11-442 / 11-642:
Search Engines

Evaluating Search Effectiveness

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Introduction to Evaluation

Given two methods, which produces better search results?

Evaluation is important to information retrieval R&D

- The theory is weak, so the field is driven by measurement
- Improved theory is good
- Improved results in several experiments is better
  - – Even if the theory is a little dodgy

Web search companies run experiments constantly

- When you search, you are a subject in an experiment

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

The Cranfield Methodology

The experimental methodology:

1. **Obtain a corpus of documents**
2. **Obtain a set of information needs**
   – Sometimes expressed as queries, sometimes as descriptions
3. **Obtain relevance judgments**
   – Which documents satisfy each information need
4. **Measure how well each method finds relevant documents**
   – Use multiple metrics to get different perspectives
5. **Compare the effectiveness of the different methods**

Relevant

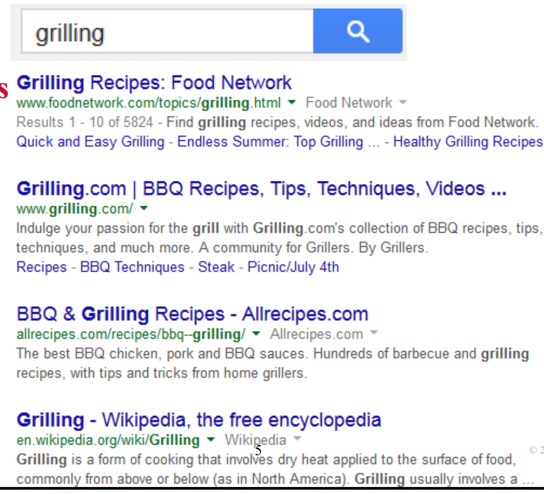
The Cranfield Methodology

Information Need Kabob recipes for dinner tonight

Manual Assessments

Query

Documents



Relevant

Not Relevant

Not Relevant

Not Relevant

5

The Cranfield Methodology

The Cranfield experimental methodology has 5 parts

- Documents
- Information needs (maybe expressed as queries)
- Relevance judgments
- Metrics
- Comparison of methods

A test collection

6

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

Test Collections

The test collection represents a real information seeking task

- A set of documents (a “corpus”)
- Typical information needs
  - And often typical queries that represent information needs
- Relevance assessments
  - The searcher’s opinion about whether the document satisfies the information need

The test collection should be as realistic as possible

- This seems obvious, but it is often neglected
- E.g., use Twitter users to do Twitter relevance assessments

Test Collections: Documents

Many standard document collections are available

- **News:** WSJ 87-92, NY Times, LA Times, Reuters 2001, …
- **Enterprise:** Tobacco litigation, patent, CSIRO, W3C, …
- **XML:** Scientific papers, patents, …
- **Email:** Enron email
- **Web:** VLC, wt10g, gov2, ClueWeb09, ClueWeb12, …
- **Wikipedia**
- **Social media:** blog06, blog08, tweets11, KBA1, KBA2, …
- **Languages:** English, European, Asian, Hindi, Arabic, …
- **…**

9

Test Collections: Documents

Types of documents

- Excellent coverage of news data
- Some coverage of enterprise data
- Good coverage of web and blog documents

Typically 50-200 information needs per document collection

- More on this later …

These collections are very useful, but each has its biases

- You must understand the biases of the collections you use

10

Test Collections: Information Needs

A person uses a search engine to satisfy an information need

- The information need is hidden in a person's head
- The query is a clue to the information need
  - The search engine may have other clues, too
  - E.g., user behavior, user history, population behavior, …
- But … the information need is never known precisely

Example information needs

- How does a septic system work?
- Kabob recipes for dinner tonight.

Test Collections: How Are Information Needs Obtained?

Ask typical users that are trying to address a real problem

- The best option, if you have access to typical users

Observe typical users trying to address real problems

- E.g., obtain a search log that contains queries, clicks, …
- Create information needs that are consistent with observations

Guess what typical users want

- Search the corpus to see what kinds of documents it has
- Create information needs that are satisfied by those documents
- The weakest option, but often the only option available

Test Collections: Relevance Assessments

A document is relevant if a person judges it to be useful in the context of a specific information need

- Different people define "useful" differently
- One person will define "useful" differently at different times
- The judgment depends upon more than the document and query
  - E.g., what the person knew before reading the document

Relevance is subjective, not objective

- It depends upon a specific individual

These are really important concepts

Test Collections: Which Documents are Relevant?

Query: Skiing near Pittsburgh

1. Skiing and Snowboarding in Western Pennsylvania – Pittsburgh
2. Boyce Park Ski Area – Allegheny County.
3. Hidden Valley Resort – Official Site
4. Best Cross Country Ski Trails Around Pittsburgh
5. Best Ski Slopes For Kids Near Pittsburgh – CBS Pittsburgh
6. Ski Resorts in Pittsburgh, Pennsylvania – USA Today
7. Skiing and Snowboarding near Pittsburgh, PA
8. Pittsburgh Cross Country Skiing
9. Fox Chapel Ski and Board
10. Cross-Country Skiing & SnowShoeing – Laurel Highlands

Test Collections: Which Documents are Relevant?

Query: Skiing near Pittsburgh

✓ 1. Skiing and Snowboarding in Western Pennsylvania – Pittsburgh
× 2. Boyce Park Ski Area – Allegheny County.
✓ 3. Hidden Valley Resort – Official Site
× 4. Best Cross Country Ski Trails Around Pittsburgh
× 5. Best Ski Slopes For Kids Near Pittsburgh – CBS Pittsburgh
? 6. Ski Resorts in Pittsburgh, Pennsylvania – USA Today
? 7. Skiing and Snowboarding near Pittsburgh, PA
× 8. Pittsburgh Cross Country Skiing
× 9. Fox Chapel Ski and Board
× 10. Cross-Country Skiing & SnowShoeing – Laurel Highlands

15

Test Collections: Which Documents Should be Judged?

Exhaustive assessment

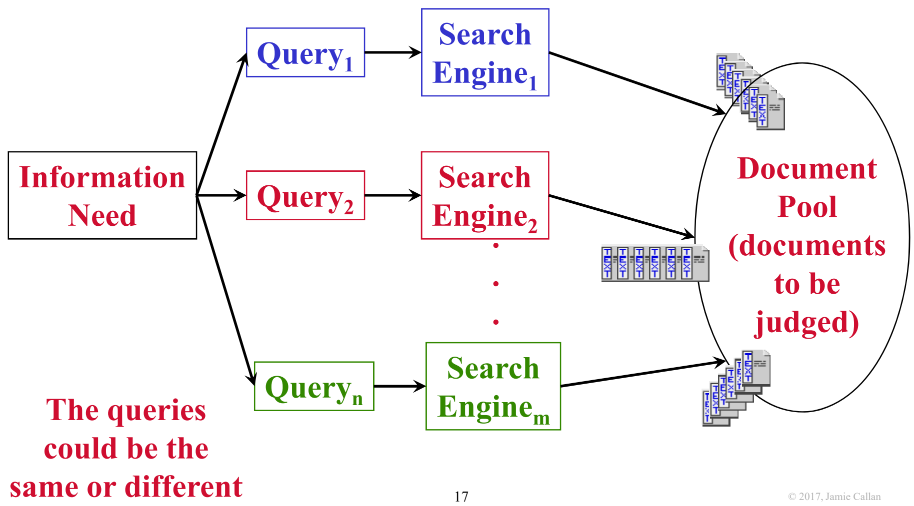
- A common approach from 1960-1990
  - – Test collections were small, so this was feasible
- Evaluate <u>all</u> documents for each query
- The baseline against which other methods are compared

Sample-based assessment ("pooling")

- Typical since 1991
- Combine results from <u>multiple</u> systems ("sample", "pool")
- Evaluate only documents in the sample ("pool")

16

Test Collections: Pooling



Test Collections: Pooling

Retrieve documents using <u>multiple</u> techniques

- Choose diverse and effective techniques, to reduce sample bias

Judge the top *n* documents for each technique

The relevant set is the union of all documents judged relevant

- This is a <u>subset</u> of the true relevant set
- The size of true relevant set can be estimated by sampling

<u>Most</u> documents are not judged

- The metrics must decide how to handle unjudged documents

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

Metrics

The IR community uses a large set of metrics to assess search engine accuracy and effectiveness

Why so many?

- Different metrics examine different types of behavior
- Different situations require different types of behavior

Today we consider several popular metrics…

Metrics: Precision and Recall

Precision and Recall measure the quality of a set

$$P = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|} = \frac{6}{10} = 60\%$$

$$R = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant}|} = \frac{6}{9} = 67\%$$

Retrieved	Not Retrieved
Relevant	Relevant
Not relevant	Relevant
Relevant	Relevant
Relevant	
Not relevant	
Relevant	
Not relevant	
Relevant	
Relevant	
Not relevant	

Search results (e.g., top 10 documents) Order is unimportant

Metrics: Precision and Recall for Ranked Retrieval

Precision and Recall are set-based measures

- In ranked retrieval, the entire collection is ranked (in theory)
- It makes no sense to calculate P & R for the entire collection

Decisions

- Where in the ranking to measure Precision & Recall
- How to combine measures from different points in the ranking

Common methods

- P@n
- Mean average precision

P@n

Precision at rank n (P@n) is a popular metric

- Easy to compute, easy to understand

P@n doesn't normalize for query difficulty

How does it behave for different queries?

- Easy query, many relevant docs
- Easy query, few relevant docs
- Hard query, few relevant docs

P@n isn't as stable as MAP (covered later)

+
-
+
+
-
+
+
-

P@5

23

Metrics: F-Measure

It is often convenient to have a single measure of effectiveness

- F-measure (Harmonic Mean) of Precision and Recall

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

- If Precision and Recall are weighted equally ($\alpha = 0.5$)

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R}$$

- F measure is used for <u>set-based</u> evaluation

24

Metrics:
Averaging Results for Multiple Queries

Micro average: Average results across documents

- Each document is equally important
  - Queries with many relevant documents dominate
- Common in machine learning, but not IR
  - Our class distribution is much more skewed

Macro average: Average results across queries

- Each query is equally important
- Most common averaging method for ad-hoc retrieval

Metrics:
Ranked Retrieval

P, R, P@n, and F_1 are defined for a set of documents

- Appropriate for the unranked Boolean retrieval model
- Appropriate for text categorization
- Less useful for ranked retrieval models

Usually we want metrics that apply to a document ranking

- Several popular metrics extend P & R to rankings
  - Average Precision (AP)
  - Mean Average Precision (MAP)
  - Interpolated Average Precision (no longer used much)

Metrics: Average Precision & Mean Average Precision

Mean Average Precision (MAP) is a popular summary metric

- **Average Precision**
  - Measure P at each relevant document for the i^{th} query
  - Average the measurements for the i^{th} query
- **Mean Average Precision**
  - The mean of the Average Precision values for all queries

Metrics: Average Precision and Mean Average Precision

Query 1

Rank	Rel?	P	R
1	**Y**	1.00	0.25
2	**Y**	1.00	0.50
3	N	0.67	0.50
4	**Y**	0.75	0.75
5	N	0.60	0.75
6	**Y**	0.67	1.00
7	Ignore everything after R = 1.00		
8			
9			
10			

4 relevant docs in corpus
Average Precision = 0.855

Query 2

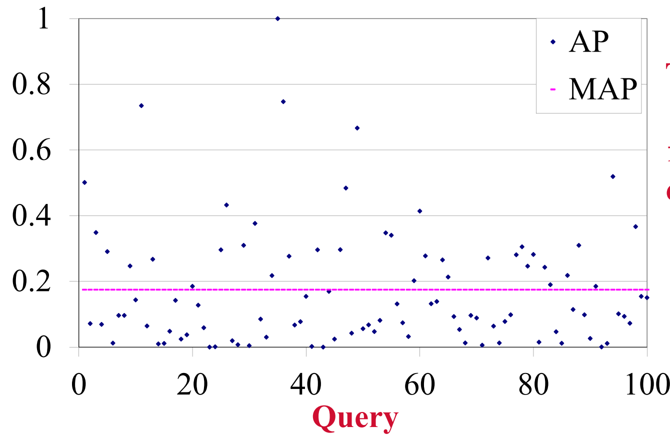
Rank	Rel?	P	R
1	N	0.00	0.0
2	**Y**	0.50	0.20
3	N	0.33	0.20
4	N	0.25	0.20
5	**Y**	0.40	0.40
6	**Y**	0.50	0.60
7	N	0.43	0.60
8	N	0.38	0.60
9	**Y**	0.44	0.80
10	**Y**	0.50	1.00

5 relevant docs in corpus
Average Precision = 0.468

Mean Average Precision
0.6615
(macro average)

Metrics:
Average Precision & Mean Average Precision

Mean Average Precision (MAP) hides a lot of variance in AP



Typical results for 100 queries

Metrics:
Mean Average Precision (MAP)

Why is Mean Average Precision (MAP) is popular?

- Single-value metrics are convenient
- MAP has been considered <u>more robust</u> than other metrics
  - If MAP (A) > MAP (B)
    Then A is likely to be better than B across <u>other metrics</u>, too
  - This is not necessarily true for other metrics, e.g., P@10

MAP is the most widely-used metric

- But, NDCG (covered later) and other metrics are slowly taking over...

Metrics: Mean Reciprocal Rank (MRR)

Sometimes we care only about the first relevant document

- E.g., retrieving home pages

Reciprocal rank

1 / rank of first relevant document

Mean reciprocal rank (MRR)

- Average of reciprocal rank values across a set of queries

A very common metric

Metrics: Characteristics of Web Search Behavior

Lower-ranked documents are less likely to be viewed

- Independent of relevance

When there are many relevant documents, graded relevance is more useful

- The best choice
- A very good choice
- Acceptable (relevant)
- Not relevant

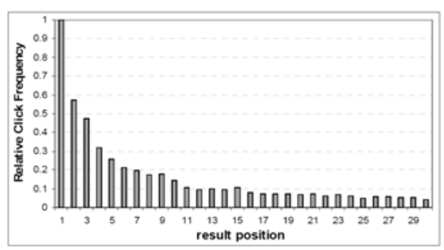


Figure 3.1: Relative click frequency for top 30 result positions over 3,500 queries and 120,000 searches.

(Agichtein, et al, 2006)

Metrics Reconsidered

Which method is better?

- AP and MAP consider them equally good

In Precision-oriented tasks, the right result is preferred

- E.g., web search

In Recall-oriented tasks, they may be equally useful

- E.g., legal search

	Method 1		Method 2	
Rank	**Rel?**	**P@n**	**Rel?**	**P@n**
1				
2				
3			Y	0.33
4	Y	0.25		
5				
6	Y	0.33		
7				
8			Y	0.25
AP		0.29		0.29

33

Metrics: Normalized Cumulative Discounted Gain

NDCG is a popular summary statistic that use multi-valued relevance assessments to measure the quality of a ranking

$$NDCG@k = Z_k \sum_{i=1}^{k} \frac{2^{R_i} - 1}{\log(1+i)}$$

← **Gain (based on relevance)**
← **Discount (based on rank)**

R_i is the relevance of the document at rank i

– E.g., 0 (non-relevant), 1 (relevant), 2 (very relevant)

Z_k normalizes so that NDCG=1 at k for a perfect ranking

– Z_k = 1 / DCG@k for the "ideal" ranking

– Required to combine scores for different queries

Popular with web search engines

34

Metrics:
Normalized Cumulative Discounted Gain

Rank i	Value R_i	Gain 2^{R_i}-1	Discount log(1+i)	Discounted Gain
1	**3**	7	0.30	23.25
2	**2**	3	0.48	6.29
3	0	0	0.60	0.00
4	**1**	1	0.70	1.43
5	0	0	0.78	0.00
6	**2**	3	0.85	3.55
7	0	0	0.90	0.00
8	0	0	0.95	0.00
9	0	0	1.00	0.00
10	0	0	1.04	0.00

DCG_{10}=34.52

Relevance scale
- **3: Best result**
- **2: Very good**
- **1: Acceptable**
- **0: Not relevant**

Ideal DCG_{10}
= 35.95.

$NDCG_{10}$
= 34.52 / 35.95
= 0.96

35

Metrics:
Rank-Biased Precision (RBP)

RBP models multi-valued relevance assessments <u>and</u> the user's persistence at examining the ranked list

$$RBP = (1-p) \cdot \sum_{i=1}^{n} R_i \cdot p^{i-1}$$

- p: A parameter that models the user's persistence
- n: Number of documents
- R_i: The relevance of the document at rank I

p (document is examined | document's rank) converges to 0
- Perhaps a more realistic model than NDCG

(Moffat and Zobel, 2008)

36

Metrics:
Rank-Biased Precision (RBP)

Rank i	Value R_i	Persistence p^{i-1}	RBP_i
1	**3**	1.000	3.000
2	**2**	0.700	1.400
3	0	0.490	0.000
4	**1**	0.343	0.343
5	0	0.240	0.000
6	**2**	0.168	0.336
7	0	0.118	0.0000
8	0	0.082	0.0000
9	0	0.058	0.0000
10	0	0.018	0.0000

Discounted Gain
23.25
6.29
0.00
1.43
0.00
3.55
0.00
0.00
0.00
0.00

Note the differences

$p = 0.7$

Metrics:
Rank-Biased Precision (RBP)

RBP is an example of a metric with a model of user behavior

- A very simple model

Recent metrics have more sophisticated models

- Of user behavior
- Of how the relevance of documents ranked higher affect the value of documents ranked lower ("cascade models")

(Moffat and Zobel, 2008)

Metrics: Summary

We covered many metrics

- Precision, Recall, P@n
  - Micro-averaging, macro-averaging
- F (and F_1)
- Average Precision (AP) and Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (NDCG)
- Rank Biased Precision (RBP)

You need to know <u>when each metric is appropriate</u> (and not)

39

Metrics: trec_eval

trec_eval is a standard evaluation tool for ad-hoc retrieval

trec_eval reports four types of information

- Basic information about the results file
- Summary statistics that apply to a complete ranking
- Quality at different positions in the ranking
- Precision at different positions in the document ranking

```
num_q all 50
num_ret all 5000
num_rel all 5061
num_rel_ret all 1082
```

```
map all 0.1825
gm_ap all 0.0707
R-prec all 0.2632
bpref all 0.2525
recip_rank all 0.6859
```

```
ircl_prn.0.00 all 0.7327
ircl_prn.0.10 all 0.4793
ircl_prn.0.20 all 0.3472
ircl_prn.0.30 all 0.2579
ircl_prn.0.40 all 0.1939
ircl_prn.0.50 all 0.1176
ircl_prn.0.60 all 0.0744
ircl_prn.0.70 all 0.0606
ircl_prn.0.80 all 0.0348
ircl_prn.0.90 all 0.0287
ircl_prn.1.00 all 0.0101
```

```
P5 all 0.5160
P10 all 0.4820
P15 all 0.4480
P20 all 0.4050
P30 all 0.3620
P100 all 0.2164
P200 all 0.1082
P500 all 0.0433
P1000 all 0.0216
```

Statistics on a <u>by-query</u> or <u>by-query-set</u> basis

40

Metrics: trec_eval

Basic information about the result file

- Primarily for documentation and error checking

Example:

num_q all 50	**There were 50 queries**
num_ret all 5000	**5,000 documents were retrieved**
num_rel all 5061	**There are 5,061 relevant documents**
num_rel_ret all 1082	**1,082 retrieved documents were relevant**

Metrics: trec_eval

Summary statistics that apply to the entire ranking

- i.e., results are averaged from different parts of the ranking

Example

map all 0.1825	**Mean average precision (MAP)**
gm_ap all 0.0707	**Avg precision using geometric mean**
R-prec all 0.2632	**R-Precision (value where P = R)**
bpref all 0.2525	**bpref (not covered)**
recip_rank all 0.6859	**Reciprocal rank (MRR)**

Metrics:
trec_eval

Quality at different points in the ranking:

Interpolated Average Precision at 11 Recall points

ircl_prn.0.00 all 0.7327	**Average Precision at 0% Recall**
ircl_prn.0.10 all 0.4793	**Average Precision at 10% Recall**
ircl_prn.0.20 all 0.3472	**Average Precision at 20% Recall**
ircl_prn.0.30 all 0.2579	**Average Precision at 30% Recall**
: : :	: : : :

Interpolated Average Precision isn't used much these days

43

Metrics:
trec_eval

Precision@n

P5 all 0.5160	**Precision at rank 5**
P10 all 0.4820	**Precision at rank 10**
P15 all 0.4480	**Precision at rank 15**
P20 all 0.4050	**Precision at rank 20**
P30 all 0.3620	**Precision at rank 30**
P100 all 0.2164	**Precision at rank 100**
P200 all 0.1082	**Precision at rank 200**
P500 all 0.0433	**Precision at rank 500**
P1000 all 0.0216	**Precision at rank 1000**

44

The Cranfield Methodology

Advantages

- Experimental conditions are clearly defined
  - Documents, information needs, relevance judgments
- Experiments can be repeated

Disadvantages

- A simple model of users
  - User is assumed to read everything
  - Relevance is independent of other documents or rank

Overview of the Evaluation Unit

Introduction to evaluation

The Cranfield methodology

- Overview and introduction
- Test collections
- Metrics

Creating test collections

- Cranfield @ TREC and other evaluation forums
- Cranfield @ work

Evaluation in a dynamic environment

For More Information

- E. Agichtein, E. Brill, S. Dumais, and R. Ragno. "Learning user interaction models for predicting Web search result preferences." *Proceedings of SIGIR 2006*. 2006.
- C. Buckley and E. M. Voorhees. "Evaluating evaluation measure stability." Proceedings of SIGIR 2000. pp. 33-40. 2000.
- C. Buckley and E. M. Voorhees. "Retrieval evaluation with incomplete information." Proceedings of SIGIR 2004. pp. 25-32. 2004.
- A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems, 27(1):1–27, 2008.
- E. M. Voorhees. "Variations in relevance judgments and the measurement of retrieval effectiveness." Proceedings of SIGIR '98. pp. 315- 323. 1998.
- E. M. Voorhees. "Evaluation by highly relevant documents." Proceedings of SIGIR 2001. pp. 74-82. 2001.
- E. M. Voorhees and C. Buckley. "The effect of topic set size on retrieval experiment error." Proceedings of SIGIR 2002. pp. 316-323. 2002.
- M. Sanderson and J. Zobel. "Information retrieval system evaluation: Effort, sensitivity, and reliability." Proceedings of SIGIR 2005. pp. 162-169.