

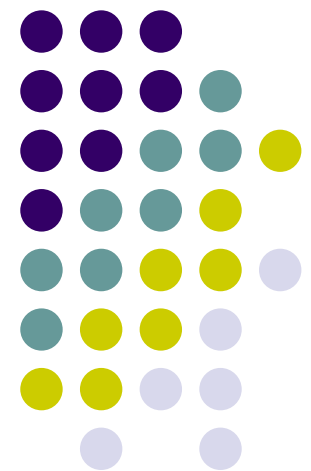
Measuring Search Effectiveness: Lessons from Interactive TREC

Gheorghe Muresan

School of Communication, Information and Library Studies

Rutgers University

<http://www.scils.rutgers.edu/~muresan/>

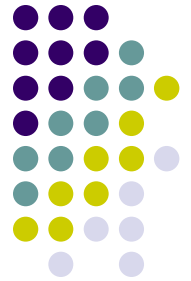




Objectives

- Discuss methodologies and measures of effectiveness that, in our experience, mainly in the TREC Interactive track, have proven successful in painting an accurate picture of the user interaction when seeking information.
- Classify the measures and discuss the contexts when they can be used.
- Attempt to provide guidelines as to which measures are appropriate in certain conditions.

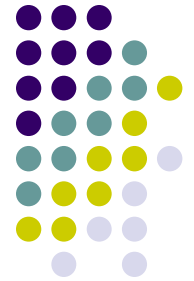
Before doing IR evaluation, ask: “What do we want from an IRS ?”



- Systemic approach
 - Goal (for a known information need):
Return as many relevant documents as possible and as few non-relevant documents as possible
- Cognitive approach
 - Goal (in an interactive information-seeking environment, with a given IRS):
Support the user’s exploration of the problem domain and the task completion.

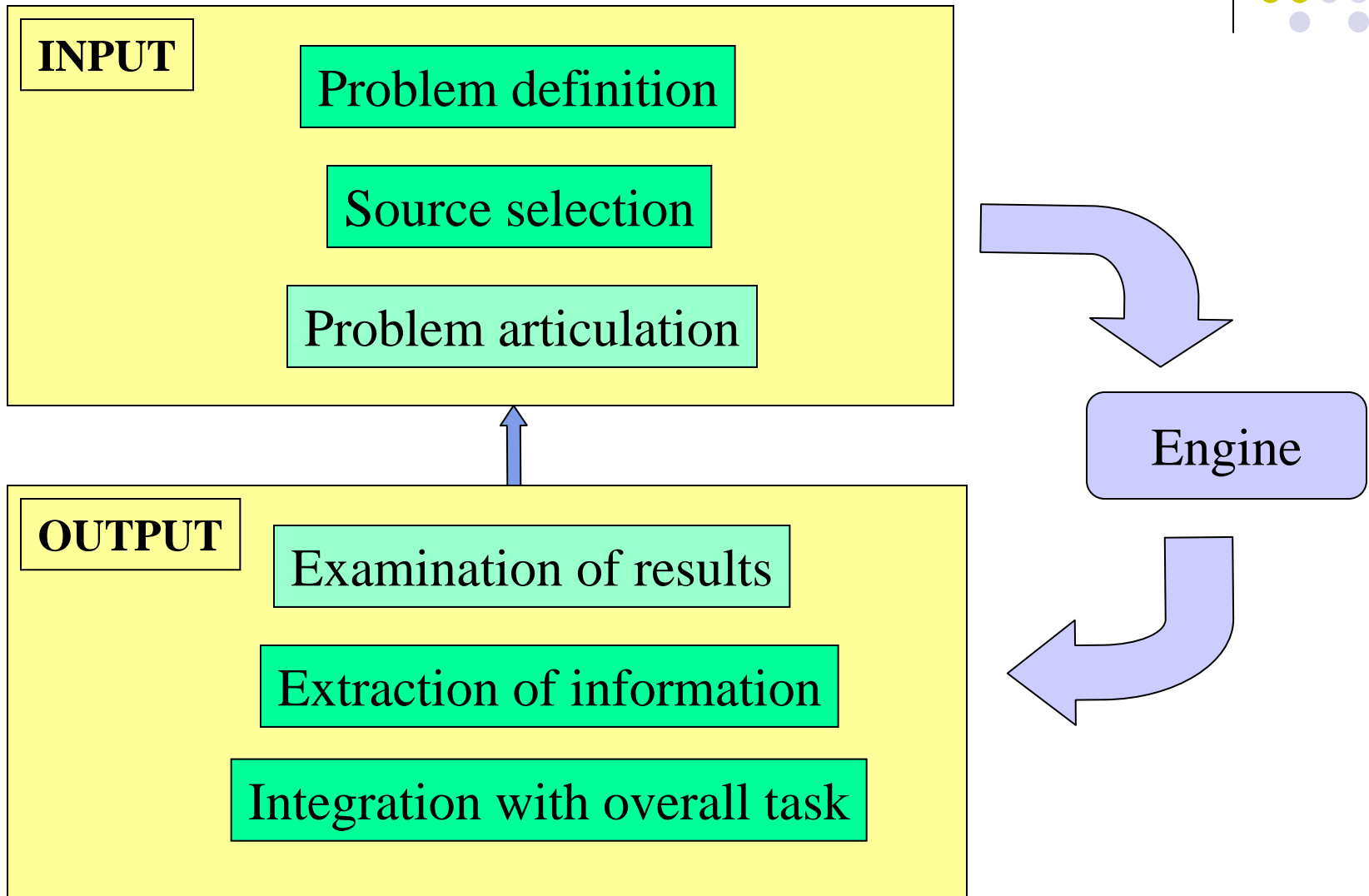
The role of an IR system

– a modern view –



- Support the user in
 - exploring a problem domain, understanding its terminology, concepts and structure
 - clarifying, refining and formulating an information need
 - finding documents that match the info need description
 - as many relevant docs as possible
 - as few non-relevant documents as possible
 - exploring the retrieved documents

Aspects to evaluate

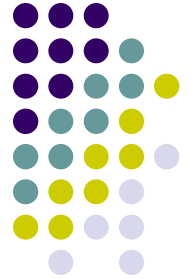




Some IR Evaluation Issues

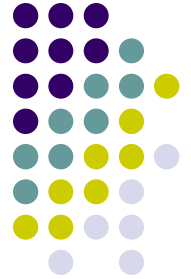
- How best to evaluate performance of the system *as a whole*
- How to be *realistic yet controlled*
- How to gather *sufficient and adequate* data from which it is possible to generalize meaningfully
- How to tailor evaluation *measures and methods* to specific contexts and tasks

Evaluation: IR specific vs. non-specific



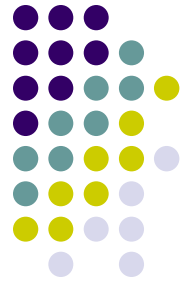
- IR-specific evaluation
 - Systemic
 - Quality of search engine
 - Influence of various modelling decisions (stopword removal, stemming, indexing, weighting scheme, ...)
 - Interaction
 - Support for query formulation
 - Support for exploration of search output
- Non-specific evaluation
 - Task-oriented evaluation
 - Usefulness, usability
 - Task completion, user satisfaction

Task-oriented evaluation (non-IR specific)



- Time to complete a task
- Time to complete a task after a specified time away from the product
- Number and type of errors per task
- Number of errors per unit of time
- Number of navigations to online help or manuals
- Number of users making a particular error
- Number of users completing task successfully

Evaluation: Qualitative vs. Quantitative



- Qualitative:
 - Heuristic evaluation, expert reviews, cognitive walkthroughs etc - preferred if the purpose of the study is to establish the usability of a system;
 - Naturalistic/ethnographic studies - preferred if the purpose of the study is to capture the behavior or preferences of a group of people in a certain setting.
- Quantitative studies:
 - Systematic studies can produce invaluable insight into the effect of various parameters, mathematical models, interaction models, or even of interface elements such as the query formulation mechanism or the layout of the search results
 - Control over experimental variables, repeatability, observability

Measures and dimensions of evaluation



Task Specificity	General	Task-specific
Interactivity		
Non-interactive (laboratory evaluation of the retrieval algorithm)	Effectiveness: Recall, Precision, E, F, Expected search length Efficiency: Time and space complexity	Question answering: mean reciprocal rank (MRR) Filtering: utility Topic distillation: coverage and accuracy
Interactive (evaluation of the interaction process and outcome)	User satisfaction User effort (clicks, iterations, scrolling, documents seen, viewed or read) Effectiveness: Expected search length, Precision at N seen Efficiency: Time to complete task	Aspect retrieval: Aspectual recall, number of saved documents Question answering: completeness and correctness of answer Topic distillation: coverage and accuracy

Interactive TREC



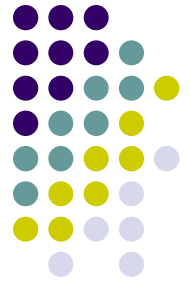
- Human in the loop
- Searcher characteristics influence performance
 - Familiarity with the topic, expertise
 - Searching skills, experience with a certain system
 - Relevance judgments (different from assessors)
- Experimental design needs to take into account user variability
- Real user searches are interactive: multiple queries are submitted, documents from multiple runs are saved
- User studies are expensive (time, effort)

Interactive TREC – a brief history



- TREC 1-8
 - Tasks: routing (initially) & ad-hoc (later)
 - “Manual” (human) intervention in query construction
 - Multiple iterations and relevance feedback was allowed
 - At some point the query was considered final and it was evaluated
 - Results:
 - Manual query formulation beats automatic formulation
 - Insights into the human query formulation and judging process are gained

Interactive TREC – a brief history



- TREC 3
 - Tasks: routing
 - Topic: title, description, narrative
 - Training provided in the form of relevance judgments
 - Results:
 - Humans do not find the routing task natural – they are better at seeking relevant information than at formulating one “best” query
 - Algorithms are better than humans at “learning” from training data

Interactive TREC – a brief history



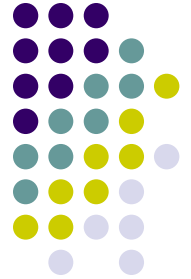
- TREC 4
 - Tasks: ad-hoc
 - Find as many relevant documents for each topic as possible, without collecting too much rubbish
 - Submit the lists of saved documents
 - “Frozen rank” evaluation conducted
 - Construct the final “best” query
 - Submit the top 1000 documents, for comparison to the automatic runs
 - Results:
 - Ad-hoc task more natural than routing
 - “Frozen ranks” difficult to evaluate
 - The main differences observed were between relevance judgments (searcher-searcher, searcher-assessor)

Interactive TREC – a brief history



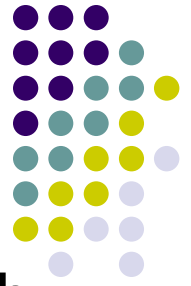
- TREC 5-6
 - Tasks: aspectual/instance recall
 - Find documents that cover as many aspects of a topic as possible; once an aspect is covered, additional documents are not needed
 - Submit: sets of documents
 - Judgments by assessors:
 - Aspectual – for each document, list of aspects covered
 - Binary judgments of relevance for each document
 - Measures of performance
 - Aspectual recall; precision
 - Experimental design:
 - Baseline system (NIST's ZPRISE) allowed inter-site comparisons
 - Results
 - Assessor's judgments inconsistent
 - The experiment is labor intensive; fatigue may have an effect

Interactive TREC – a brief history



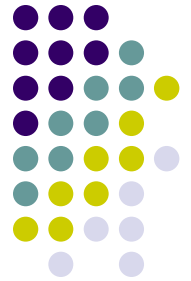
- TREC 7-9
 - Task: aspectual recall
 - Experimental design:
 - Inter-site comparison dropped
 - Participating groups encouraged to evaluate various research hypotheses
 - Number of queries and overall duration reduced
 - Measures:
 - Aspectual recall, aspectual precision, elapsed time
 - Results:
 - No significant differences between baselines and experimental systems
 - Decision to use two-year cycle:
 - Observational (qualitative) studies to identify key issues and generate research questions
 - Detailed metric-based evaluations of research questions

Interactive TREC – a brief history



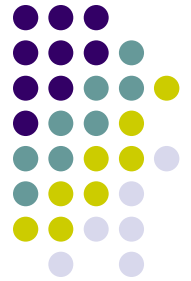
- TREC 10-11, TREC 12 Interactive sub-track of Web track
 - Task: aspectual recall
 - Experimental design:
 - No inter-site comparison
 - Participating groups encouraged to pursue own interests
 - Support in query formulation, effect of output layout, etc
 - Measures:
 - Aspectual recall, aspectual precision, elapsed time, effort
 - Results:
 - Specific to each participating group
 - Development of experimental design, and instruments (questionnaires, interviews etc) widely used in IR use studies

Lessons Learned from the TREC Experience



- IR is inherently interactive
 - measures of search effectiveness alone are insufficient
- Information seeking is engaged in for many different purposes, in many different contexts, to accomplish many different tasks
 - one (or one set of) measure(s) for evaluating IR in general is a Chimera
- It may not be a good idea to rely on external “objective” judgments for evaluation purposes
- Experimental methods can be used successfully in user-centered evaluation of interactive IR

Some conclusions or recommendations



- *Perceptions* of performance are as important as “objective” measures; both should be interpreted w.r.t. measures of the *search process*
- Different measures need to be established w.r.t. goals of different *tasks*
- Specific experimental tasks should be designed so that the *subjects’ performance* in the task, and the *subjects’ own evaluation* of performance, are the criteria for the evaluation measures



References

- **Nicholas J. Belkin and Gheorghe Muresan** *Measuring Web Search Effectiveness: Rutgers at Interactive TREC*, in [Measuring Web Search Effectiveness: The User Perspective](#), workshop at [WWW 2004](#), May 2004, New York ([paper](#), [presentation](#)).
- **Ellen M. Voorhees and Donna Harman** *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005, ISBN 0-262-22073-3.
 - Ch.3: “Retrieval System Evaluation” by Chris Buckley and Ellen Voorhees
 - Ch.6: “The TREC Interactive Tracks: Putting the User into Search” by Susan T. Dumais and Nicholas J. Belkin