

Information Retrieval Evaluation

Jing He

hejing@iro.umontreal.ca

October 21, 2012

- Excellent
- Very good
- Good
- Average
- Poor

Outline

- Background and Problem
- Evaluation Methods
 - User Study
 - Cranfield Paradigm (Test Collections)
 - Implicit Feedback
- Summary

Outline

- Background and Problem
- Evaluation Methods
 - User Study
 - Cranfield Paradigm
 - Implicit Feedback
- Summary

Commercial Search Engines

Google™

YAHOO!

msn

Ask™
.com
UK

AOL

overture
search performance

Netscape
software

altavista
THE SEARCH COMPANY

LYCOS

iwon.

dmoz

looksmart

TEOMA

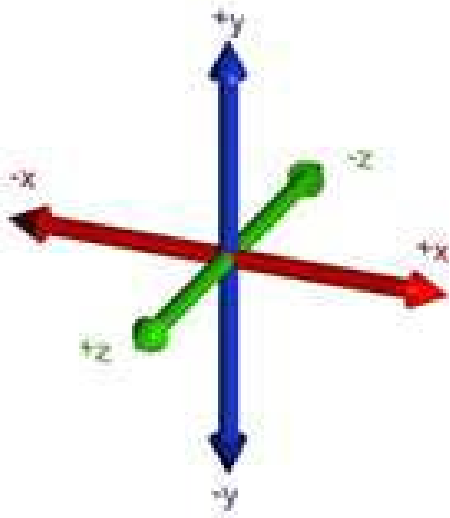


alltheweb***
all the web. all the time

inktomi

Windows Live

Information Retrieval Algorithms



vector space model

```
...  
text 0.2  
mining 0.1  
association 0.01  
clustering 0.02  
...  
food 0.00001  
...
```

language model

probabilistic model



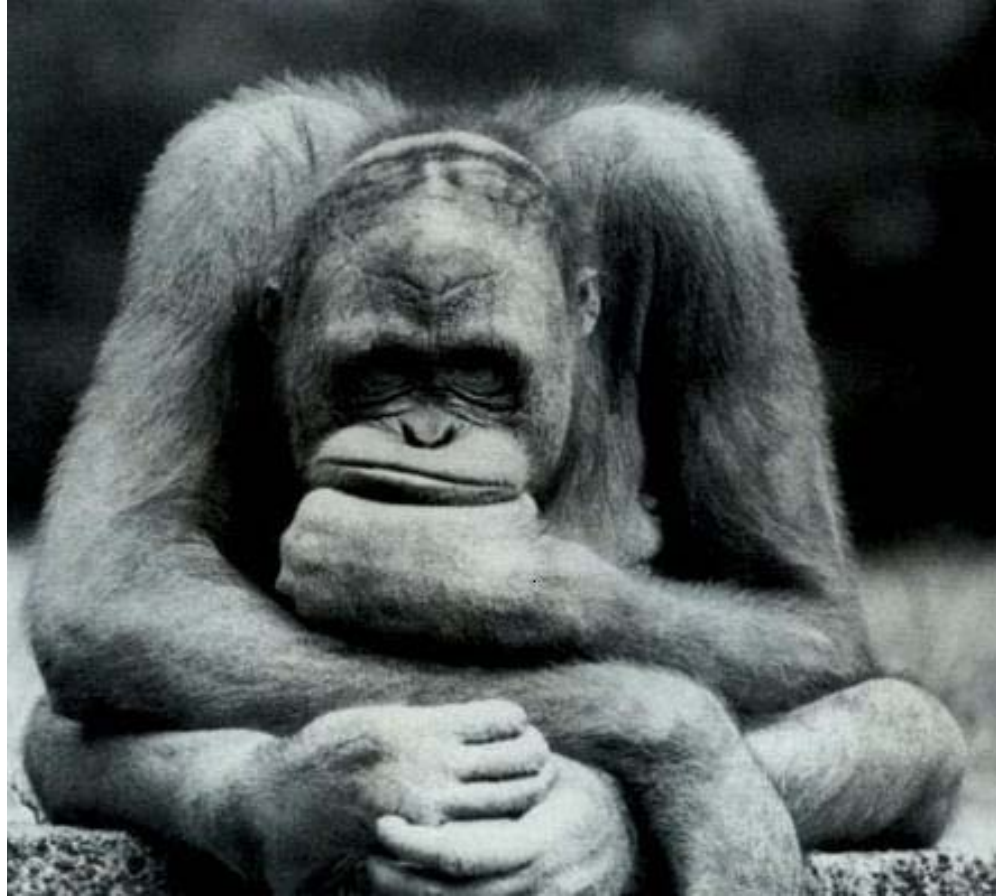
Problem

What is a better search engine (IR system) ?

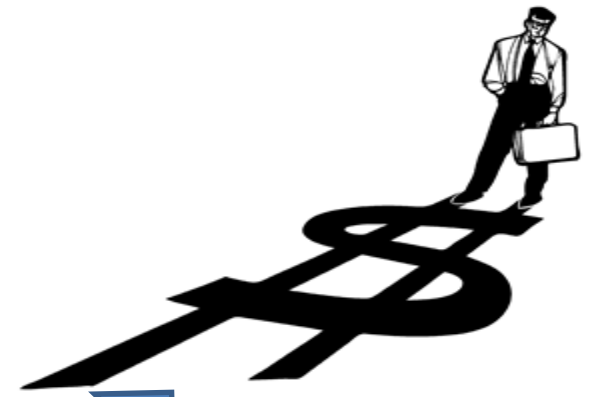


Wait.....Better?

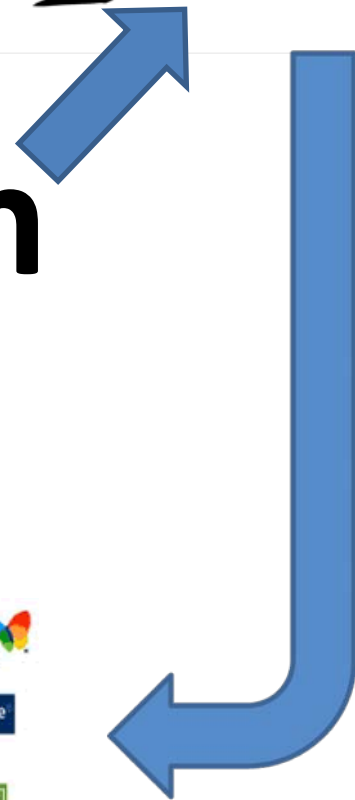
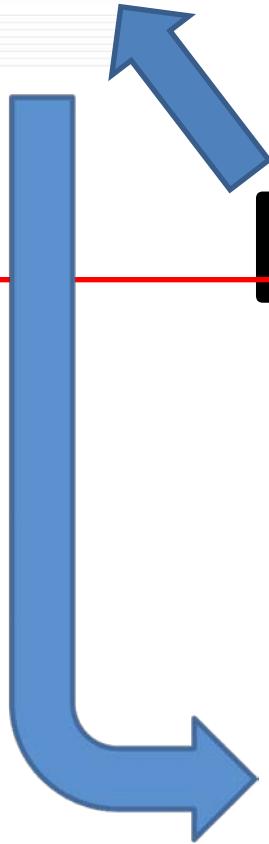
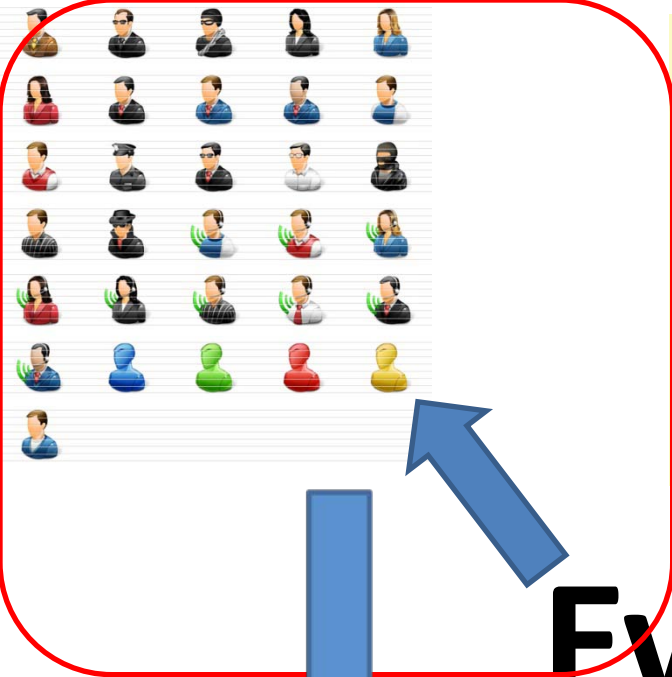
What do you mean?



Three different parties have different needs for a good system



Evaluation



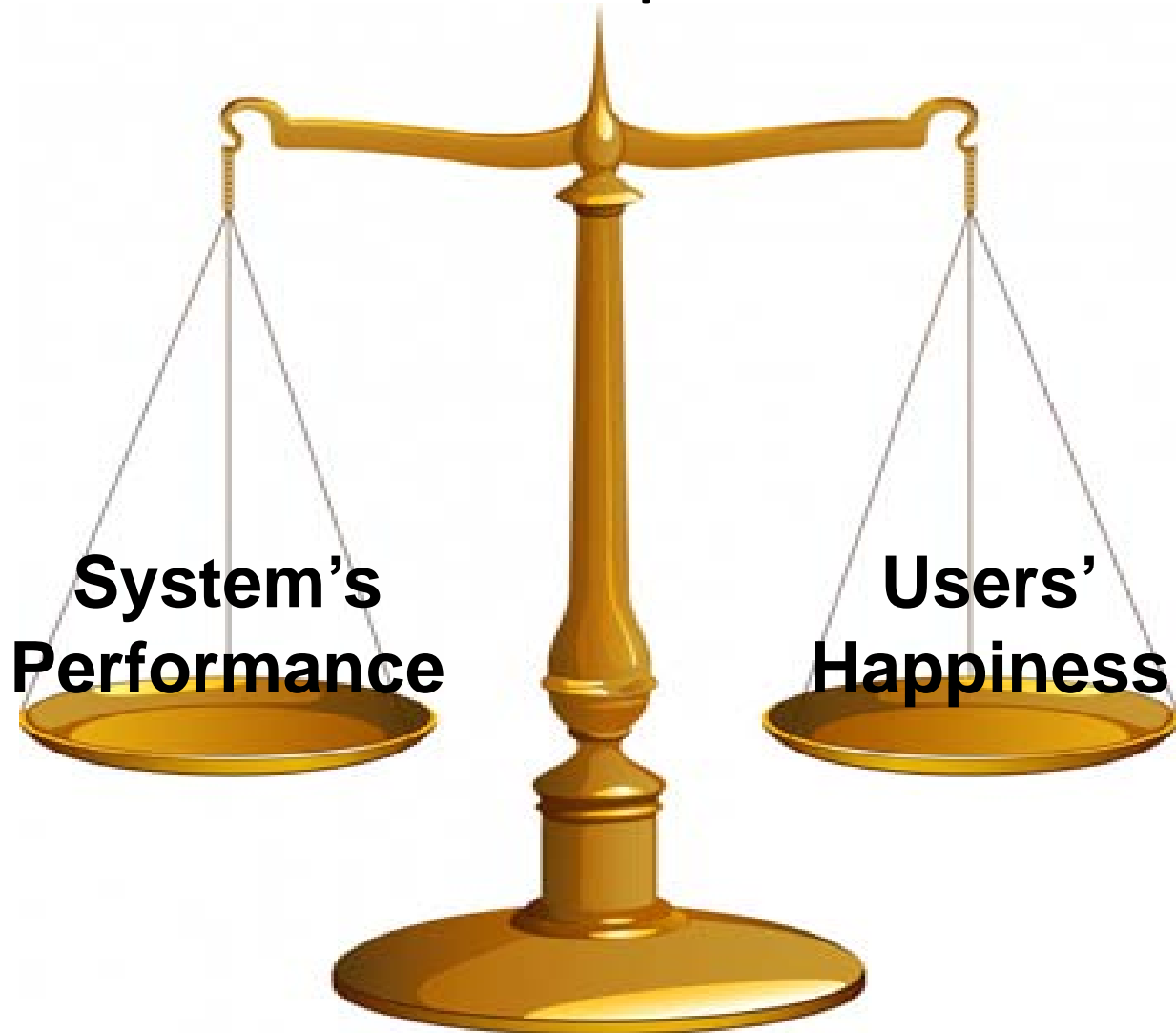
Outline

- Background and Problem
- **IR Evaluation**
 - User Study
 - Cranfield Paradigm
 - Implicit Feedback
- Summary

Outline

- Background and Problem
- IR Evaluation
 - User Study
 - Cranfield Paradigm
 - Implicit Feedback
- Summary

User Study Assumption



User Study

- Process
 - Actual users are hired
 - They use the systems to complete some tasks
 - They report their subjective feeling



User Study

- Strength
 - Close to real
- Weakness
 - Too subjective
 - Too expensive → Small Scale → Bias




User Study

- Strength
 - Close to real
- Weakness
 - Too subjective
 - Too expensive → Small Scale → Bias



User Study [Kagolovsky et al., 03]

- Process
 - Actual users are hired
 - They use the systems to finish a task
 - Their performance is measured
 - # of relevant documents found in a given time
 -  of finding required answers



IR Evaluation: User Study

- Strength
 - Close to real
- Weakness
 - Too expensive → Small Scale → Bias



Outline

- Background and Problem
- IR Evaluation
 - User Study
 - Cranfield Paradigm (Test Collection)
 - Implicit Feedback
- Summary

Satisfaction/Happiness: Divide and Conquer

- Efficiency
 - Response Time
 - Throughput
- Effectiveness
 - Quality of the returned list
- Interface
 - e.g. faceted search
 - Usually rely on the user study



Google

Web Show options...

Scholarly articles for [information retrieval](#)

[Information retrieval, data structures and algorithms](#) - Frakes - Cited 1
[Modern information retrieval](#) - Baeza-Yates - Cited by 6656
[Information storage and retrieval](#) - Korfhage - Cited by 612

[Information retrieval - Wikipedia, the free encyclopedia](#) - 2 visits -
Information retrieval (IR) is the science of searching for documents, for infor
documents and for metadata about documents, as well as that of ...
[History](#) - [Overview](#) - [Performance measures](#) - [Model types](#)
en.wikipedia.org/wiki/[Information_retrieval](#) - [Cached](#) - [Similar](#) -

[Introduction to Information Retrieval](#) - 3 visits - Jun 14
The book aims to provide a modern approach to [information retrieval](#) from a c
science perspective. It is based on a course we have been teaching in ...
[www-csli.stanford.edu/~hinrich/information-retrieval-book.html](#) - [Cached](#) - [Similar](#) -

[Information Retrieval Resources](#)
Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. S
UP, 2008. Classical and web [information retrieval](#) systems: ...
[www-csli.stanford.edu/~hinrich/information-retrieval.html](#) - [Cached](#) - [Si](#)

[Information Retrieval](#) - 7 visits - Aug 26
[Information Retrieval](#) - The Journal of [Information Retrieval](#) is an internation
... ..



Efficiency



- Same as any database/Architecture/Software
- benchmark/test collection
 - Document collection
 - Query set
- Because the test collection is **reusable**, so
 - Cheap
 - Easy for Error Analysis

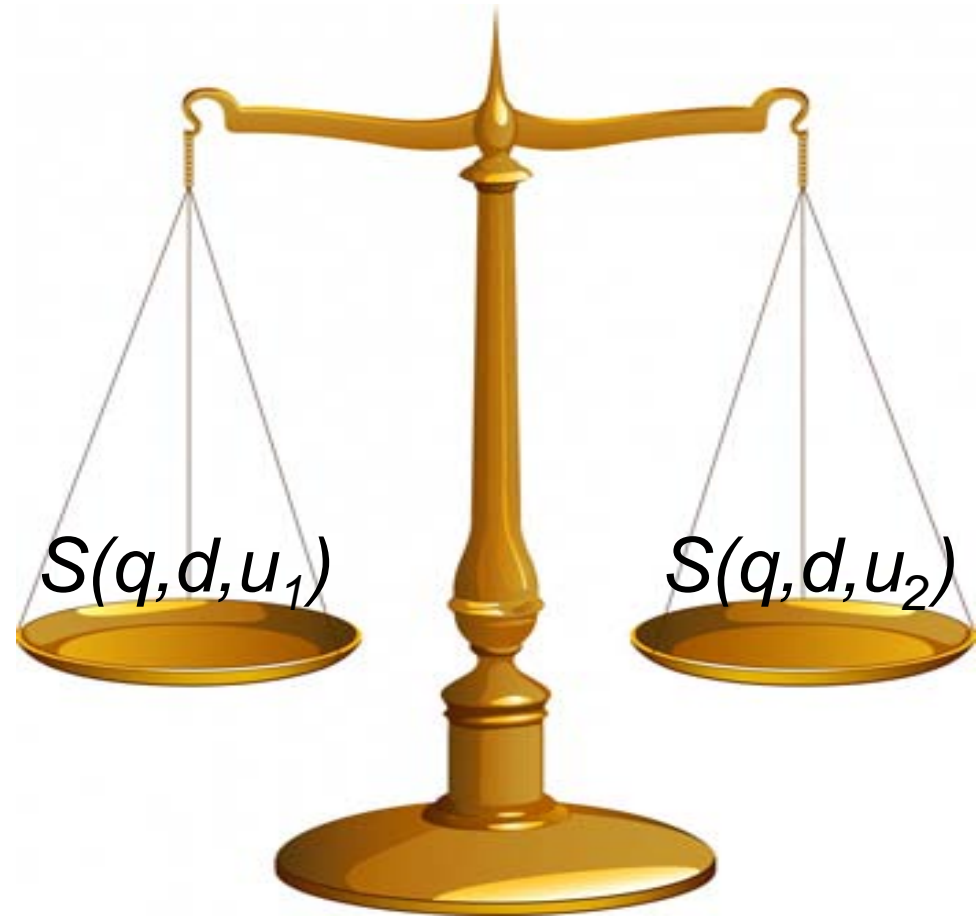
Effectiveness

- A reusable test collection for effectiveness ?

Effectiveness Evaluation

Assumption

- Information need q
- Document d
- User u
- Satisfaction $S(q,d,u)$



Cranfield Paradigm

- A test collection
 - Document collection D
 - Topic set T
 - Relevance Judgments R
- A retrieval system runs
 - Retrieve lists L from D for topic T
- A measure is used to score the system
 - score = $f(R, L)$



Cranfield Paradigm: Process

- Given
 - a) A test collection (T, D, R)
 - b) A retrieval run for the test collection : a doc-list L_t for each topic t in T
- For each topic t in T
 - Use a measure (e.g. $P@10$) to compute the quality of L_t
- Combine scores
 - e.g., arithmetic average



Test Collection/Benchmark



Document Collection



$R(d,q)=\text{True?}$



Relevance Judgments



Query Set

Assumption

$$R(d, q, u1) == R(d, q, u2)$$

Organizations for Standard Test Collections

- Cranfield
 - Cranfield College, UK, 1950s
- TREC (Text REtrieval Conference)
 - by U.S. National Institute of Standards and Technology
 - 1992-now
- NTCIR (NII Test Collection for IR Systems)
 - East Asian languages
- CLEF (Cross Language Evaluation Forum)
 - European languages

Cranfield Paradigm: Process

- Given
 - a) A test collection (T, D, R)
 - b) A retrieval run for the test collection : a doc-list L_t for each topic t in T
- For each topic t in T
 - Use a **measure** (e.g. P@10) to compute the quality of L_t
- Combine scores
 - e.g., arithmetic average



Measures

- Binary Judgment Measures

- Unranked Results

$$J: Q \times D \rightarrow \{0,1\}$$

- Ranked Results Measures

- Graded Judgment Measures

$$J: Q \times D \rightarrow \{0,1,2,3\}$$

Measures

- Binary Judgment Measures
 - Unranked Results: a document set
 - Ranked Results: a document list
- Graded Measures

Measures

- Binary Judgment Measures
 - Unranked Results: a document set
 - Precision
 - Recall
 - F-score
 - Ranked Results: a document list
- Graded Measures

Measures: 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

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

Measures: Precision and Recall

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

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

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

- Trade-off between precision and recall
 - Return more docs → higher recall, (usually) lower precision

Measures: Combining Precision and Recall

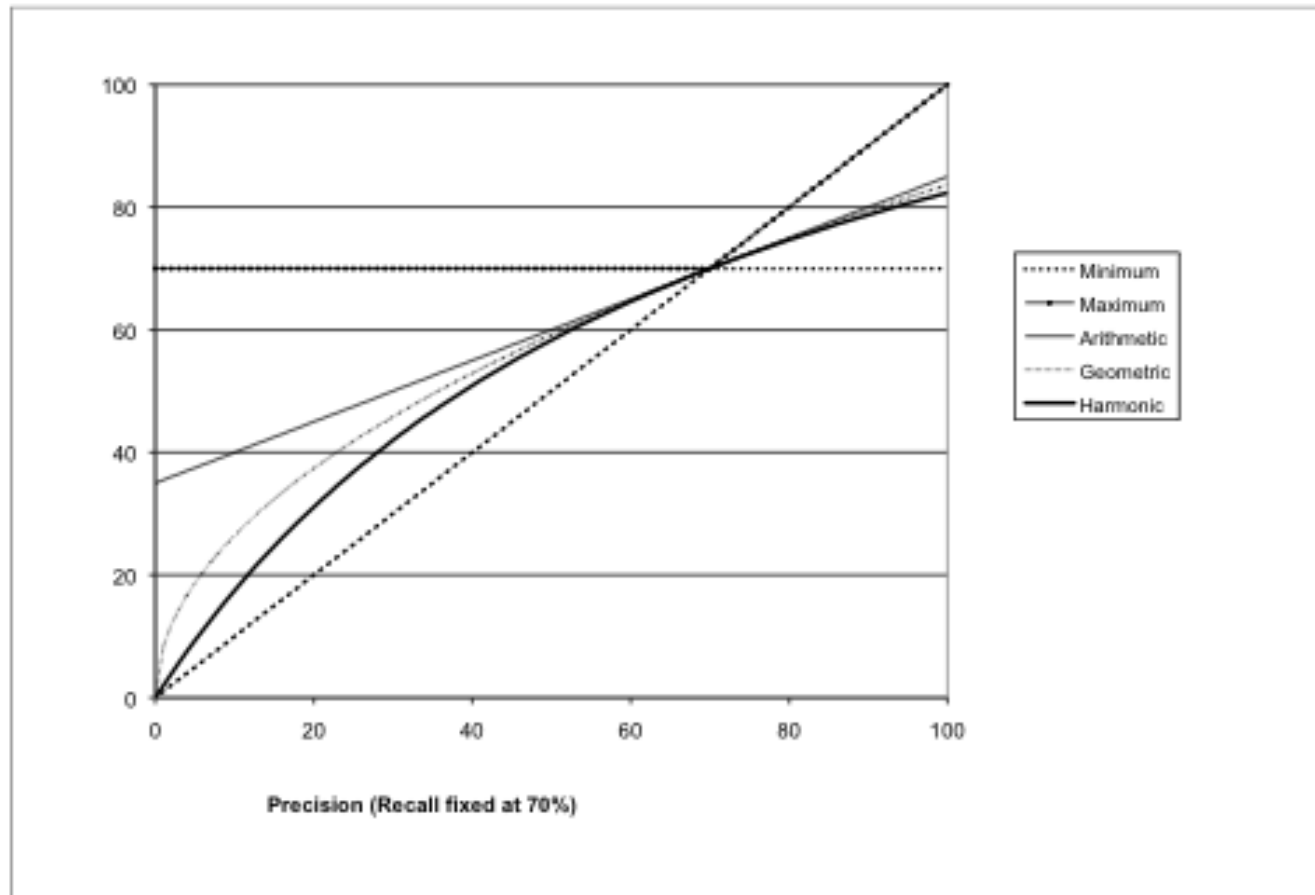
- Combine precision and recall in F-score

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

- $\alpha \in [0, 1]$ is used to control the relative importance of precision/recall
 - Precision is more important for Web search
 - Recall is more important for patent search
- When $\alpha=0.5$, it is the harmonic mean

Why harmonic average?

- A kind of soft-minimum



Measures: a Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- $P = 20 / (20 + 40) = 1/3$

- $R = 20 / (20 + 60) = 1/4$

- $F_1 = 2 \frac{1}{\frac{1}{3} + \frac{1}{4}} = 2/7$

Measures: a Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- Why not using accuracy?

Measures

- Binary Judgment Measures
 - Unranked Results: a document set
 - Ranked Results: a document list
 - $P@n$, $R@n$, precision-recall curve, MRR, MAP
- Graded Measures

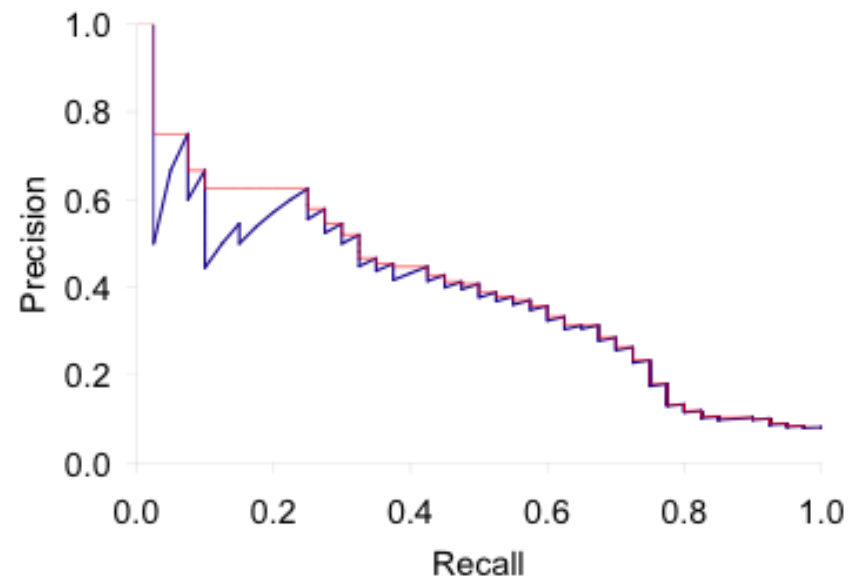
Measures: $P@n$ and $R@n$

- For each cutoff n , take top n docs as a set
- Drawback
 - Only contains incomplete information of a list
 - Insensitive to the rank of relevant docs
 - e.g. $P@5$ values are identical for the following lists
 - 1,1,0,0,0
 - 0,0,0,1,1
- P-R curve
 - Contains complete information

Measures: P-R curve

- For each cut off n , get a $(R@n, P@n)$ pair
- Take $R@n$ as x-axis, and $P@n$ as y-axis, we get the P-R curve


Interpolation (in red): Take maximum of all future points

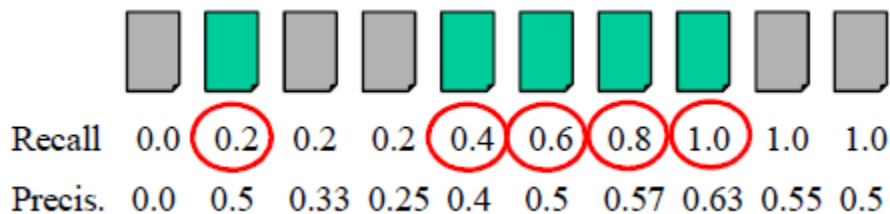


- P-R curve is usually only plotting for Recall (0.0, 0.1, ..., 0.9, 1.0) – for easy combination

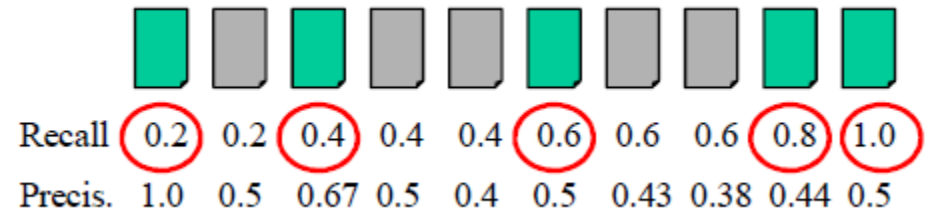
Measures: Average Precision

- Not easy to compare systems by P-R curves
- Approximate area under the P-R: Average Precision
 - Average the precision at the positions of relevant docs

 = the relevant documents



AvgPrec= 62.2%



AvgPrec= 52.0%

Measures: MRR

- Mean Reciprocal Rank
 - Reciprocal of rank of the first relevant doc
- Used for some kinds of queries
 - Navigational Queries
 - “glassdoor”
 - Specific Informational Queries
 - “when was the first Olympic Game?”

Measures

- Binary Judgment Measures
 - Unranked Results: a document set
 - Ranked Results: a document list
- Graded Judgment Measures
 - nDCG

Measures: nDCG

- Graded Judgment
 - Relevant documents can provide different amount of useful information
 - Highly relevant doc vs. Marginal relevant doc
- Gain from a doc (G)
 - Determined by its relevance degree

Measures: nDCG

- Cumulated Gain (CG)
 - Sum of gain from docs in the list
- Discounted Cumulated Gain (DCG)
 - Top ranked docs are more important for users
 - Top ranked docs should be weighted highly

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

Measures: nDCG

- Gain

3, 2, 3, 0, 0, 1, 2, 2, 3, 0

- Discounted Gain

3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

- Discounted Cumulated Gain

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

Measures: nDCG

- Normalized Discounted Cumulated Gain (nDCG)
 - Why normalizing?
 - Value ranges for queries are quite different
 - e.g.
 - q1 has only 1 relevant doc in D
 - q2 has 1000 relevant docs in D
 - The average score of DCG will be dominated by q2
- Normalized Factor
 - DCG value for an ideal (best) doc list

Measures: nDCG

- G and DCG (assume it contains all rel docs)

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

- Ideal G and DCG

3, 3, 3, 2, 2, 2, 1, 0, 0, 0

3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

- nDCG

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

Measures

- Binary Judgment Measures
 - Unranked Results: a document set
 - Precision, Recall, F
 - Ranked Results: a document list
 - $P@n$, $R@n$, P-R curve, Average Precision, MRR
- Graded Judgment Measures
 - nDCG

Cranfield Paradigm: Process

- Given
 - a) A test collection (T, D, R)
 - b) A retrieval run for the test collection : a doc-list L_t for each topic t in T
- For each topic t in T
 - Use a measure (e.g. $P@10$) to compute the quality of L_t
- **Combine scores**
 - e.g., arithmetic average



Combine Scores and Compare

- Two systems (A and B), which is better?
- Compare the arithmetic average score?
 - Difference between scores
 - Sample size
- Principle Comparison: Significant Test
 - For comparison: One-sided test
 - Widely used: t-test, Wilcoxon signed-rank test

Cranfield Paradigm

- Strength



- Cheap

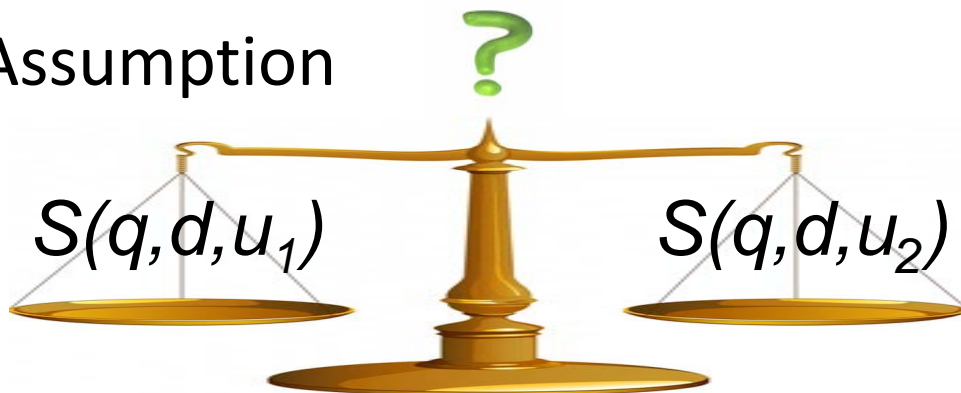
- Easy for Error Analysis

- Large Sample for More Confidence

- Repeatable

Cranfield Paradigm: Weakness

- test collection
 - Document collection D
 - Topic set T
 - Relevance Judgments R
- Weakness
 - Relevance Judgments are expensive \rightarrow incomplete
 - Assumption



Problem of Relevance Judgments

- Collect Relevance Judgments from Real User?

Outline

- Background and Problem
- IR Evaluation
 - User Study
 - Cranfield Paradigm
 - **Implicit Feedback**
- Summary

Implicit Feedback

- User Behavior → Relevance Judgments



Implicit Feedback

- Strength
 - Real User
 - Cheaper than cranfield paradigm
 - Much Larger sample size
- Challenge
 - User behavior noise
 - Long-tail search



Implicit Feedback

- A/B test
 - Use a small proportion of traffic (1%) for evaluation
 - Option 1: Show results from different retrieval methods alternatively
 - Option 2: Merge results in a doc list
 - Compare the clickthrough-rate of two results

Outline

- Background and Problem
- IR Evaluation
 - User Study
 - Cranfield Paradigm
 - Implicit Feedback
- Summary

Summary

- Real users are ground-truth
- Evaluation of methods can be decomposed
- Reusable test collection is useful
- User behavior (log) is really a kind of wealth



