Introduction to Information Retrieval

(Manning, Raghavan, Schutze)

Chapter 8
Evaluation and Result Summaries

Content

- Results summaries:
  - Making our good results usable to a user

- How do we know if our results are any good?
  - Evaluating a search engine
    - Benchmarks
    - Precision and recall

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 (snippet)

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

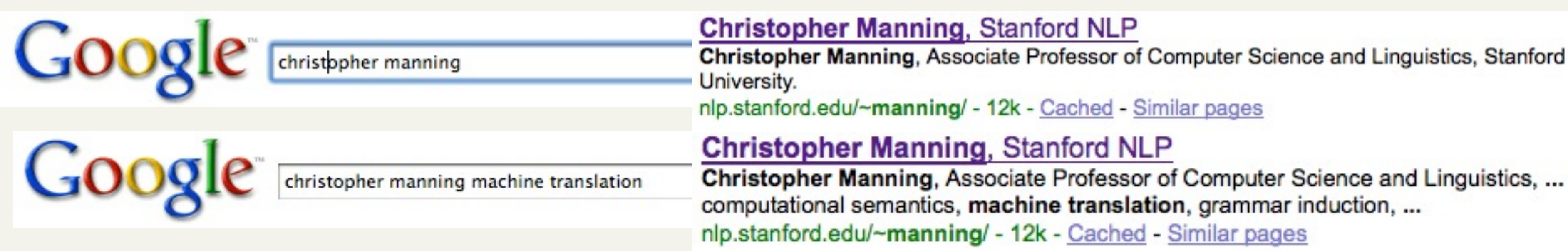
- The title is typically automatically extracted from document metadata. What about the summaries?
  - This description is crucial.
  - User can identify good/relevant hits based on description.
- Two basic kinds:
  - Static
  - Dynamic
- A **static summary** of a document is always the same, regardless 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
- Generated in conjunction with scoring
  - If query found as a phrase, all or some occurrences of the phrase in the doc
  - If not, document windows that contain multiple query terms
- The summary itself gives the entire content of the window – all terms, not only the query terms – how?

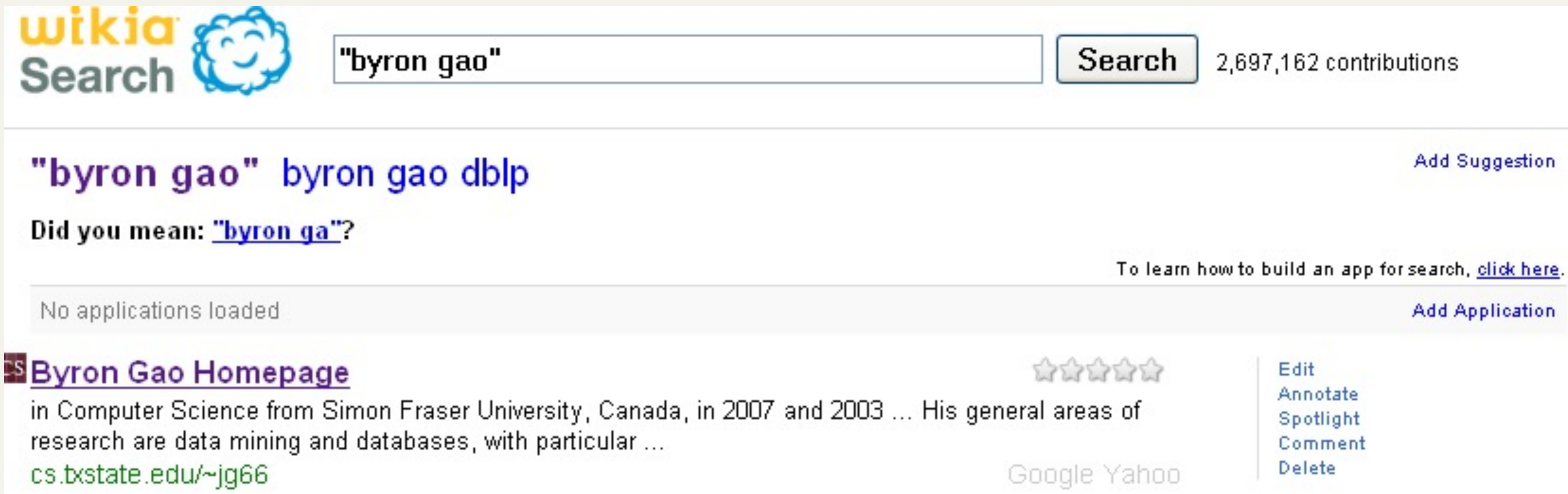


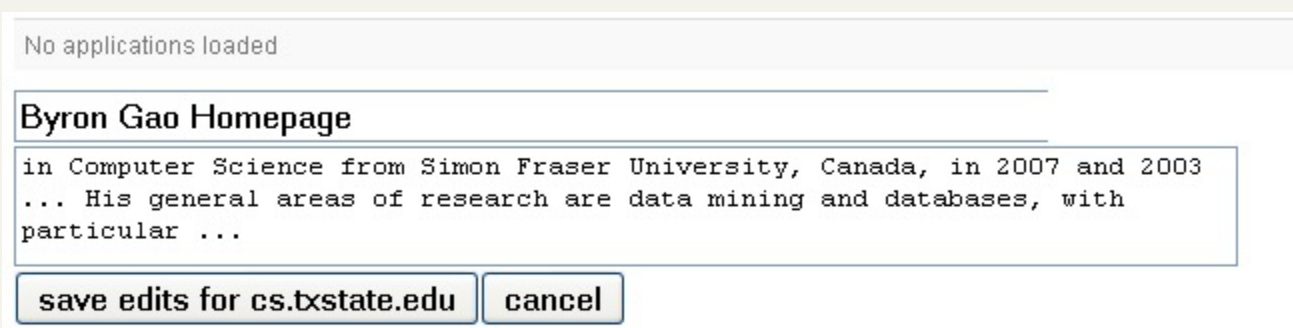
Generating dynamic summaries

- If we have only a positional index, we cannot (easily) reconstruct the context surrounding search engine hits
- If we *cache the documents* at index time, can find windows in it, cueing from 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
- Generating snippets must be fast
  - Most often, cache only a fixed-size prefix of the doc
- Note: Cached copy can be outdated
- Users really like snippets, even if they complicate IR system design

Alternative snippets

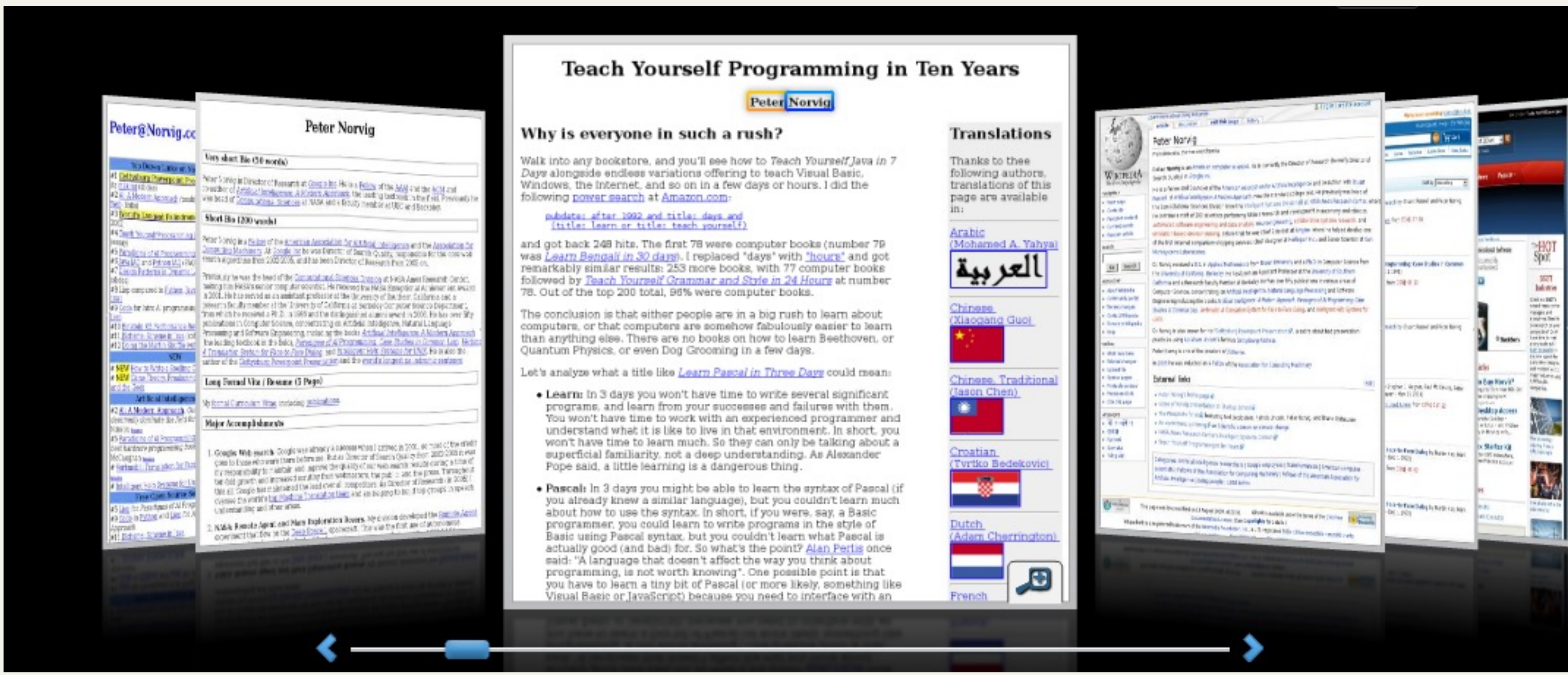
- http://search.wikia.com/
  - Mass collaboration, allow user editing





Alternative results presentations?

- An active area of HCI research
- An alternative: [http://www.searchme.com /](http://www.searchme.com/) copies the idea of Apple's Cover Flow for search results



Evaluating search engines

Measures for a search engine

- How fast does it index
  - Number of documents/hour
  - (Average document size)
- How fast does it search
  - Latency as a function of index size
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries
- Uncluttered UI
- Is it free?

Measures for a search engine

- All of the preceding criteria are *measurable*: we can quantify speed/size; we can make expressiveness precise
- However, the key measure: user happiness
  - What is user happiness?
  - Factors include: speed of response/size of index/UI …
- Elusive to measure happiness, but the most common definition is: **relevance**

How to measure relevance

- Standard methodology in IR requires 3 elements:
  1. A benchmark document collection
  2. A benchmark suite of queries
  3. A usually binary assessment of either <u>Relevant</u> or <u>Nonrelevant</u> for each query-document pair
     - Some work on more-than-binary, but not the standard

Standard relevance benchmarks

- TREC - National Institute of Standards and Technology (NIST) has run a large IR test bed for many years
- Reuters and other benchmark doc collections

- Human experts mark, for each query and for each doc, <u>Relevant</u> or <u>Nonrelevant</u>
  - or at least for subset of docs that some system returned for that query

TREC

- TREC Ad Hoc task from first 8 TRECs is standard IR task
  - 50 detailed information needs a year
  - Human evaluation of pooled results returned
  - More recently other related things: Web track, HARD
- A TREC query (TREC 5)

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>

Standard relevance benchmarks: Others

- 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

Relevance to what?

- Relevance is assessed relative to the **information need** *not* the **query**
- E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- Query: ***wine red white heart attack effective***
- You evaluate whether the doc addresses the information need, not whether it has these words

- Our terminology is sloppy: we talk about query-document relevance judgment although we mean information-need-document relevance judgment

Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved)
- **Recall**: fraction of relevant docs that are retrieved = P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Should we instead use the accuracy measure for evaluation?

- Given a query, an engine classifies each doc as "Relevant" or "Nonrelevant"
- The **accuracy** of an engine: the fraction of these classifications that are correct
- **Accuracy** is a commonly used evaluation measure in machine learning classification work

- Why is this not a very useful evaluation measure in IR?

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

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

Precision/Recall tradeoff

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved

- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation

A combined measure: *F*

- Combined measure that assesses precision/recall tradeoff is **F measure**:

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

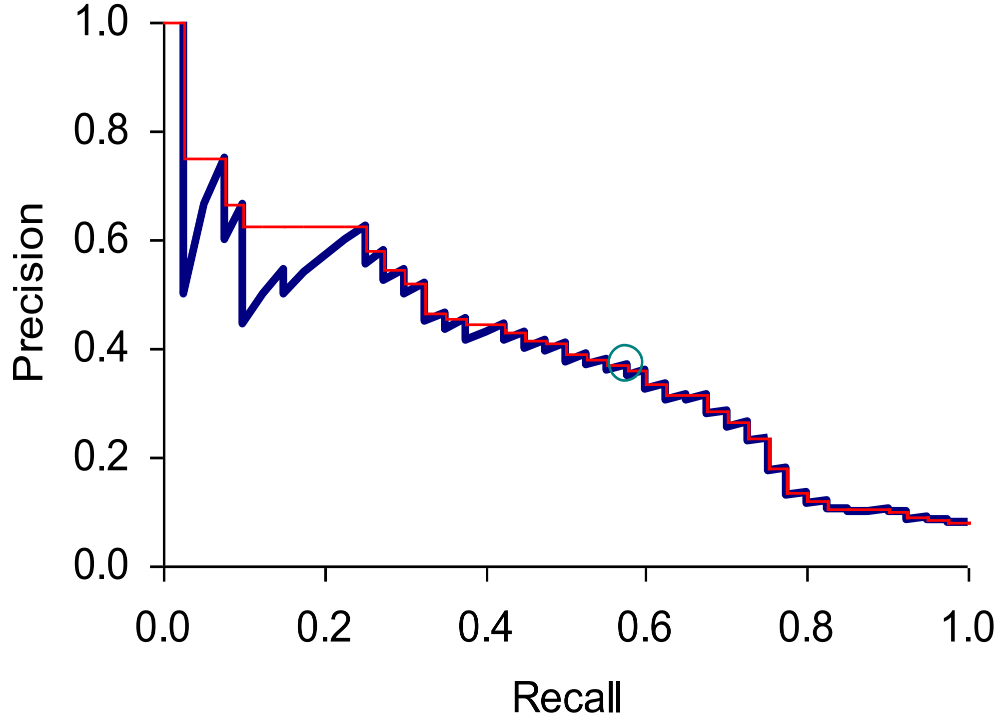
- Weighted harmonic mean of P and R: $\frac{1}{F} = \alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}$
- People usually use balanced *F* measure
  - F_1 ; or $F_{\beta=1}$;
  - with $\beta = 1$ or $\alpha = ½$; harmonic mean: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- $\beta < 1$ emphasizes P or R?
- Either P or R is bad -> bad F

Evaluating ranked results

- P/R/F are measured for unranked sets
- We can easily turn set measures into measures for ranked results
  - The system can return any number of results
  - Just use the set measures for each "prefix", the top 1, top 2, top 3, top 4, etc., results
  - Doing this for precision and recall produces a *precision-recall curve*, where a "prefix" corresponds to a level of recall

A precision-recall curve

- Sawtooth shape:
  - If the (k+1)th doc is non-relevant, R is the same as for the top k docs, but P has dropped
  - If it is relevant, then both P and R increase, and the curve jags up and to the right



- Often useful to remove the jiggles: interpolation
  - Take maximum precision of all future points

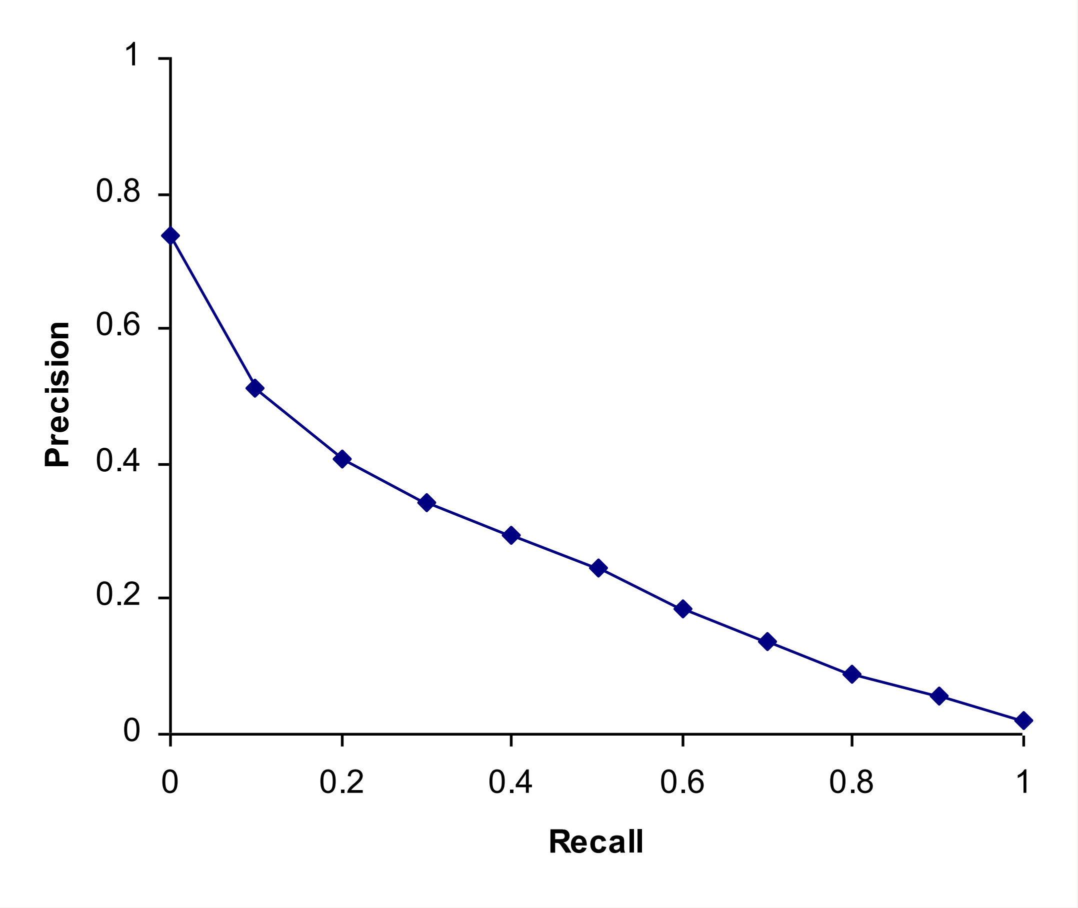
11-point interpolated average precision

- Entire precision-recall graph is very informative, but there is often a desire to boil this information down to a few numbers, even a single number
- 11-point interpolated average precision
  - The standard measure in the early TREC competitions
  - take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation, and average over queries
  - Evaluates performance at all recall levels

Recall	Interpolated 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

Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



Mean average precision (MAP)

- Recently, other measures have become more common. Most standard among TREC community is MAP
  - A single figure measure of quality across recall levels
  - Good discrimination and stability
- For a single information need, *average precision* is the average of precision value obtained for the top k docs each time a relevant doc is retrieved
  - Approximates the area under the un-interpolated precision-recall curve
- Then, this value (average precision) is averaged over many information needs to get MAP
  - Approximates the average area under the precision-curve for a set of queries

Yet more evaluation measures…

- The above ones factor in precision at all recall levels
- For many prominent applications, e.g., web search, this may not be appropriate, where what matters is rather how many good results there are on the first page or the first 3 pages!
  - Leads to measuring precision at fixed low levels (e.g., 10 or 30) of retrieved results
- Precision at k: precision of top k results
  - Standard for web search
  - Cons: the least stable among commonly used measures; does not average well because the total number of relevant docs for a query has strong influence on precision at k
- R-precision alleviates this problem
  - But may not be feasible for web search

R-precision

- If have known (though perhaps incomplete) set of relevant documents of size *Rel,* then calculate precision of top *Rel* docs returned
  - Averaging the measure across queries makes more sense
- If there are |*Rel*| relevant docs for query, we examine the top |*Rel*| results, and find r are relevant. Then,
  - recall = precision = r / |*Rel*|
  - Thus, R-precision is identical to the *break-even point*
- Empirically, highly correlated with MAP

Note

- In practice, many queries are used in evaluating a system.
  - Need to take average
- In assignments/exams, maybe only one
  - Average over this one query

Critique of pure relevance

- Assumption: relevance of one doc is treated as independent of relevance of other docs in the collection
  - But a document can be redundant (e.g., duplicates) even if it is highly relevant
  - Duplicates
- Marginal Relevance: concerns whether a doc still have distinctive usefulness after the user has looked at certain other documents … (Carbonell and Goldstein 1998)
- Maximizing marginal relevance requires returning documents that exhibit diversity and novelty

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k, 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.