Digital Libraries: Ranked Evaluation

Raffaella Bernardi

Università degli Studi di Trento

P.zza Venezia, Room: 2.05, e-mail: bernardi@disi.unitn.it

Contents

1. Why Ranking

- Problems with *no ranked retrieval*:
  - Users want to look at a few results – not thousands.
  - It's very hard to write queries that produce a few results.
  - Even for expert searchers
  - → Ranking is important because it effectively *reduces a large set of results to a very small one.*
- Even if users get more data, they only look at a few results. In the vast majority of cases they only examine 1, 2, or 3 results.

1.1. Empirical investigation of the effect of ranking

How can we measure how important ranking is?

Observe what searchers do when they are searching in a controlled setting

- Videotape them
- Ask them to “think aloud”
- Interview them
- Eye-track them
- Time them
- Record and count their clicks

The following slides are from Dan Russell’s JCDL talk – Dan Russell is the “Über Tech Lead for Search Quality & User Happiness” at Google.

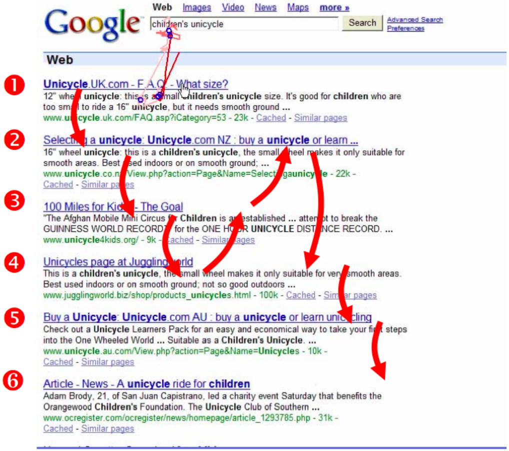
Rapidly scanning the results

Note scan pattern:

Page 3:
Result 1
Result 2
Result 3
Result 4
Result 3
Result 2
Result 4
Result 5
Result 6 <click>

Q: Why do this?

A: What's learned later influences judgment of earlier content.



Web Images Video News Maps more »

children's unicycle Search Advanced Search Preferences

Web

1 Unicycle.UK.com - F.A.Q. - What size?
12" wheel unicycle: this is a small children's unicycle size. It's good for children who are too small to ride a 16" unicycle, but it needs smooth ground ...
www.unicycle.uk.com/FAQ.asp?iCategory=53 - 23k - Cached - Similar pages

2 Selecting a unicycle: Unicycle.com NZ : buy a unicycle or learn ...
16" wheel unicycle: this is a children's unicycle, the small wheel makes it only suitable for smooth areas. Best used indoors or on smooth ground; ...
www.unicycle.co.nz/View.php?action=Page&Name=Selectingaunicycle - 22k -
Cached - Similar pages

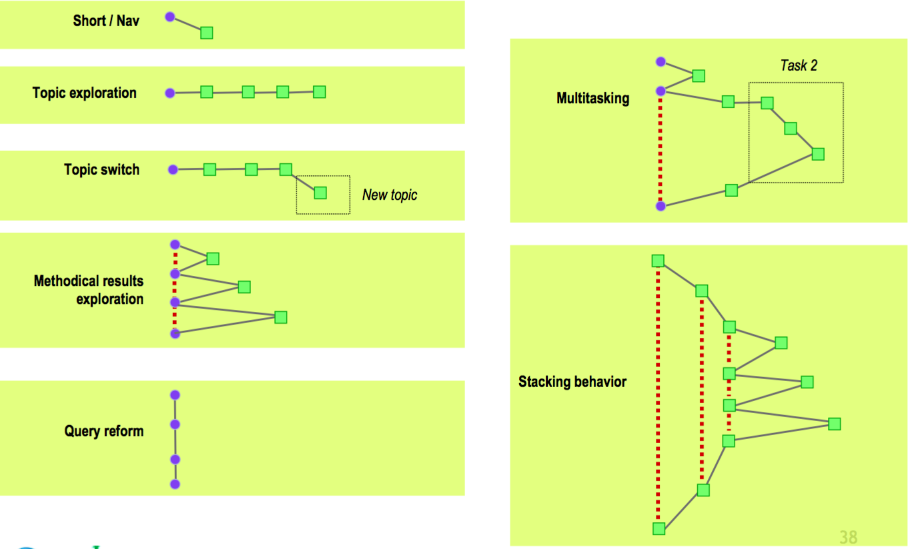
3 100 Miles for Kids - The Goal
"The Afghan Mobile Mini Circus for Children is an established ... attempt to break the GUINNESS WORLD RECORD for the ONE HOUR UNICYCLE DISTANCE RECORD. ...
www.unicycle4kids.org/ - 9k - Cached - Similar pages

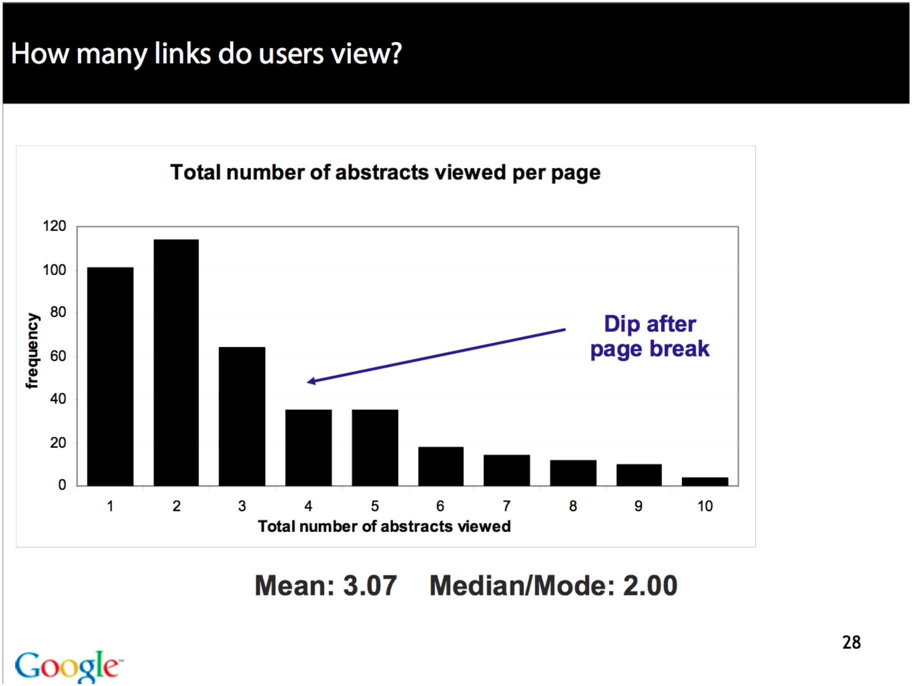
4 Unicycles page at Jugglingworld
This is a children's unicycle, the small wheel makes it only suitable for very smooth areas. Best used indoors or on smooth ground; not so good outdoors ...
www.jugglingworld.biz/shop/products_unicycles.html - 100k - Cached - Similar pages

5 Buy a Unicycle: Unicycle.com AU : buy a unicycle or learn unicycling
Check out a Unicycle Learners Pack for an easy and economical way to take your first steps into the One Wheeled World ... Suitable as a Children's Unicycle. ...
www.unicycle.au.com/View.php?action=Page&Name=Unicycles - 10k -
Cached - Similar pages

6 Article - News - A unicycle ride for children
Adam Brody, 21, of San Juan Capistrano, led a charity event Saturday that benefits the Orangewood Children's Foundation. The Unicycle Club of Southern ...
www.ocregister.com/ocregister/news/homepage/article_1293785.php - 31k -
Cached - Similar pages

Kinds of behaviors we see in the data





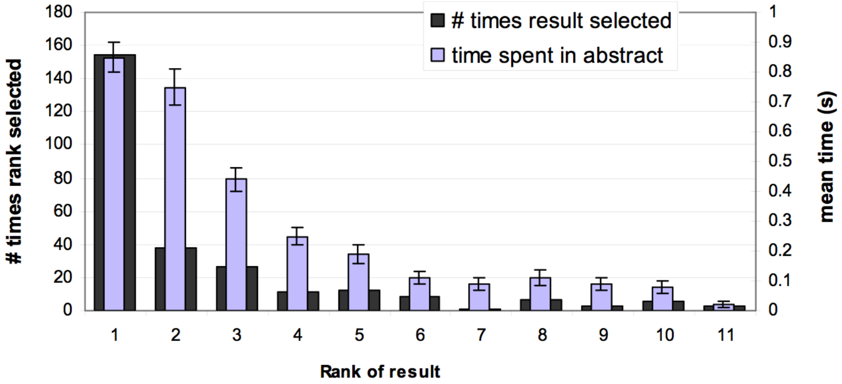
How many links do users view?

Total number of abstracts viewed per page

Dip after page break

Mean: 3.07 Median/Mode: 2.00

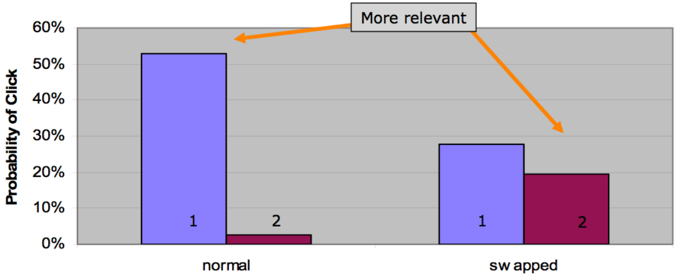
Looking vs. Clicking



- Users view results one and two more often / thoroughly
- Users click most frequently on result one

Presentation bias – reversed results

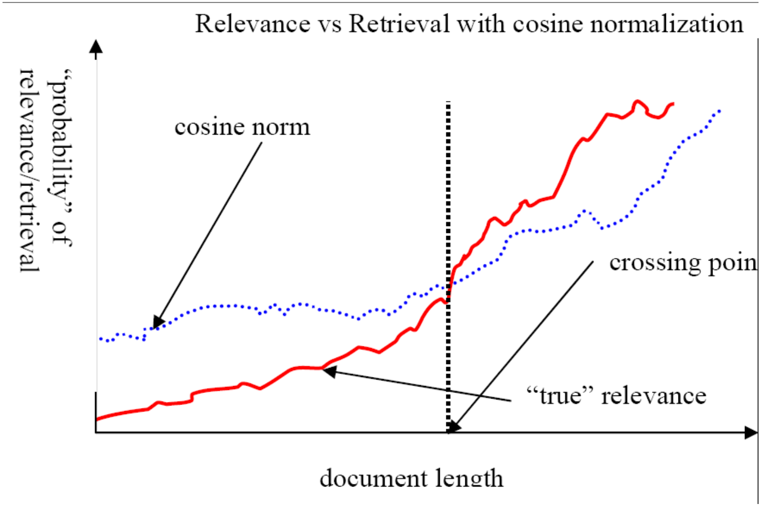
- Order of presentation influences where users look **AND** where they click



1.2. Importance of ranking: Summary

- *Viewing abstracts:* Users are a lot more likely to read the abstracts of the top-ranked pages (1, 2, 3, 4) than the abstracts of the lower ranked pages (7, 8, 9, 10).
- *Clicking:* Distribution is even more skewed for clicking.
- In 1 out of 2 cases, users click on the top-ranked page.
- Even if the top-ranked page is not relevant, 30% of users will click on it.
- → Getting the *ranking right* is very important.
- → Getting the *top-ranked* page right is most important.

2. Predicted and true probability of relevance



Hence cosine similarity normalization has been fine tuned.

3. Evaluation

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair

3.1. Relevance: query vs. information need

- Relevance to *what?*
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- *Information need i*: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
- *Query q*: [red wine white wine heart attack]
- Consider document d': *At heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' is an excellent match for query q . . .
- d' is *not* relevant to the information need i.

3.2. Relevance: query vs. information need

- User happiness can only be measured by relevance to an information need, not by relevance to queries.
- Our terminology is sloppy in these slides: we talk about query-document relevance judgments even though we mean information-need-document relevance judgments.

3.3. Reminder: Precision and Recall

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

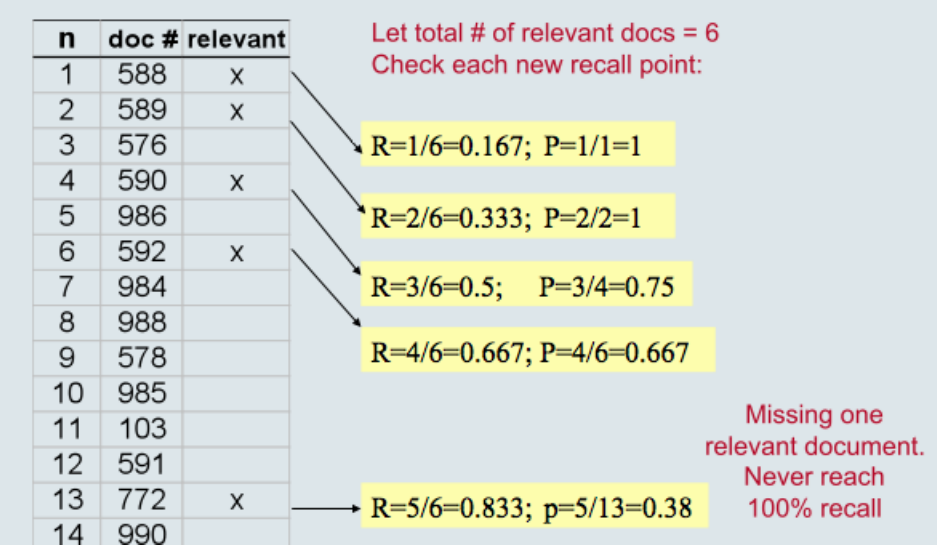
$$\begin{aligned} P &= TP/(TP+FP) \\ R &= TP/(TP+FN) \end{aligned}$$

- You can increase recall by returning more docs. docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.

4. Ranked evaluation

- Precision/recall/F are measures for *unranked sets*.
- We can easily turn set measures into measures of *ranked lists*.
- Just compute the set measure for each "prefix": the top 1, top 2, top 3, top 4 etc results
- Doing this for precision and recall gives you a *precision-recall curve*.

4.1. Precision/Recall: at position



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

$P = TP/(TP + FP)$ $R = TP/(TP + FN)$

$P = TP/\text{retrieved}$ $R = TP/\text{relevant}$

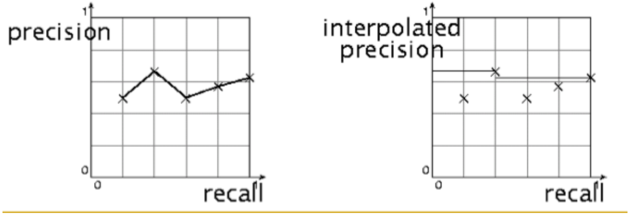
4.2. Interpolated Precision

Note: if the $(k+1)^{th}$ document is not relevant, then recall is the same as for the top k documents, but precision has dropped; if it is relevant, then both precision and recall increase, and the curve jags up and to the right.

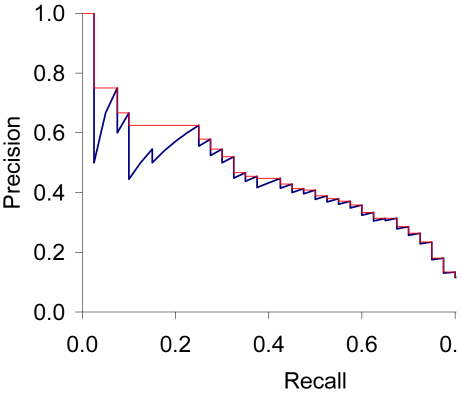
It's good to remove these jiggles. The standard way to do so is with *interpolated precision* (P_{interp})

$$P_{interp}(r) = max_{r' \geqslant r}\, P(r')$$

the interpolated precision at a certain recall level r is defined as the highest precision found for any recall level $r' \geqslant r$.



4.3. Precision-recall curve



- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \ldots$).
- Interpolation (in red): Take maximum of all future points
- Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.

4.4. From P-R curve to measure

Examining the entire P-R curve is very informative, but having a *single measure* is more desirable.

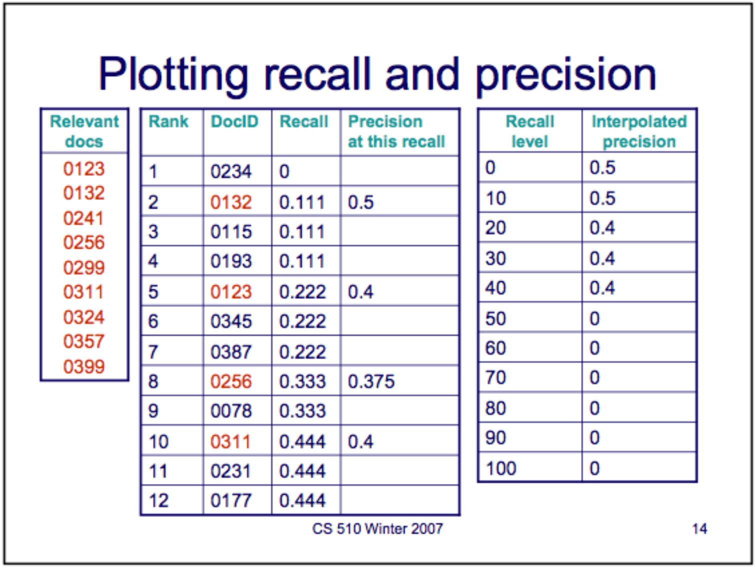
The traditional way to obtain this is the *eleven-point interpolated average precision*:

The interpolated precision is measured at *11 recall levels* of 0.0, 0.1., 0.2 . . . 1.0. (0, 10, 20 . . . 100 percent)

For each recall level, we then calculate the *arithmetic mean* of the interpolated precision at that recall level for each information need in the test collection.

4.5. 11-point interpolated precision

P = TP/retrieved, R = TP/relevant relevant = 9



Plotting recall and precision

Relevant docs
0123
0132
0241
0256
0299
0311
0324
0357
0399

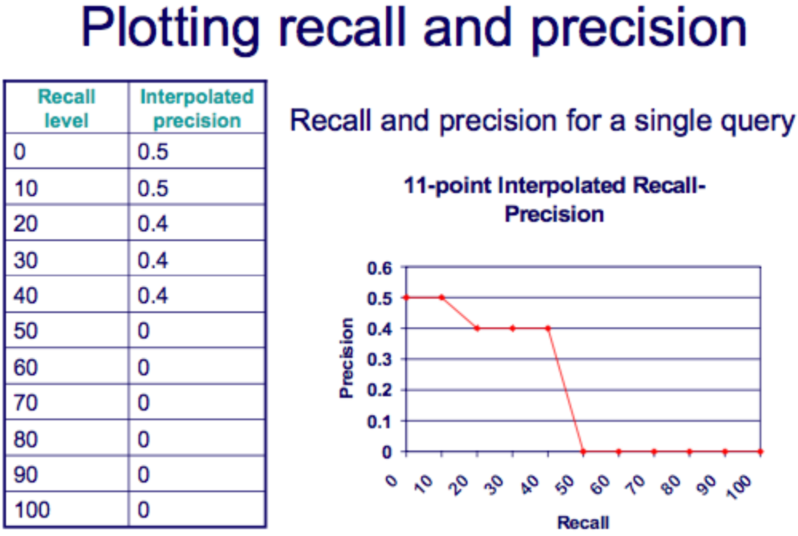
Rank	DocID	Recall	Precision at this recall
1	0234	0	
2	0132	0.111	0.5
3	0115	0.111	
4	0193	0.111	
5	0123	0.222	0.4
6	0345	0.222	
7	0387	0.222	
8	0256	0.333	0.375
9	0078	0.333	
10	0311	0.444	0.4
11	0231	0.444	
12	0177	0.444	

Recall level	Interpolated precision
0	0.5
10	0.5
20	0.4
30	0.4
40	0.4
50	0
60	0
70	0
80	0
90	0
100	0

CS 510 Winter 2007 14

(c) Susan Price and David Maier

4.6. 11-point interpolated precision: Graph



Plotting recall and precision

Recall level	Interpolated precision
0	0.5
10	0.5
20	0.4
30	0.4
40	0.4
50	0
60	0
70	0
80	0
90	0
100	0

Recall and precision for a single query

11-point Interpolated Recall-Precision

4.7. 11-point interpolated average precision

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

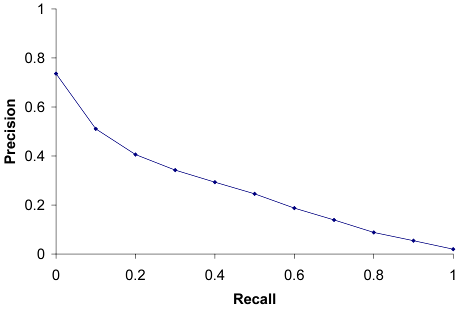
11-point average: ≈ 0.425

100+67+63+55+45+41+36+29+13+10+8 = 467
467/11 = 42.45454545454545454545

4.8. Average over quries

Single query performance is not necessarily representative of the system's performance:

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
- The curve below is typical of performance levels at TREC (50 queries).



4.9. Mean Average Precision

For *one information need*, it is the average of the precision value obtained for the top k documents each time a relevant document is retrieved.

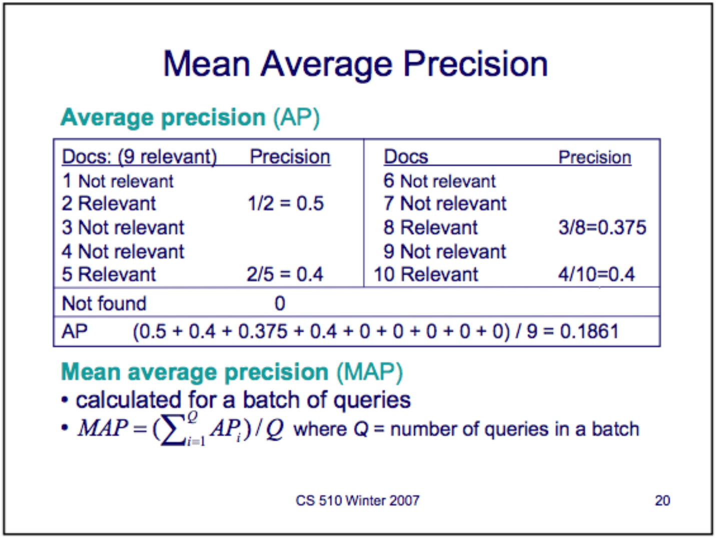
- *No* used of *fixed recall levels*. No interpolation.
- When no relevant doc is retrieved, the precision value is taken to be 0.

The MAP value for a *test collection* is then the arithmetic mean of MAP values for individual information needs. Each query counts equally.

MAP scores vary widely across information needs.

This means that a set of test information needs must be *large* and *diverse* enough to be representative of system effectiveness across different queries.

4.10. MAP: Example



Mean Average Precision

Average precision (AP)

Docs: (9 relevant)	Precision	Docs	Precision
1 Not relevant		6 Not relevant	
2 Relevant	1/2 = 0.5	7 Not relevant	
3 Not relevant		8 Relevant	3/8=0.375
4 Not relevant		9 Not relevant	
5 Relevant	2/5 = 0.4	10 Relevant	4/10=0.4
Not found	0		
AP	(0.5 + 0.4 + 0.375 + 0.4 + 0 + 0 + 0 + 0 + 0) / 9 = 0.1861		

Mean average precision (MAP)

- calculated for a batch of queries
- $MAP = (\sum_{i=1}^{Q} AP_i)/Q$ where Q = number of queries in a batch

CS 510 Winter 2007 20

4.11. Precision and Recall at position k

Some time what matters is how many good results there are on the first page or first three pages.

This brings to other measures based on fixed low level of retrieved results (10 or 30 documents.).

Precision@k Precision on the top k retrieved documents. It's appropriate for Web search engines: Most user scan only the first few (e.g. 10) hyperlinks that are presented.

Recall@k Recall on the top k retrieved documents. Appropriate for archival retrieval systems: what fraction of total number of relevant documents did a user find after scanning the first (e.g. 100) documents?

Advantage: there is no need to estimate the size of the set of relevant documents. *Disadvantage*: it is the least stable measure.

4.12. R-Precision

Alternative. It requires having a set of known relevant document (*Rel*), (though perhaps incomplete).

It computes the precision on the top *Rel* retrieved documents.

It's better to average this measure over queries.

R-Precision is highly correlated with MAP.

4.13. Variance of measures like precision/recall

- For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).

- Indeed, it is usually the case that the variance of the same system across queries is much greater than the variance of different systems on the same query.

- That is, there are easy information needs and hard ones.

4.14. Evaluation Measures: Summing up

Evaluation of effectiveness based on Relevance:

Ranking-Ignorant Measures Accuracy, Precision and Recall, F measure.

Ranking-Aware Measures Precision and Recall curve, 11 Point, MAP, P/R@k, R-Precision, PRBEP, ROC Curve.

5. Evaluation benchmarks

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - . . . which we will often incorrectly refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.
- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.

5.1. Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs
- Too small, too untypical for serious IR evaluation today

5.2. Standard relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- Best known: TREC Ad Hoc, used for first 8 TREC evaluations between 1992 and 1999
- 1.89 million documents, mainly newswire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
- Rather, NIST assessors' relevance judgments are available only for the documents that were among the top *k* returned for some system which was entered in the TREC evaluation for which the information need was developed.

5.3. Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Used to be 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.
- CLEF-TEL.
- Trento: http://www.celct.it/ (E. Pianta)

5.4. Validity of relevance assessments

- Relevance assessments are only usable if they are consistent.
- If they are not consistent, then there is no “truth” and experiments are not repeatable.
- How can we measure this consistency or agreement among judges?
- → Kappa measure

5.5. Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement (e.g. marginal statistics: sum up raw/column)
- $P(A)$ = proportion of time judges agree
- $P(E)$ = what agreement would we get by chance

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

Values of κ in the interval

- $[0.8 - 1]$ (good agreement),
- $[0.67 - 0.8]$ (fair agreement),
- $[\cdot - 0.67]$ (dubious basis for an evaluation).

5.6. Calculating the kappa statistic

		Judge 2 Relevance		
		Yes	No	Total
Judge 1	Yes	300	20	320
Relevance	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 $\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$
(still in acceptable range)

5.7. Interjudge agreement at TREC

information need	number of docs judged	disagreements
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

5.8. Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
- *Virtually no impact on ranking of systems*
- Supposes we want to know if algorithm A is better than algorithm B
- An information retrieval experiment will give us a reliable answer to this question even if there is a lot of disagreement between judges.

5.9. 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

5.10. Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The *marginal relevance* of a document at position k in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \dots d_{k-1}$.
- But even if a document is highly relevant its information can be completely redundant with other documents that have already been considered (e.g. duplicates!).

6. Result summaries

How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
- This description is crucial. The user often can identify good hits (= relevant hits) based on the description.
- No need to “click” on all documents sequentially

Doc description in result list Most commonly: doc title, url, some metadata . . . and a summary. How do we “compute” the summary (not considered in this course.)

7. General comment

- Systems usually have various weights (*parameters*) that can be adjusted to tune system performance.
- It is wrong to report results on a test collection that were obtained by tuning these parameters to maximize performance on that collection.
- Such tuning overstates the expected performance of the system, because the weights will be set to *maximize the performance on one particular set of queries* rather than for a random sample of queries.
- The correct procedure is to have one or more development test collection, and to tune the parameters on the development test collection.
- The tester then runs the system with those weights on the test collection and reports the results on that collection as an *unbiased* estimate of performance.