Beyond Classical Measures: How to Evaluate the Effectiveness of Interactive Information Retrieval System?

Azzah Al-Maskari
Dept. of Information Studies University of Sheffield
Sheffield, S1 4DP, UK
Lip05aaa@shef.ac.uk

ABSTRACT

This research explores the relationship between Information Retrieval (IR) systems' effectiveness and users' performance (accuracy and speed) and their satisfaction with the retrieved results (precision of the results, completeness of the results and overall system success). Previous studies have concluded that improvements in IR systems based on increase in IR effectiveness measures do not reflect on improvement in users' performance. This work aims at exploiting factors that can possibly be considered as confounding variables in Interactive Information Retrieval (IIR) evaluation. In this research, we look at substantive approaches to evaluate IIR systems. We aim to build an interactive evaluation framework that brings together aspects of systems' effectiveness and users' performance and satisfaction. This research also involves developing methods for capturing users' satisfaction with the retrieved results of IR systems, as well as examination how users assess their own performance in task completion. Furthermore, we are also interested in identifying evaluation measures which are used in batch mode (non-interactive experiment), but correlate well in interactive IR systems. Thus, by the end of this research, we hope to develop a valid and reliable metrics for IIR evaluation.

A first study was set up to explore the relationship between system effectiveness as quantified by traditional measures, such as precision and recall, and users' effectiveness and satisfaction of the results, though this study was limited to few users. The tasks involve finding images for recall-based tasks. It was concluded that no direct relationship between system effectiveness and users' performance. People learn to adapt to a system regardless to its effectiveness. This study recommends that a combination of measures (e.g. system effectiveness, user performance and satisfaction) to be used to evaluate IIR systems. Based on our observation from this study, we found that users' familiarity of the search topic has increased their performance.

Thus, we set up a second experiment to investigate how users' satisfaction correlate with some IR effectiveness measures such as precision and the suite of Cumulative Gain measures (CG, DCG, NDCG) in simple web searching tasks. Results from this study have shown that CG and Precision are better than NDCG at reflecting users' satisfaction with the results of an IR system. We have also concluded that users of web search engines, in the context of simple search task, are more concerned with precision than completeness of the search. This stemmed from the stronger correlation between users' satisfaction with the success of overall search and their satisfaction with the accuracy of the search results than with their satisfaction with the completeness of the search.

Many scholars such as [1], [2], [3], and [4] have recommended considering perceptions of the users as important as IR effectiveness measures, and both should be interpreted as measures of effectiveness. Therefore, the issue in IIR evaluation should not be focusing on maximizing the retrieval performance, by refining IR techniques alone, but also understanding users' satisfaction, behaviors and information needs. This raises the need for more investigation on measures that translate users' performance and satisfaction as the criterion of a system. Indeed, the need for effective and efficient evaluation of IIR is very important.

Future plans are to incorporate variables domain knowledge, motivation, task complexity and search behaviours on user performance and users evaluation of IR system performance when evaluating interactive IR systems; this is in an attempt to explore the suitability of different measures in IIR evaluation. Thus, the proposed approach adopts a systematic and multidimensional approach to evaluation including not only classical traditional evaluation measures, such as precision and recall, but also interactive non-traditional measures, such as users' characteristics and their satisfaction.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Performance

REFERENCES

[1] Belkin, N. J., Muresan, G. & Zhang, X.-M. Using User's Context for IR Personalization. *SIGIR-04.* Sheffield, UK. 2004

[2] Järvelin, K. & Ingwersen, P. Information seeking research needs extension towards tasks and technology. *Information Research,* 10**,** 212. 2004

[3] Turpin, A. & Scholer, F. User Performance versus Precision Measures for Simple Search Tasks. *SIGIR.* Seatle, Washington, USA. 2006

[4] Su, L. T. (1992) Evaluation measures for interactive information retrieval. *Information Processing & Management,* 28**,** 503-516.

SIGIR' 07, July 23-27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007