

# Evaluating information retrieval systems – parallels with evaluating image processing techniques in medicine

Henning Müller  
Division of Medical Informatics  
University Hospitals of Geneva  
5.5.2003

## Overview

- Why evaluation?
- Information retrieval system evaluation
  - TREC – Text REtrieval Conference
  - Multimedia retrieval
- Evaluation of clinical systems
  - Objectivist and subjectivist approach
  - Verification, validation, assessment of human factors, clinical effects
- Current needs
- Conclusions

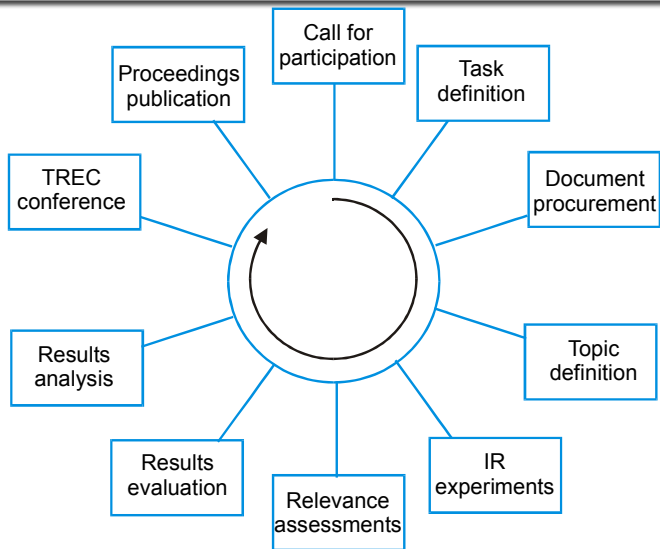
## Evaluation

- Measure the performance of an algorithm/program
- Possibility to show progress in a discipline of research
  - Without standardized evaluation not much can be shown
  - Propagate the use of a system (arguments)
- Benchmarks can lead to large improvements
  - Text retrieval, compression, etc.
  - There are also a few negative effects
- Evaluation of systems in clinical practice has many different aspects
  - Hard factors
  - Soft factors

## IR system evaluation

- Standardized evaluation since the early 1960s
- System comparisons since 40 years
  - Standard data sets
  - Standard performance measures
- Economically interesting area
  - American case law
  - Digital libraries
  - DoD
- TREC is a standard since 1992 (performed by the NIST)
  - Much money is available
  - Many analyses of data have been performed
  - Several different tasks have been introduced

## TREC (circle of events)



## TREC



- Very large databases using millions of textual documents
  - Realistically-sized
  - Old databases for testing, but new database for evaluation each year
  - Large number of varying query tasks/topics (200)
  - Human assessors for ground truth
    - Expensive but the only solution, pooling helps
  - THE standard in the field
- Results of systems are published every year
- Conference itself takes places before the publication of the results
- Choice of presentations is not based on the performance but mainly on originality

## Results of TREC

- Improvements in the domain can be shown
  - Performance more than doubled over less than ten years
- «Friendly» meeting to compare different technologies
- Generation of large amounts of test data
- Many statistical tests on problems and advantages of the evaluation measures and methods
- Inclusion of various domains
  - Video retrieval
  - Query answering
  - Web searching
  - Multi-language retrieval

## Multimedia retrieval system evaluation

- Similar to the evaluation of text retrieval systems
  - Less economic and military interest
  - No databases available free of charge
  - More subjectivity in similarity assessment
- Efforts on the way
  - Creation of reference databases (free of charge)
    - Including ground truth and query tasks
  - Research on performance measures
  - <http://www.benchathlon.net/>
- No accepted standard to evaluate systems
  - No comparison between any two systems is possible
  - Advances in the field cannot be shown



## Evaluation of clinical systems

- There is no real standard and a standard will be hard to find due to large variety and number of people involved
  - Much literature on evaluation
- Objectivist
  - Definition of a gold standard with which all systems are compared
    - Quantitative measures
  - This assumes an agreement on important system properties
- Subjectivist
  - Qualitative measures
  - Assumes that different contexts and observers have varying opinions
  - This is often necessary for complex clinical systems

## Evaluation of clinical systems (2)

- Verification
  - Is the system build correctly and does it what we want it to do?
  - Already during design phase
- Validation
  - Is the system leading to the correct results?
  - Comparison with a gold standard
- Assessment of human factors
  - Will the system be accepted and used?
  - Measurement of user satisfaction
  - User interface
- Clinical effects
  - Length of patient stay, morbidity, resources used

## Conclusions

- Evaluation is absolutely necessary in all research domains
  - To show progress and to have arguments for the use of certain programs or technologies
- Clinical evaluation of imaging systems are complex to pursue
  - Large number of people involved
  - More than just a technical proof
  - Human factors are extremely important for the success of a method
  - We can learn from other domains
- Standard databases and ground truths need to be created, clinically useful tasks need to be defined