Revised Tutorial Submission for ECIR 2009

Current Developments in Information Retrieval Evaluation

Thomas Mandl
University of Hildesheim, Germany
mandl@uni-hildesheim.de

The tutorial introduces and summarizes recent research on the validity of evaluation experiments in information retrieval.

Abstract: In the last decade, many evaluation results have been created within the evaluation initiatives like TREC, NTCIR and CLEF. The large amount of data available has led to substantial research on the validity of the evaluation procedure. An evaluation based on the Cranfield paradigm requires basically topics as descriptions of information needs, a document collection, systems to compare, human jurors to judge the documents retrieved by the systems against the information needs descriptions and some metric to compare the systems. For all these elements, there has been a scientific discussion. How many topics, systems, jurors and juror decisions are necessary to achieve valid results? How can the validity be measured? Which metrics are the most reliable ones and which metrics are appropriate from a user perspective? Examples from current CLEF experiments are used to illustrate some of the issues.

User based evaluations confront test users with the results of search systems and let them solve information tasks given in the experiment. In such a test setting, the performance of the user can be measured by observing the number of relevant documents he finds. This measure can be compared to a gold standard of relevance for the search topic to see if the perceived performance correlates with an objective notion of relevance defined by a juror. In addition, the user can be asked about his satisfaction with the search system and its results. In recent years, there has a growing concern on how well the results of batch and user studies correlate. When systems improve in a batch comparison and bring more relevant documents into the results list, do users get a benefit from this improvement? Are users more satisfied with better result lists and do better systems enable them to find more relevant documents? Some studies could not confirm this relation between system performance and user satisfaction.

Outline

90 minutes:

- Introduction: Repetition of Recall and Precision, Evaluation initiatives
- Presentation: Perspectives on the Cranfield paradigm
- Activity: Relevance judgments for one topic and 15 documents from GeoCLEF, Comparison and Discussion of Results

90 minutes:

- Presentation: Metrics for System Comparison
- Activity: Analysis of real evaluation results (robust CLEF), generation of different system rankings based on different measures and averages, calculation of correlation between system rankings

Lunch break

90 minutes:

- Presentation: Topic Difficulty and Topic Specific Treatment
- Activity: Analysis of real evaluation results (robust CLEF or GeoCLEF), exploration of definitions of topic difficulty

90 minutes:

- Presentation: User Studies and their relation to Cranfield style experiments
- Activity: Discussion on the validity of Cranfield style experiments and the design of user tests

Perspectives on the Cranfield paradigm

Some important questions driving research are: Are relevance judgments worth the money spent? Do they lead to reliable system comparisons? Can fewer judgments also lead to the same results? The role of the human jurors has been explored by measuring their subjectivity. The interrater reliability can be shown to affect the absolute performance values but it only marginally modifies the ranking of the systems. Effort can be saved either by having fewer topics or by judging fewer documents per topics. Obviously, if more topics are developed, the reliability of the results is higher. Research needs to find an optimal balance between reliability of a test and the cost involved.

Research on the following issues will be presented:

- comparison of rankings and correlation between rankings [*Buckley & Voorhees 2005*]
- subjectivity of the jurors and interrater reliability [*Buckley & Voorhees 2005*]
- number of topics necessary for validity [*Zobel & Sanderson 2005*]
- number of systems necessary for validity [*Webber et al. 2008*]
- amount of relevance judgments necessary [*Carterette 2007*]
- active learning [*Moffat et al. 2007*]

Metrics for System Comparison

Many new measures have been introduced within the last few years. The tutorial presents some of them like BPref [*Buckley & Voorhees 2004*], RPrec and NDCG. Their behavior and some current results with these measures from evaluation initiatives are presented [e.g. *Bompada et al. 2007, Sakai 2008*]. Measures for Diversity and Novelty will be briefly mentioned [*Clarke et al. 2008*].

In addition, alternative approaches to aggregate the performance values of individual topics are also discussed. Retrieval systems as well as evaluation measures are desired to be robust. Robust IR means the capability of an IR system to work well (and reach at least a minimal performance) under a variety of conditions (topics, difficulty, collections, users, languages …). Robustness might be measured with the geometric mean of a set of topics instead of the mean average [*Robertson 2006, Mandl et al. 2009*].

Topic Difficulty and Topic Specific Treatment

The evaluation of robust retrieval has been motivated by the fact that the variance for topics has been very large even for top performing systems. Even these good systems achieve only poor results for some topics. Improving on these topics would greatly enhance their overall quality as perceived by the user. Users remember poor performance often better than excellent performance. It is important to "ensure that all topics obtain minimum effectiveness levels" [*Voorhees 2005*]. Systems could also try to guess which topics might be difficult and apply appropriate methods to them.

Typical distributions of topic performances are shown. Categorizations for reasons for failure are presented [*Harman & Buckley 2004, Mandl et al. 2006, Savoy 2007*]. Systems which adapt to the features of a query in order to optimize the results are very promising [*Kwok 2005, Zaragoza 2009*].

User Studies

User studies try to compare retrieval systems by measuring user satisfaction or performance directly without introducing jurors who try to act as "average" users. The methodology for user tests is taken from human-computer interaction. Some recent studies tried to test the validity of user studies by checking whether users notice that one system is better than the other. Some studies could not confirm this relation between system performance and user satisfaction [*Turpin & Scholer 2006, Turpin & Hersh 2001, Al-Maskari et al. 2006*]. These experiments are presented and discussed. A large study of the author complements this section. Expectation as defined in models of customer satisfaction is introduced as a factor which influences the satisfaction. Like previous studies we revealed that users significantly relax their relevance criteria when faced with a bad system and compensate for low performance [*Lamm et al. 2009*]. Other studies on the modification of the user behavior facing different systems are included [*Smith & Kantor 2008, Scholer & Turpin 2008*]. A brief look on click-through data analysis concludes the section.

Biographical sketch

Thomas Mandl is assistant professor at the University of Hildesheim in Germany where he is teaching in the programme *International Information Management*. He has received a doctorate degree on neural networks in information retrieval and a post doctoral degree on quality in web information retrieval. His research interests also include human-computer interaction, internationalization of information systems and applications of machine learning. He has been coordinating the GeoCLEF evaluation track on geographic queries and the robust task at the Cross Language Evaluation Forum (CLEF) and is currently developing the LogCLEF track on log file analysis.