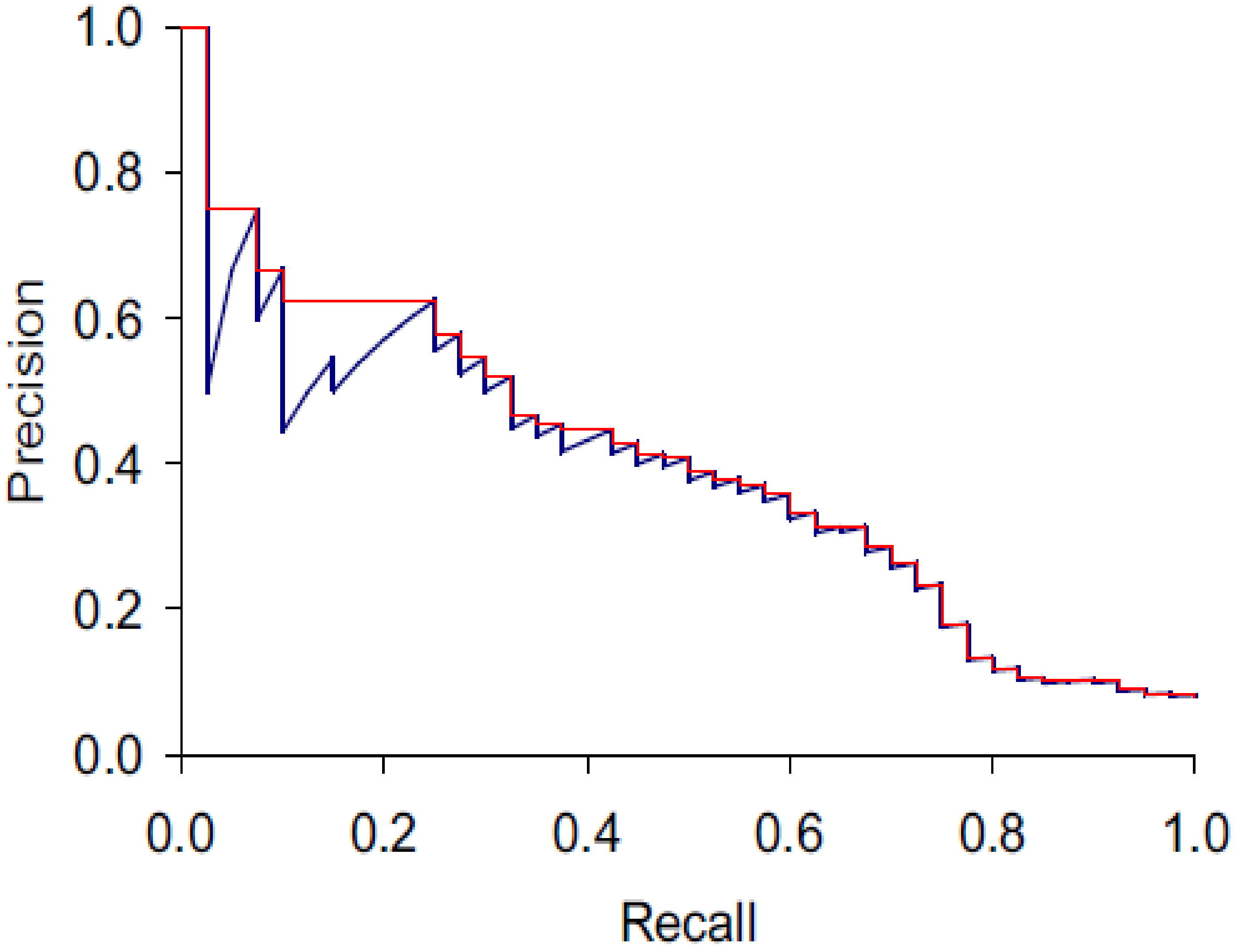
COPLAND
Maker and one line of products: stereo and multi-channel valve amplifier, stereo and multi-channel power amplifier and cd player.
www.**copland**.dk/ - Cached - Similar

P=2/4, R=2/1000

Aaron Copland | American Composer
4 Jan 2010 ... Lucidcafé's profile noting life, works, and style with photograph and links.
www.lucidcafe.com/library/95nov/**copland**.html - Cached - Similar

P=3/5, R=3/1000

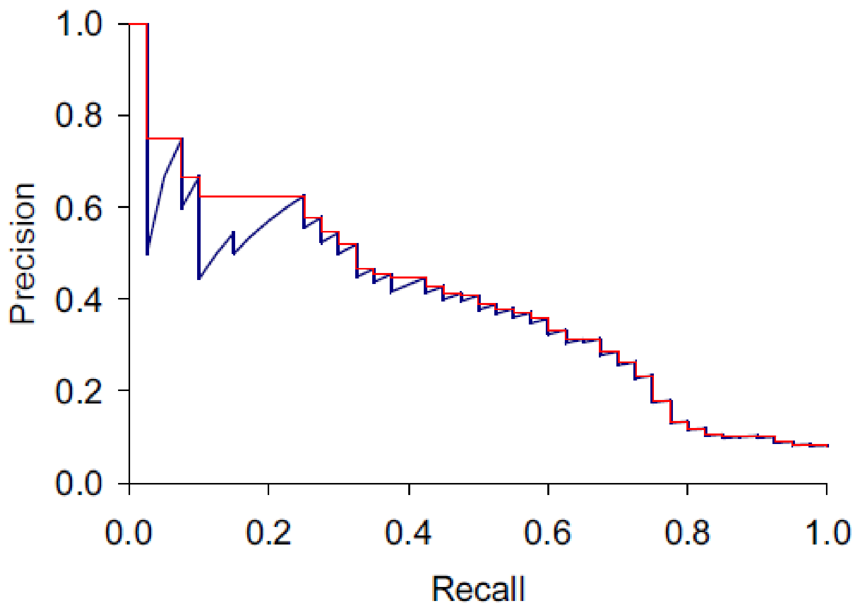
Classical Net - Basic Repertoire List - **Copland**
As much as anyone, Aaron **Copland** established American concert music through his



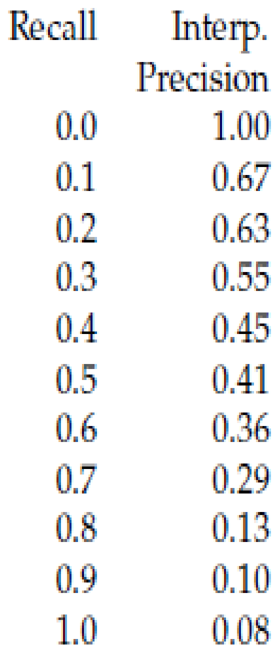
► **Figure 8.2** Precision/ recall graph.

11-point interpolated average precision

- Examining the entire precision-recall curve is very informative, but there is often a desire to boil this information down to a few numbers, or perhaps even a single number.
- The traditional way of doing this is the *11-point interpolated average precision*.
- The standard measure in the early TREC competitions
- Take the *interpolated* precision at 11 levels of recall varying from 0 to 1 by tenths
- Then **average them**
- Evaluates performance at all recall levels.



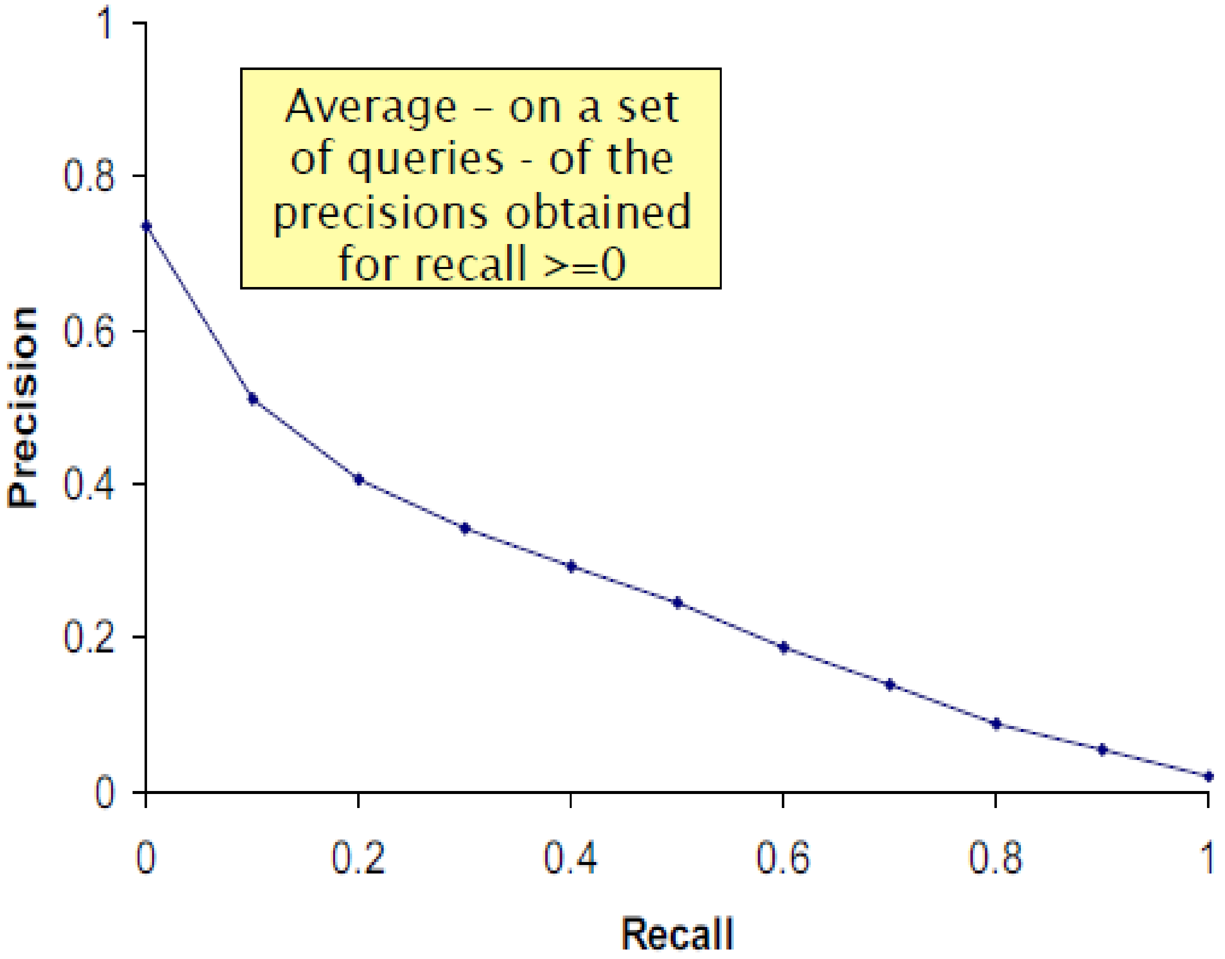
▸ **Figure 8.2** Precision/recall graph.



Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

▸ **Table 8.1** Calculation of 11-point Interpolated Average Precision. This is for the precision-recall curve shown in Figure 8.2.

- A composite precision recall curve showing 11 points can then be graphed.



Average – on a set of queries - of the precisions obtained for recall >=0

▶ **Figure 8.3** Averaged 11-point precision/recall graph across 50 queries for a representative TREC system. The Mean Average Precision for this system is 0.2553.

Mean Average Precision (MAP)

- Average of the precision values obtained for increasing values of *K*, for the top *K* documents, each time a new relevant doc is retrieved
- Avoids interpolation, use of fixed recall levels
- MAP for a query collection is arithmetic average
  - Macro-averaging: each query counts equally
- **Definition:** if the set of relevant documents for an information need q_j is $\{d_1, \ldots, d_{m\_j}\}$ and R_{jk} is the set of documents retrieved until you get d_k, then:

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

Example

Q1		Q2	
🟩	1/1	🟩	1/1
🔴		🟩	2/2
🟩	2/3	🔴	
🔴		🔴	
🔴		🔴	
🔴		🟩	3/6
🟩	3/7	🟩	4/7
🔴		🔴	
⋮		⋮	
🔴		🔴	

(1+2/3+3/7)/3 = 0.69

(1+1+3/6+4/7)/4 = 0.76

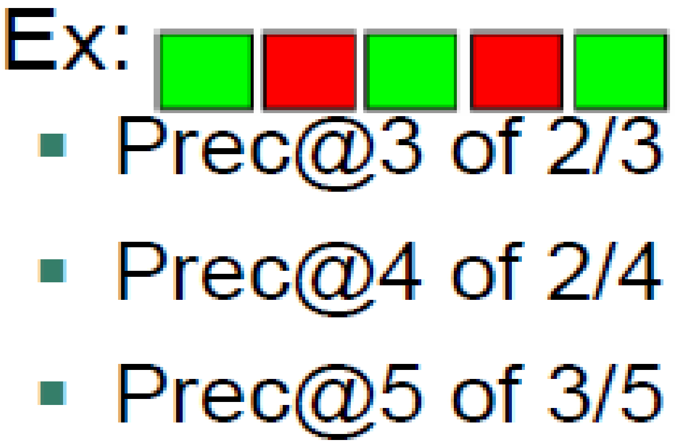
Average precision =
(0.69 +0.76)/2 = 0.72

🔴 nonrelevant

🟩 relevant

Precision at *k*

- Many measures factor in precision at all recall levels.
- For many prominent applications, particularly web search, this may not be useful to users.
- What matters is rather how many good results there are on the first page or the first three pages.
- This leads to measuring precision at fixed low levels of retrieved results, such as 10 or 30 documents. This is referred to as “Precision at *k*”, for example “Precision at 10”.
- It has the advantage of not requiring any estimate of the size of the set of relevant documents



Ex: 🟩🟥🟩🟥🟩

- Prec@3 of 2/3
- Prec@4 of 2/4
- Prec@5 of 3/5

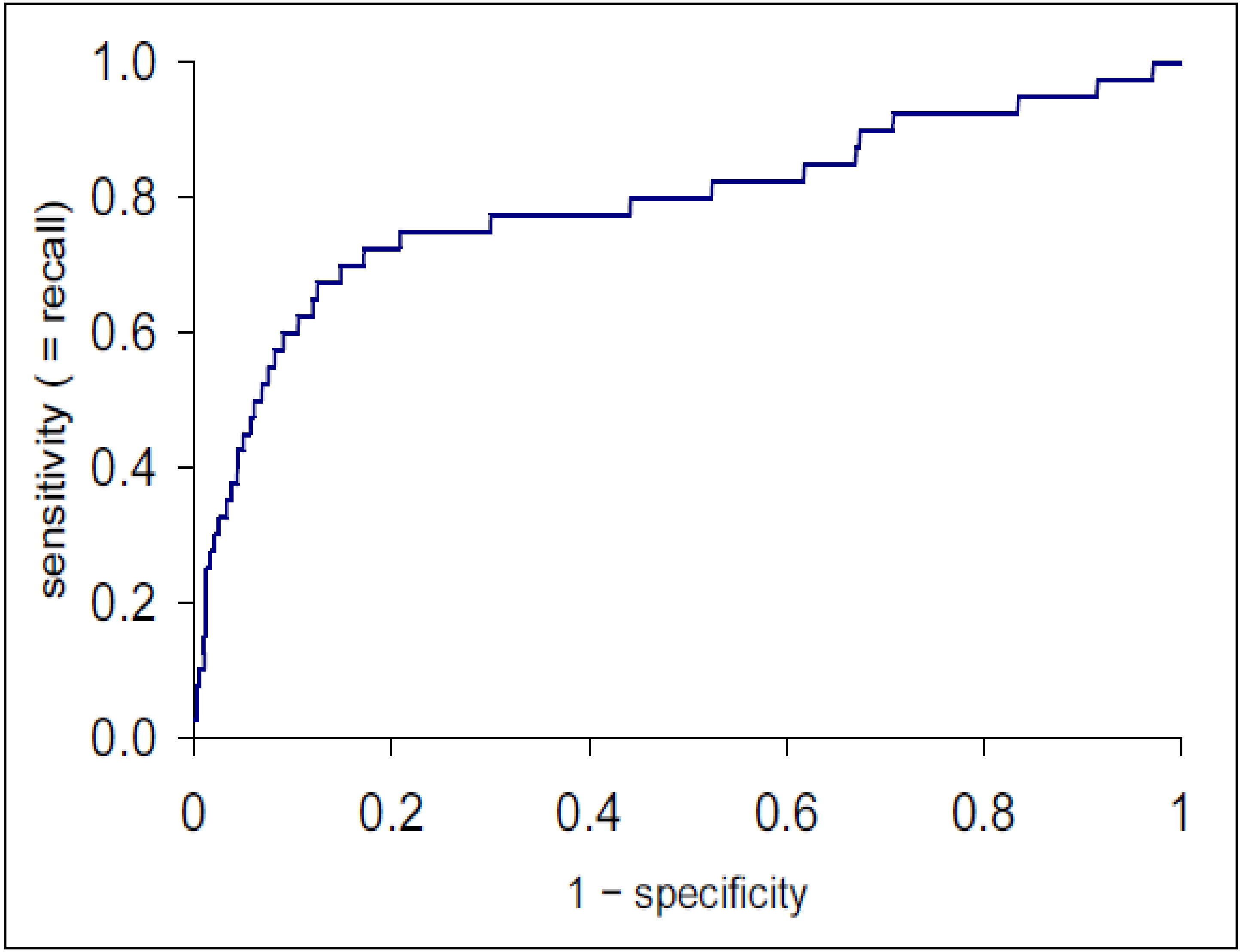
- The disadvantage is it does not average well, since the total number of relevant documents for a query has a strong influence on precision at *k*.

R-Precision

- It requires having a set of known relevant documents *Rel*, from which we calculate the precision of the top *Rel* documents returned.
- R-precision adjusts for the size of the set of relevant documents
- If there are |*Rel*| relevant documents for a query, we examine the top |*Rel*| results of a system, and find that *r* are relevant, then by definition, not only is the precision *r*/|*Rel*|, but the recall of this result set is also *r*/|*Rel*|.
- R-precision turns out to be identical to the *break-even point*, another measure which is sometimes used, defined in terms of this equality relationship holding.
- Like Precision at *k*, R-precision describes only one point on the precision-recall curve,
- Example
- Suppose in your collection there are 100 documents in total, 30 of which are relevant, the rest irrelevant.
- So you retrieve the first 30 documents (because 30 are relevant in total in your collection) and, say, 10 of them are relevant.
- Your R-Precision is then 10/30 =1/3

ROC Curve

- Another concept sometimes used in evaluation is an *ROC curve*. (“ROC” stands for “Receiver Operating Characteristics”.
- An ROC curve plots the true positive rate or sensitivity against the false positive rate or (1 − specificity).
- Here, *sensitivity* is just another term for recall.
- *Specificity*, given by $tn/(fp+tn)$, 1- specificity = $fp/(fp+tn)$.
- An ROC curve always goes from the bottom left to the top right of the graph.
- For a good system, the graph climbs steeply on the left side.
- For unranked result sets, *specificity*, given by $tn/(fp+tn)$,was not seen as a very useful notion. Because the set of true negatives is always so large, its value would be almost 1 for all information needs

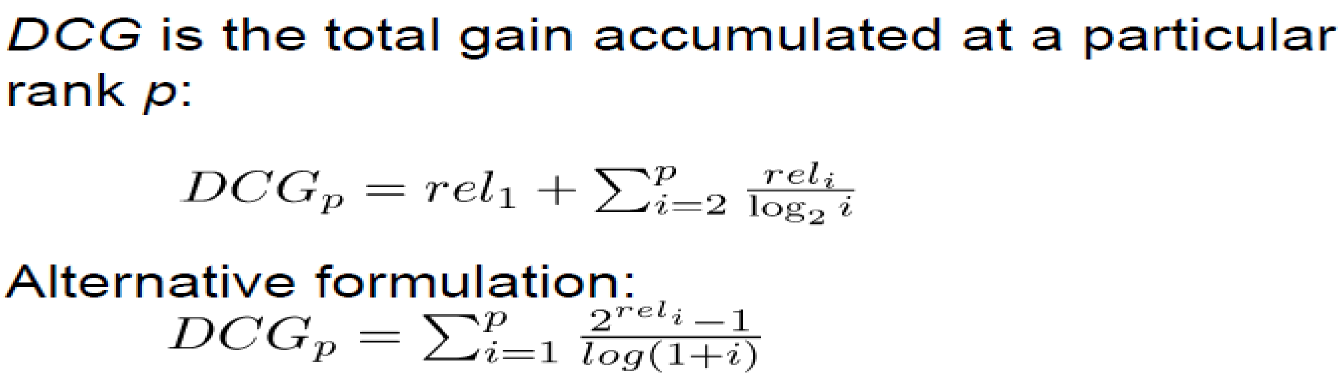


▶ **Figure 8.4** The ROC curve corresponding to the precision-recall curve in Figure 8.2.

Normalized discounted cumulative gain

- NDCG is designed for situations of non-binary notions of relevance
- Uses *graded relevance* as a measure of usefulness, or *gain*
- What if relevance judgments are in a scale of [0,r]? r>2
- 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
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r1, r2, …rn (in ranked order)
  - CG = r1+r2+…rn
  - Gain is accumulated starting at the top of the ranking and may be

DCG :



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:

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

- Discounted Cumulative Gain (DCG) at rank n
- DCG = r1 + r2/$\log_2 2$ + r3/$\log_2 3$ + … rn/$\log_2 n$
- We may use any base for the logarithm, e.g., base=b

- The ideal ranking first returns the documents with the highest relevance level, then the next highest relevance level, etc

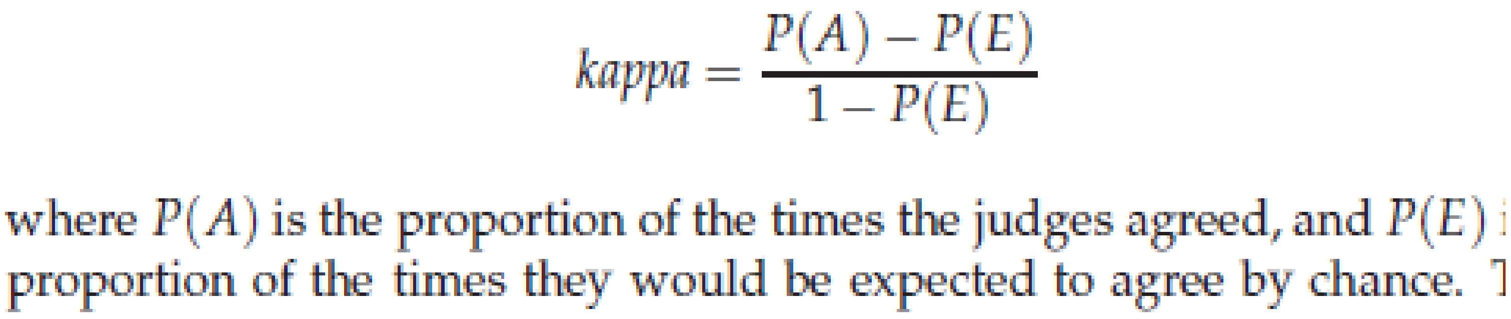
$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

Assessing relevance

- To properly evaluate a system, our test information needs must be relevant to the documents in the test document collection, and appropriate for predicted usage of the system
- Random combinations of query terms as an information need is generally not a good idea because typically they will not resemble the actual distribution of information needs
- Given information needs and documents, you need to collect relevance assessments. This is a time-consuming and expensive process involving human beings.
- For large modern collections, it is usual for relevance to be assessed only for a subset of the documents for each query. The most standard approach is *pooling*, where relevance is assessed over a subset of the collection that is formed from the top *k* documents returned by a number of different IR systems (usually the ones to be evaluated)
- It is interesting to consider and measure how much agreement between judges there is on relevance judgments.

Kappa Statistic

- In the social sciences, a common measure for agreement between judges is the KAPPA STATISTIC

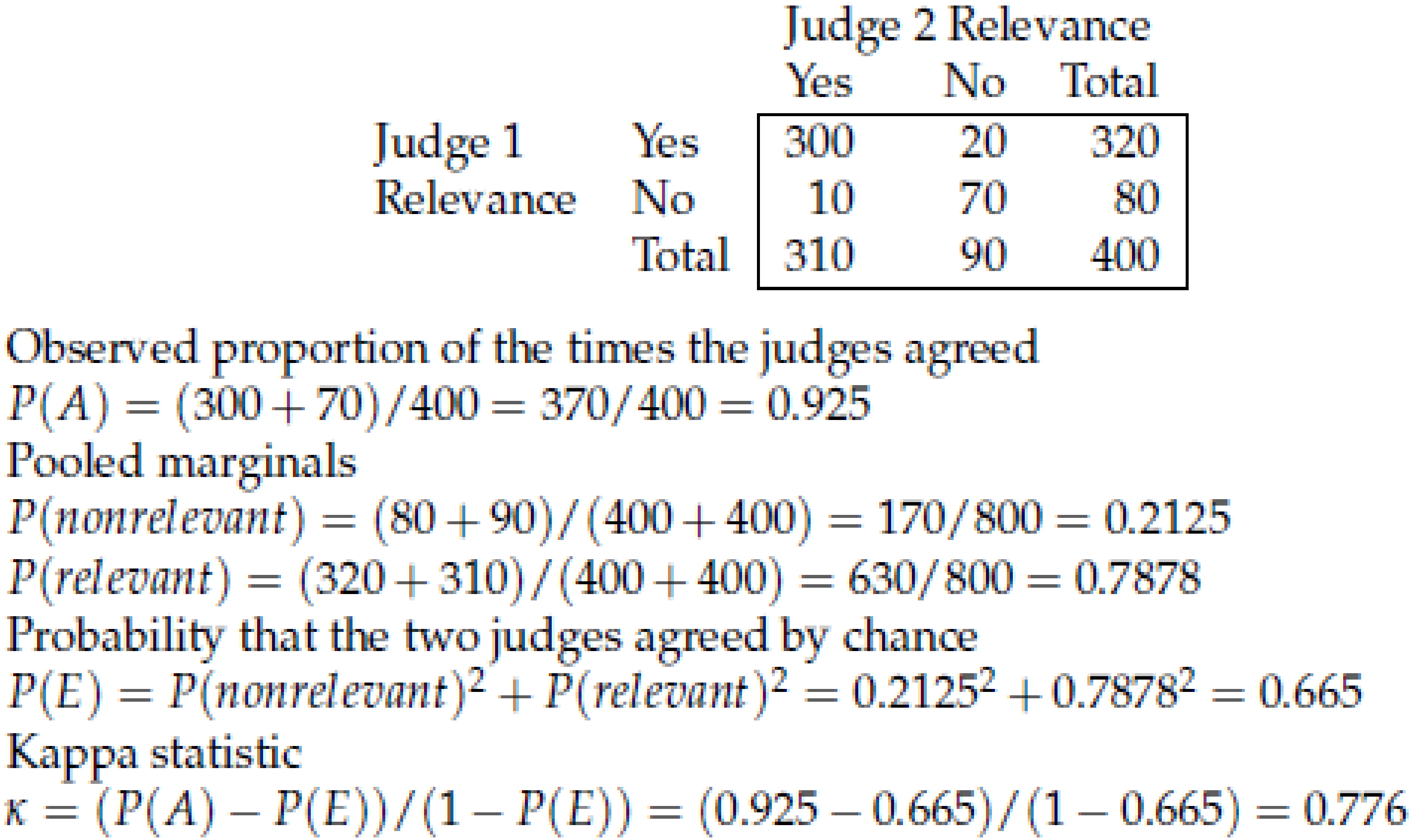


$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of the times the judges agreed, and $P(E)$ i proportion of the times they would be expected to agree by chance. T

- Kappa > 0.8 = good agreement

Kappa Statistic Example



		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Observed proportion of the times the judges agreed

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals

$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

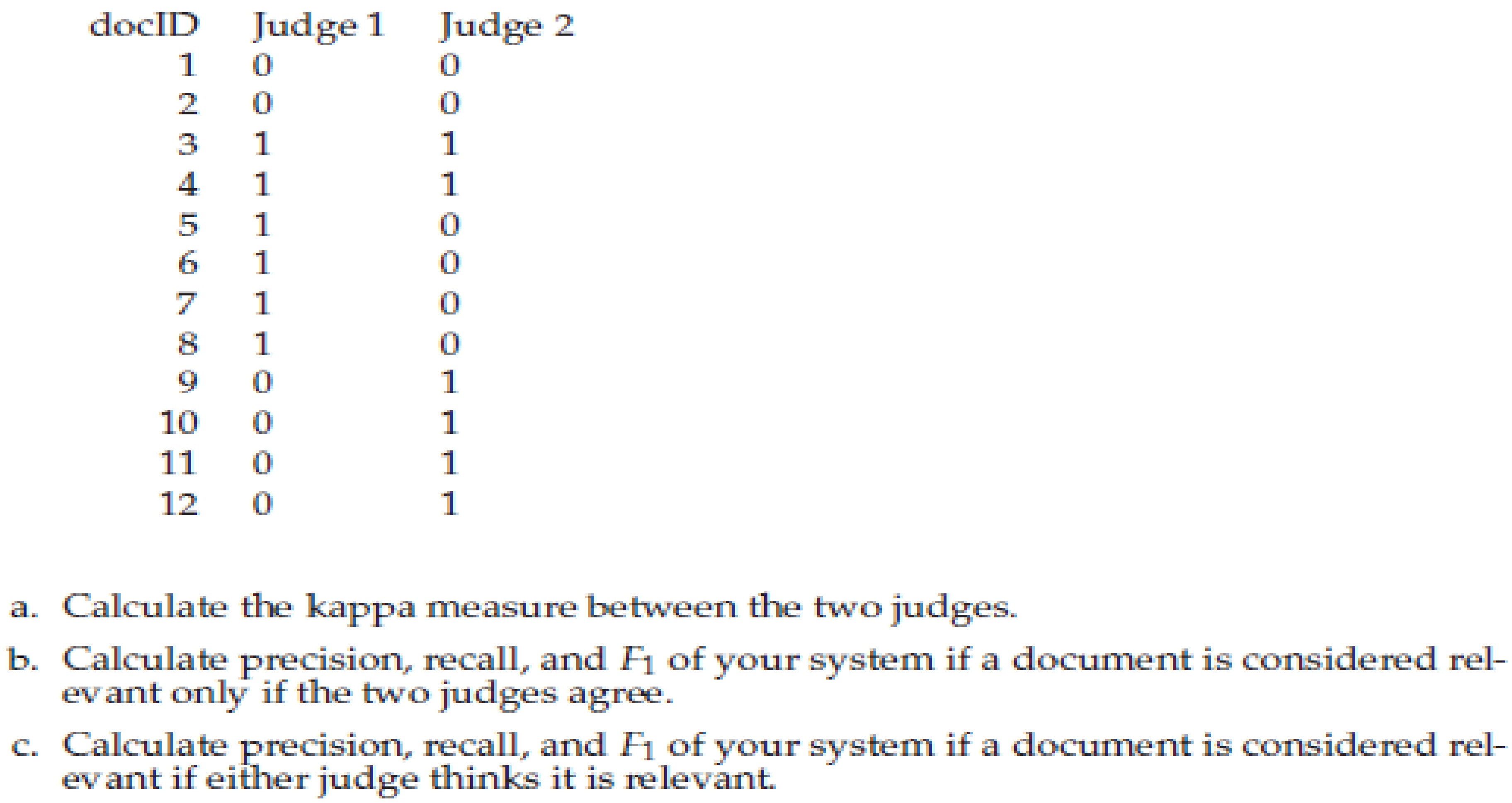
Probability that the two judges agreed by chance

$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic

$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

Example : Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you’vewritten an IR systemthat for this query returns the set of documents {4, 5, 6, 7, 8}.



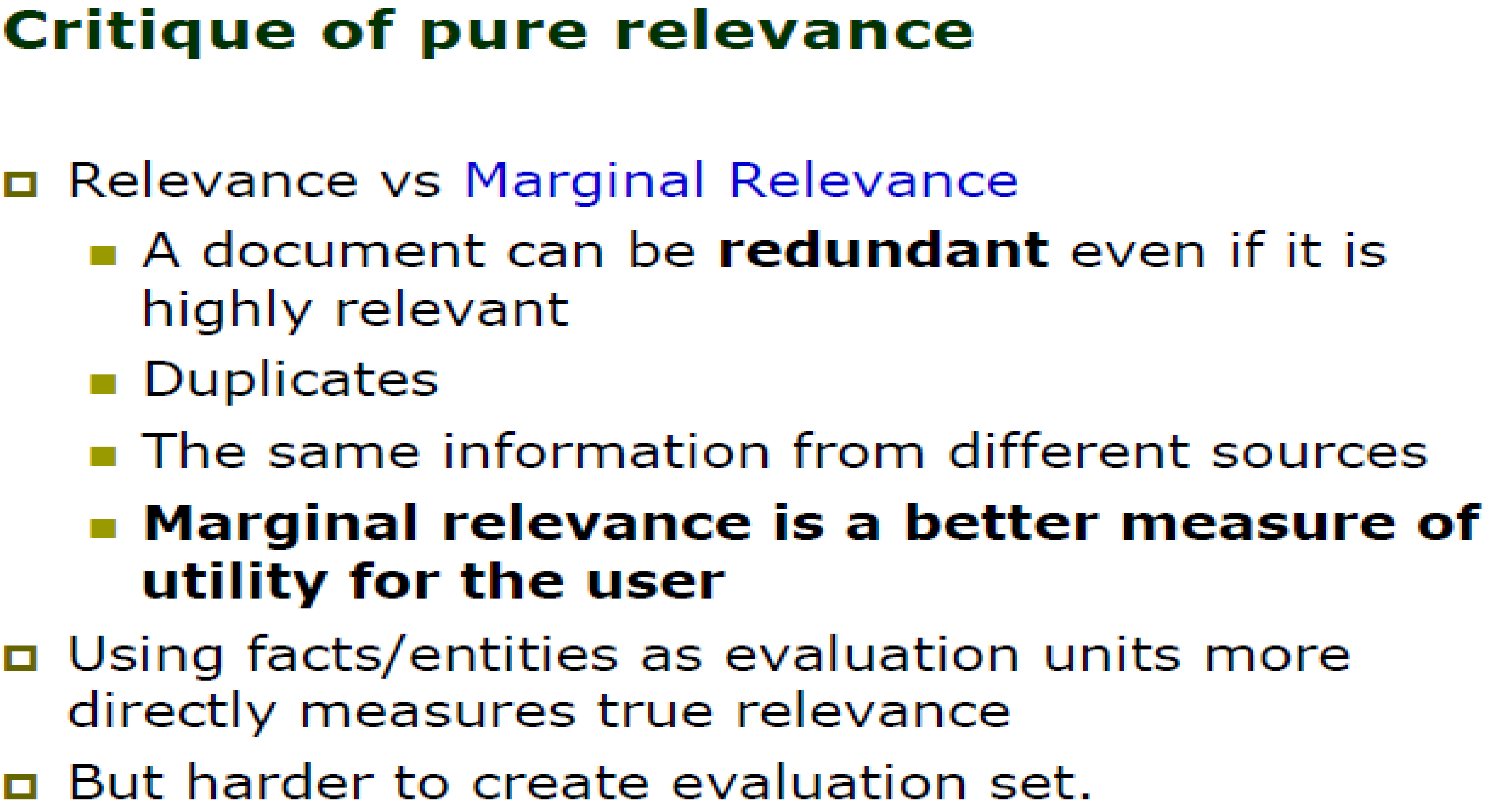
docID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

a. Calculate the kappa measure between the two judges.

b. Calculate precision, recall, and F_1 of your system if a document is considered relevant only if the two judges agree.

c. Calculate precision, recall, and F_1 of your system if a document is considered relevant if either judge thinks it is relevant.

Critiques and justifications of the concept of relevance



Critique of pure relevance

- Relevance vs Marginal Relevance
  - A document can be **redundant** even if it is highly relevant
  - Duplicates
  - The same information from different sources
  - **Marginal relevance is a better measure of utility for the user**
- Using facts/entities as evaluation units more directly measures true relevance
- But harder to create evaluation set.

Maximal Marginal Relevance (MMR)

- Challenge of presenting users with a diverse yet highly relevant set of items
- Maximal Marginal Relevance (MMR) is a powerful technique that addresses this challenge by striking a balance between an item's relevance and its diversity compared to previously selected items.
- MMR operates on the principle of maximizing the relevance of selected documents while minimizing redundancy.
- Benefits of MMR
  - **Enhanced User Experience**: By providing a diverse set of relevant documents, MMR improves the likelihood that users will find the information they need.
  - **Reduction of Redundancy**: MMR effectively reduces the chances of presenting similar documents, which can lead to a more engaging and informative experience.
- Practical Applications
- MMR is widely used in various applications, including:
  - **Search Engines**: To present a diverse set of results for ambiguous queries.
  - **Recommendation Systems**: To suggest items that are not only relevant but also varied, enhancing user satisfaction.
  - **Document Summarization**: To select sentences or paragraphs that cover different aspects of a topic without redundancy.

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web (why?)
- Search engines often use top k precision, e.g., k=10
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right: NDCG (**Normalized Cumulative Discounted Gain**)
- Search engines also use non-relevance-based measures:
    - **Clickthrough on first result:** Not very reliable if you look at a single clickthrough … but pretty reliable in the aggregate
    - Studies of **user behavior in the lab**
    - **A/B testing**.

A/B testing

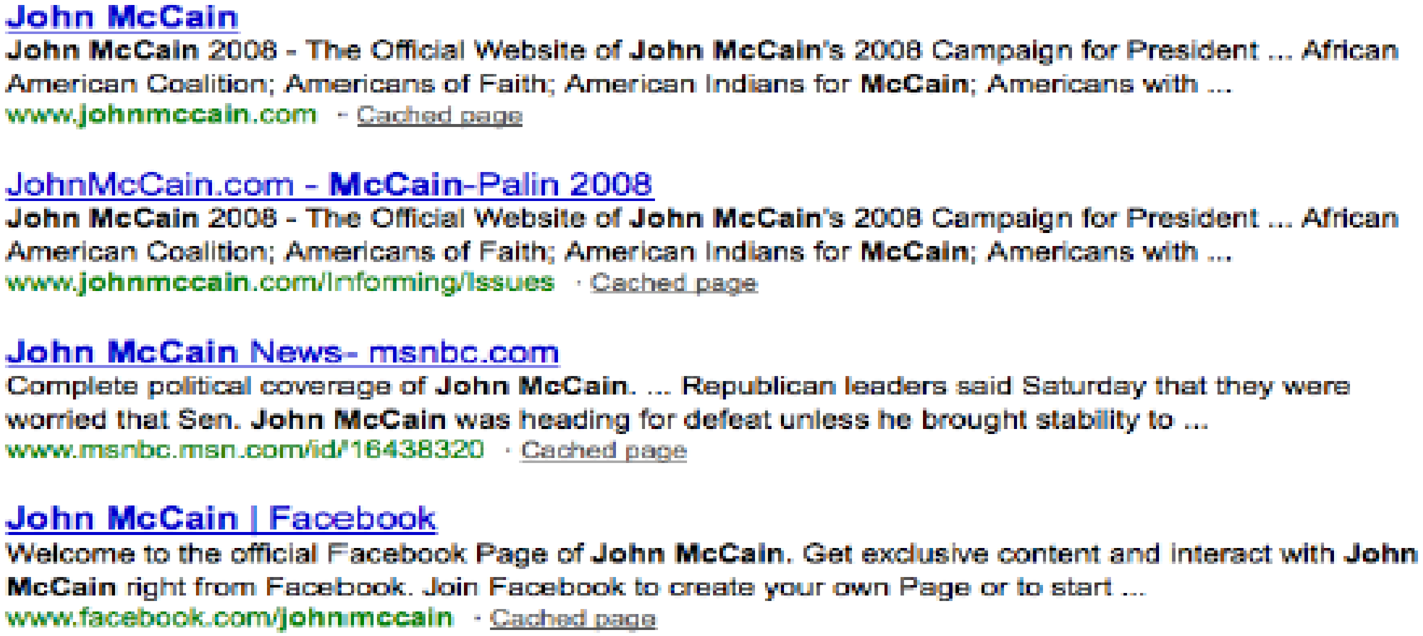
- **Purpose:** Test a single innovation
- **Prerequisite:** You have a large search engine up and running.
- Have **most users use old system**
- **Divert a small proportion of traffic** (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness
- Probably the evaluation methodology that large search engines trust most (true also for RecSys).

A/B testing

- In practice, this is how A/B testing works:
- Creating two versions of a page - the original (control or A) and a modified version (variation or B)
- Randomly splitting your traffic between these versions
- Measuring user engagement through a dashboard
- Analyzing results to determine if the changes had positive, negative, or neutral effects

Result Summaries

- Having ranked the documents matching a query, we wish to present a results list
- Most commonly, a list of the document titles plus a short summary, aka “10 blue links”



John McCain
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com - Cached page

JohnMcCain.com - **McCain**-Palin 2008
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com/Informing/Issues · Cached page

John McCain News- msnbc.com
Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...
www.msnbc.msn.com/id/16438320 · Cached page

John McCain | Facebook
Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...
www.facebook.com/**johnmccain** · Cached page

Summaries

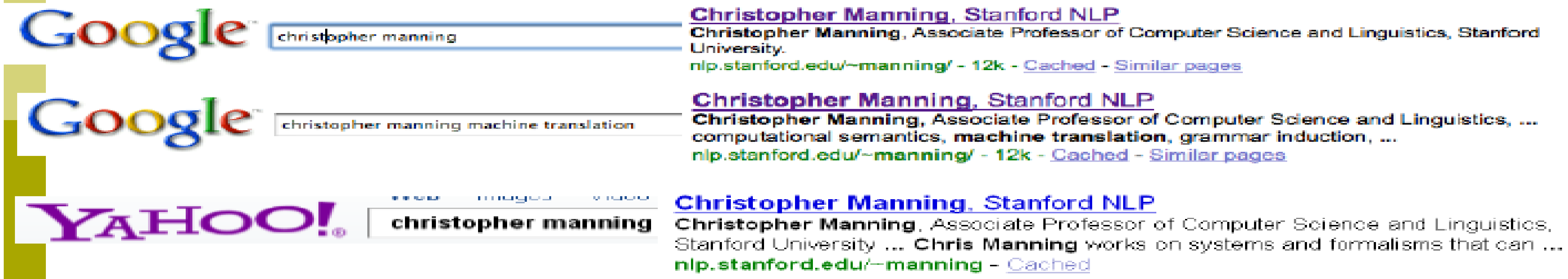
- The title is often automatically extracted from document metadata. What about the summaries?
  - This description is crucial
  - User can identify good/relevant hits based on description
- Two basic kinds:
  - Static
  - Dynamic
- A **static summary** of a document is always the same, regardless of the query that hit the doc
- A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand.

Static summaries

- In typical systems, the static summary **is a subset of the document**
- **Simplest heuristic:** the first 50 (or so – this can be varied) 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.

Dynamic summaries

- Present one or more "windows" within the document that contain several of the query terms
  - "KWIC" snippets: Keyword in Context presentation



christopher manning

Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University.
nlp.stanford.edu/~manning/ - 12k - Cached - Similar pages

christopher manning machine translation

Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, ...
computational semantics, **machine translation**, grammar induction, ...
nlp.stanford.edu/~**manning**/ - 12k - Cached - Similar pages

christopher manning

Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University ... **Chris Manning** works on systems and formalisms that can ...
nlp.stanford.edu/~**manning** - Cached

- Prefer snippets where query terms occurred as a phrase or jointly in a small window (e.g., paragraph).
- The summary that is computed this way gives the entire content of the window – all terms, not just the query terms.

Query: “new guinea economic development”

Snippets (in bold) that were extracted from a document:

> … **In recent years, Papua New Guinea has faced severe economic difficulties and** economic growth has slowed, partly as a result of weak governance and civil war, and partly as a result of external factors such as the Bougainville civil war which led to the closure in 1989 of the Panguna mine (at that time the most important foreign exchange earner and contributor to Government finances), the Asian financial crisis, a decline in the prices of gold and copper, and a fall in the production of oil. **PNG’s economic development record over the past few years is evidence that** governance issues underly many of the country’s problems. Good governance, which may be defined as the transparent and accountable management of human, natural, economic and financial resources for the purposes of equitable and sustainable development, flows from proper public sector management, efficient fiscal and accounting mechanisms, and a willingness to make service delivery a priority in practice. …

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
- Note that the cached copy can be outdated
- Don't cache very long documents – just cache a short prefix

- Space on the search result page is limited.
- The snippets must be short but also long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally:
  - linguistically well-formed snippets
  - should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because ...

  ... we can quickly scan them to find the relevant document to click on.

  ... in many cases, we don't have to click at all and save time.