

LBSC 796/INFM 718R: Week 2

## Evaluation



**Jimmy Lin**  
College of Information Studies  
University of Maryland

Monday, February 6, 2006

## IR is an experimental science!

- Formulate a research question: the hypothesis
- Design an experiment to answer the question
- Perform the experiment
  - Compare with a baseline "control"
- Does the experiment answer the question?
  - Are the results significant? Or is it just luck?
- Report the results!
- Rinse, repeat...

## Questions About the Black Box

- Example "questions":
  - Does morphological analysis improve retrieval performance?
  - Does expanding the query with synonyms improve retrieval performance?
- Corresponding experiments:
  - Build a "stemmed" index and compare against "unstemmed" baseline
  - Expand queries with synonyms and compare against baseline unexpanded queries

## Questions That Involve Users

- Example "questions":
  - Does keyword highlighting help users evaluate document relevance?
  - Is letting users weight search terms a good idea?
- Corresponding experiments:
  - Build two different interfaces, one with keyword highlighting, one without; run a user study
  - Build two different interfaces, one with term weighting functionality, and one without; run a user study

## The Importance of Evaluation

- The ability to measure differences underlies experimental science
  - How well do our systems work?
  - Is A better than B?
  - Is it really?
  - Under what conditions?
- Evaluation drives what to research
  - Identify techniques that work and don't work
  - Formative vs. summative evaluations

## Desiderata for Evaluations

- Insightful
- Affordable
- Repeatable
- Explainable

## Summary

- Qualitative user studies suggest what to build
- Decomposition breaks larger tasks into smaller components
- Automated evaluation helps to refine components
- Quantitative user studies show how well everything works together

## Outline

- Evaluating the IR black box
  - How do we conduct experiments with reusable test collections?
  - What exactly do we measure?
  - Where do these test collections come from?
- Studying the user and the system
  - What sorts of (different) things do we measure when a human is in the loop?
- Coming up with the right questions
  - How do we know what to evaluate and study?

## Types of Evaluation Strategies

- System-centered studies
  - Given documents, queries, and relevance judgments
  - Try several variations of the system
  - Measure which system returns the "best" hit list
- User-centered studies
  - Given several users, and at least two retrieval systems
  - Have each user try the same task on both systems
  - Measure which system works the "best"

## Evaluation Criteria

- Effectiveness
  - How "good" are the documents that are returned?
  - System only, human + system
- Efficiency
  - Retrieval time, indexing time, index size
- Usability
  - Learnability, frustration
  - Novice vs. expert users

## Good Effectiveness Measures

- Should capture some aspect of what the user wants
  - That is, the measure should be meaningful
- Should have predictive value for other situations
  - What happens with different queries on a different document collection?
- Should be easily replicated by other researchers
- Should be easily comparable
  - Optimally, expressed as a single number

## The Notion of Relevance

- IR systems essentially facilitate communication between a user and document collections
- **Relevance** is a measure of the effectiveness of communication
  - Logic and philosophy present other approaches
- Relevance is a relation... but between what?

## What is relevance?

|                           |                        |                            |
|---------------------------|------------------------|----------------------------|
| <b>Relevance is the</b>   | measure of a           | correspondence             |
|                           | degree                 | utility                    |
|                           | dimension              | connection                 |
|                           | estimate               | satisfaction               |
|                           | appraisal              | fit                        |
|                           | relation               | bearing                    |
|                           |                        | matching                   |
| <b>existing between a</b> | document               | <b>and a</b> query         |
|                           | article                | request                    |
|                           | textual form           | information used           |
|                           | reference              | point of view              |
|                           | information provided   | information need statement |
|                           | fact                   |                            |
| <b>as determined by</b>   | person                 |                            |
|                           | judge                  |                            |
|                           | user                   |                            |
|                           | requester              |                            |
|                           | Information specialist |                            |

**Does this help?**

Tefko Saracevic, (1975) Relevance: A Review of and a Framework for Thinking on the Notion in Information Science. Journal of the American Society for Information Science, 26(6), 321-343.

## Mizzaro's Model of Relevance

- Four dimensions of relevance
- Dimension 1: Information Resources
  - Information
  - Document
  - Surrogate
- Dimension 2: Representation of User Problem
  - Real information needs (RIN) = visceral need
  - Perceived information needs (PIN) = conscious need
  - Request = formalized need
  - Query = compromised need

Stefano Mizzaro, (1999) How Many Relevance in Information Retrieval? Interacting With Computers, 10(3), 305-322.

## Time and Relevance

- Dimension 3: Time

## Components and Relevance

- Dimension 4: Components
  - Topic
  - Task
  - Context

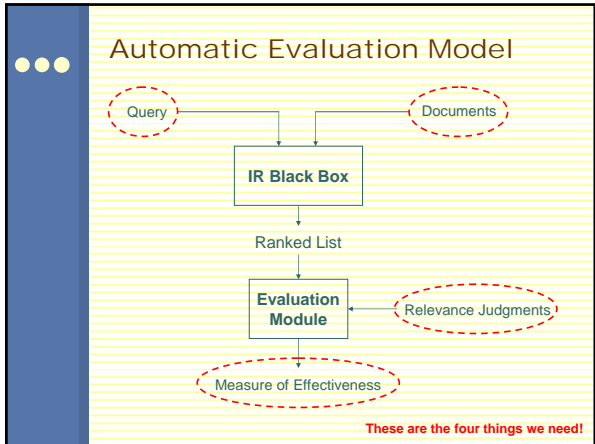
## What are we after?

- Ultimately, relevance of the information
  - With respect to the real information need
  - At the conclusion of the information seeking process
  - Taking into consideration topic, task, and context
$$Rel(\text{Information, RIN, } t(t), \{\text{Topic, Task, Context}\})$$
- In system-based evaluations, what do we settle for?
  - $Rel(\text{surrogate, request, } t(t), \text{Topic})$
  - $Rel(\text{document, request, } t(t), \text{Topic})$

## Evaluating the Black Box

## Evolution of the Evaluation

- Evaluation by **inspection** of examples
- Evaluation by **demonstration**
- Evaluation by **improvised** demonstration
- Evaluation on **data** using a figure of merit
- Evaluation on **test data**
- Evaluation on **common** test data
- Evaluation on common, **unseen** test data



## Test Collections

- Reusable test collections consist of:
  - Collection of documents
    - Should be "representative"
    - Things to consider: size, sources, genre, topics, ...
  - Sample of information needs
    - Should be "randomized" and "representative"
    - Usually formalized topic statements
  - Known relevance judgments
    - Assessed by humans, for each topic-document pair (topic, not query!)
    - Binary judgments make evaluation easier
- Measure of effectiveness
  - Usually a numeric score for quantifying "performance"
  - Used to compare different systems

## Which is the Best Rank Order?

■ = relevant document

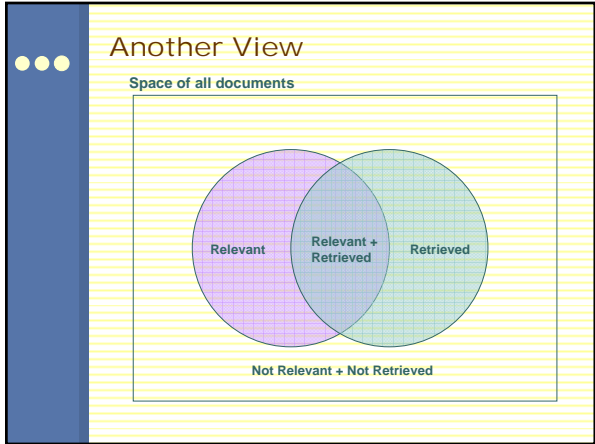
## Set-Based Measures

|               |          |              |
|---------------|----------|--------------|
|               | Relevant | Not relevant |
| Retrieved     | A        | B            |
| Not retrieved | C        | D            |

Collection size = A+B+C+D  
Relevant = A+C  
Retrieved = A+B

- **Precision** =  $A \div (A+B)$
- **Recall** =  $A \div (A+C)$
- **Miss** =  $C \div (A+C)$
- **False alarm (fallout)** =  $B \div (B+D)$

When is precision important?  
When is recall important?



## F-Measure

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

- Harmonic mean of recall and precision
- Beta controls relative importance of precision and recall
  - Beta = 1, precision and recall equally important
  - Beta = 5, recall five times more important than precision

What if no relevant documents exist?

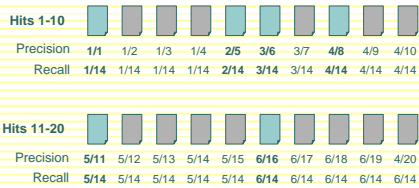
## Single-Valued Measures

- Precision at a fixed number of documents
  - Precision at 10 docs is often useful for Web search
- R-precision
  - Precision at  $r$  documents, where  $r$  is the total number of relevant documents
- Expected search length
  - Average rank of the first relevant document

How do we take into account the relationship between precision and recall?

## Measuring Precision and Recall

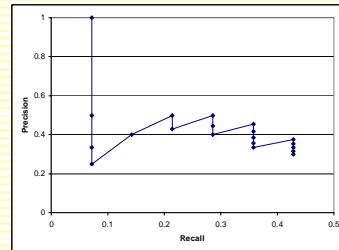
Assume there are a total of 14 relevant documents



 = relevant document

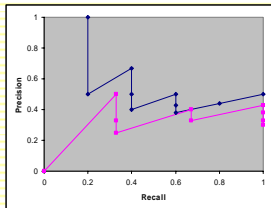
## Graphing Precision and Recall

- Plot each (recall, precision) point on a graph
- Visually represent the precision/recall tradeoff



## Need for Interpolation

- Two issues:
  - How do you compare performance across queries?
  - Is the sawtooth shape intuitive of what's going on?



Solution: Interpolation!

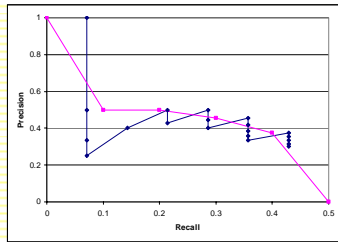
## Interpolation

- Why?
  - We have no observed data between the data points
  - Strange sawtooth shape doesn't make sense
- It is an empirical fact that *on average* as recall increases, precision decreases
- Interpolate at 11 standard recall levels
  - 100%, 90%, 80%, ... 30%, 20%, 10%, 0% (!)
- How?

$$P(R) = \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

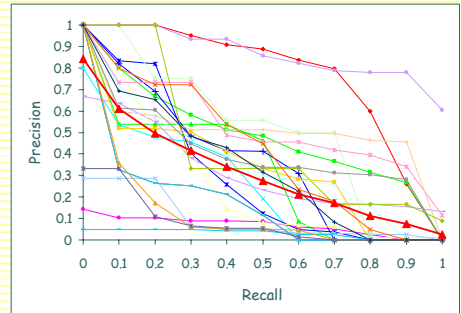
where  $S$  is the set of all observed  $(P, R)$  points

## Result of Interpolation



We can also average precision across the 11 standard recall levels

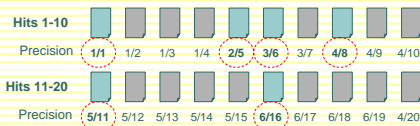
## How do we compare curves?



Adapted from a presentation by Ellen Voorhees at the University of Maryland, March 29, 1999

## The MOAM

- Mean average precision (MAP)
  - Average of precision at each retrieved relevant document
  - Relevant documents not retrieved contribute zero to score



Assume total of 14 relevant documents: 8 relevant documents not retrieved contribute eight zeros

MAP = .2307

= relevant document

## Building Test Collections

- Where do test collections come from?
  - Someone goes out and builds them (expensive)
  - As the byproduct of large scale evaluations
- TREC = Text REtrieval Conferences
  - Sponsored by NIST
  - Series of annual evaluations, started in 1992
  - Organized into "tracks"
  - Larger tracks may draw a few dozen participants

See proceedings online at <http://trec.nist.gov/>

## Ad Hoc Topics

- In TREC, a statement of information need is called a *topic*

**Title:** Health and Computer Terminals

**Description:** Is it hazardous to the health of individuals to work with computer terminals on a daily basis?

**Narrative:** Relevant documents would contain any information that expands on any physical disorder/problems that may be associated with the daily working with computer terminals. Such things as carpal tunnel, cataracts, and fatigue have been said to be associated, but how widespread are these or other problems and what is being done to alleviate any health problems.

## Obtaining Judgments

- Exhaustive assessment is usually impractical
  - TREC has 50 queries
  - Collection has >1 million documents
- Random sampling won't work
  - If relevant docs are rare, none may be found!
- IR systems can help focus the sample
  - Each system finds some relevant documents
  - Different systems find different relevant documents
  - Together, enough systems will find most of them
  - Leverages cooperative evaluations

## Pooling Methodology

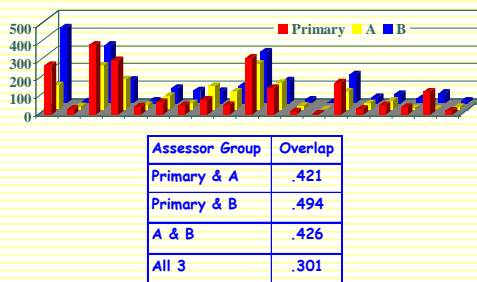
- Systems submit top 1000 documents per topic
- Top 100 documents from each are judged
  - Single pool, duplicates removed, arbitrary order
  - Judged by the person who developed the topic
- Treat unevaluated documents as not relevant
- Compute MAP down to 1000 documents
- To make pooling work:
  - Systems must do reasonable well
  - Systems must not all "do the same thing"
- Gather topics and relevance judgments to create a reusable test collection

## Does pooling work?

- But judgments can't possibly be exhaustive!
  - It doesn't matter: relative rankings remain the same!  
Chris Buckley and Ellen M. Voorhees. (2004) Retrieval Evaluation with Incomplete Information. SIGIR 2004.
- But this is only one person's opinion about relevance
  - It doesn't matter: relative rankings remain the same!  
Ellen Voorhees. (1998) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. SIGIR 1998.
- But what about hits 101 to 1000?
  - It doesn't matter: relative rankings remain the same!
- But we can't possibly use judgments to evaluate a system that didn't participate in the evaluation!
  - Actually, we can!  
Justin Zobel. (1998) How Reliable Are the Results of Large-Scale Information Retrieval Experiments? SIGIR 1998.

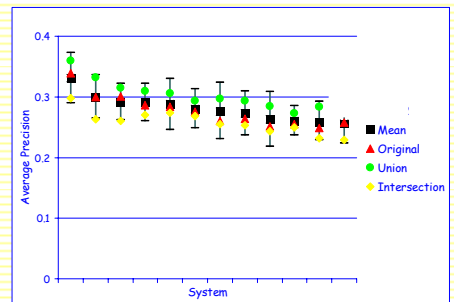
## Agreement on Relevance

# Relevant per Topic by Assessor



Adapted from a presentation by Ellen Voorhees at the University of Maryland, March 29, 1999

## Effect of Different Judgments



Adapted from a presentation by Ellen Voorhees at the University of Maryland, March 29, 1999

## Lessons From TREC

- Absolute scores are not trustworthy
  - Who's doing the relevance judgment?
  - How complete are the judgments?
- Relative rankings are stable
  - Comparative conclusions are most valuable
- Cooperative evaluations produce reliable test collections
- Evaluation technology is predictive

## Alternative Methods

- Search-guided relevance assessment
  - Iterate between topic research/search/assessment
- Known-item judgments have the lowest cost
  - Tailor queries to retrieve a single known document
  - Useful as a first cut to see if a new technique is viable

## Why significance tests?

- System A and B identical on all but one query:
  - Is it just a lucky query for System A?
  - Need A to beat B frequently to believe it is really better
  - Need as many queries as possible

Empirical research suggests 25 is minimum needed  
TREC tracks generally aim for at least 50 queries
- System A beats system B on every query:
  - But only does so by 0.00001%
  - Does that mean much?
- Significance tests consider those issues
  - What's a p-value?

## Averages Can Deceive

| Experiment 1 |          |          | Experiment 2 |          |          |
|--------------|----------|----------|--------------|----------|----------|
| Query        | System A | System B | Query        | System A | System B |
| 1            | 0.20     | 0.40     | 1            | 0.02     | 0.76     |
| 2            | 0.21     | 0.41     | 2            | 0.39     | 0.07     |
| 3            | 0.22     | 0.42     | 3            | 0.16     | 0.37     |
| 4            | 0.19     | 0.39     | 4            | 0.58     | 0.21     |
| 5            | 0.17     | 0.37     | 5            | 0.04     | 0.02     |
| 6            | 0.20     | 0.40     | 6            | 0.09     | 0.91     |
| 7            | 0.21     | 0.41     | 7            | 0.12     | 0.46     |
| Average      | 0.20     | 0.40     | Average      | 0.20     | 0.40     |

## How Much is Enough?

- Measuring improvement
  - Achieve a meaningful improvement
    - Guideline: 0.05 is noticeable, 0.1 makes a difference (in MAP)
  - Achieve reliable improvement on "typical" queries
    - Wilcoxon signed rank test for paired samples
- Know when to stop!
  - Inter-assessor agreement places a limit on human performance

## The Evaluation Commandments

1. Thou shalt define insightful evaluation metrics
2. Thou shalt define replicable evaluation metrics
3. Thou shalt report all relevant system parameters
4. Thou shalt establish upper bounds on performance
5. Thou shalt establish lower bounds on performance

## The Evaluation Commandments

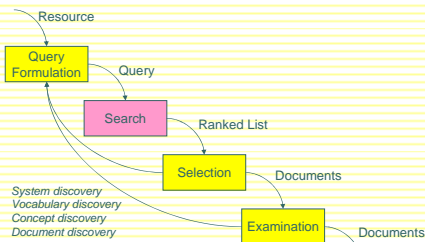
6. Thou shalt test differences for statistical significance
7. Thou shalt say whether differences are meaningful
8. Thou shalt not mingle training data with test data
9. Thou shalt not mingle training data with test data
10. Thou shalt not mingle training data with test data

## Recap: Automatic Evaluation

- Test collections abstract the evaluation problem
  - Places focus on the IR black box
- Automatic evaluation is one shot
  - Ignores the richness of human interaction
- Evaluation measures focus on *one* notion of performance
  - But users care about other things
- Goal is to compare systems
  - Values may vary, but relative differences are stable
- Mean values obscure important phenomena
  - Augment with failure analysis and significance tests



## Putting the User in the Loop



## User Studies

- Goal is to account for interaction
  - By studying the interface component
  - By studying the complete system
- Two different types of evaluation
  - Formative: provides a basis for system development
  - Summative: designed to assess performance

## Quantitative User Studies

- Select independent variable(s)
  - e.g., what info to display in selection interface
- Select dependent variable(s)
  - e.g., time to find a known relevant document
- Run subjects in different orders
  - Average out learning and fatigue effects
- Compute statistical significance

## Additional Effects to Consider

- Learning
  - Vary topic presentation order
- Fatigue
  - Vary system presentation order
- Expertise
  - Ask about prior knowledge of each topic

## Blair and Maron (1985)

- A classic study of retrieval effectiveness
  - Earlier studies were on unrealistically small collections
- Studied an archive of documents for a law suit
  - 40,000 documents, ~350,000 pages of text
  - 40 different queries
  - Used IBM's STAIRS full-text system
- Approach:
  - Lawyers stipulated that they must be able to retrieve at least 75% of all relevant documents
  - Search facilitated by paralegals
  - Precision and recall evaluated only after the lawyers were satisfied with the results

David C. Blair and M. E. Maron. (1984) An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM*, 28(3), 289-299.

## Blair and Maron's Results

- Average precision: 79%
- Average recall: 20% (!!)
- Why recall was low?
  - Users can't anticipate terms used in relevant documents
    - "accident" might be referred to as "event", "incident", "situation", "problem," ...
  - Differing technical terminology
  - Slang, misspellings
- Other findings:
  - Searches by both lawyers had similar performance
  - Lawyer's recall was not much different from paralegal's

## Batch vs. User Evaluations

- Do batch (black box) and user evaluations give the same results? If not, why?
- Two different tasks:
  - Instance recall (6 topics)

What countries import Cuban sugar?  
What tropical storms, hurricanes, and typhoons have caused property damage or loss of life?
  - Question answering (8 topics)

Which painting did Edvard Munch complete first, "Vampire" or "Puberty"?  
Is Denmark larger or smaller in population than Norway?

Andrew Turpin and William Hersh, (2001) Why Batch and User Evaluations Do No Give the Same Results. *Proceedings of SIGIR 2001*.

## The Study

- Compared of two systems:
  - a baseline system
  - an improved system that was provably better in batch evaluations
- Results:

|                         | Instance Recall |             | Question Answering |               |
|-------------------------|-----------------|-------------|--------------------|---------------|
|                         | Batch MAP       | User recall | Batch MAP          | User accuracy |
| Baseline                | 0.2753          | 0.3230      | 0.2696             | 66%           |
| Improved                | 0.3239          | 0.3728      | 0.3544             | 60%           |
| Change                  | +18%            | +15%        | +32%               | -6%           |
| p-value (paired t-test) | 0.24            | 0.27        | 0.06               | 0.41          |

## Analysis

- A "better" IR black box doesn't necessary lead to "better" end-to-end performance!
- Why?
- Are we measuring the right things?

## Qualitative User Studies

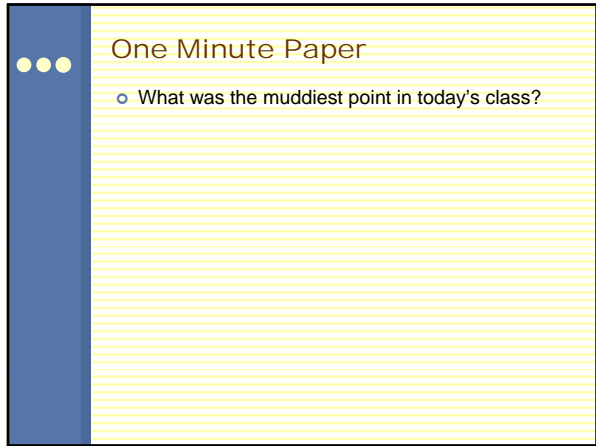
- How do we discover the right research questions to ask?
- Observe user behavior
  - Instrumented software, eye trackers, etc.
  - Face and keyboard cameras
  - Think-aloud protocols
  - Interviews and focus groups
- Organize the data
- Look for patterns and themes
- Develop a "grounded theory"

## Questionnaires

- Demographic data
  - Age, education, computer experience, etc.
  - Basis for interpreting results
- Subjective self-assessment
  - Which system did they think was more effective?
  - Often at variance with objective results!
- Preference
  - Which system did they prefer? Why?

## Summary

- Qualitative user studies suggest what to build
- Decomposition breaks larger tasks into smaller components
- Automated evaluation helps to refine components
- Quantitative user studies show how well everything works together



One Minute Paper

- o What was the muddiest point in today's class?