Task based evaluation of exploratory search systems

Wessel Kraaij
TNO Information and Communication Technology
Brassersplein 2
Delft, The Netherlands
kraaijw@acm.org

Wilfried Post
TNO Human Factors
Kampweg 5
Soesterberg, The Netherlands
wilfried.post@tno.nl

ABSTRACT

Evaluation of interactive search systems has always been time-consuming and complex, which probably explains the relative low level of interest from IR researchers for this type of evaluation in the past. Yet the limitations of batch-style system evaluations cannot be ignored anymore. We present some case studies of evaluations in interactive settings. Several of these evaluations offer valuable new insights about system adequacy. This more than compensates for the reduced ability to reproduce results. We distinguish system centered evaluations focusing on performance and user centered (task based) evaluations focusing on adequacy. The latter take the natural task of a user as starting point. Task based evaluations suggest that proper HCI design is probably a more important factor for user satisfaction than the quality of statistical indexing and ranking methods. User centered and system centered evaluations of interactive systems measure different aspects of quality. The challenge is to design an evaluation where the different components that determine system adequacy and performance can be identified and their relationship can be quantified.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*User-centered design*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*evaluation/methodology*

General Terms

Measurement, Performance, Human Factors

Keywords

Task based evaluation, interactive search, meetings

SIGIR 2006 workshop, Evaluating Exploratory Search Systems Seattle, USA

1. INTRODUCTION

Modern information professionals are used to access and share information in a multitude of ways using various repositories. A lookup search query in a search engine is not the predominant search method anymore, search is often accompanied by browsing for more complex tasks like learning and investigation [8]. Search engines experiment with interactive functions, become context aware and get increasingly personalized. Techniques for structuring result lists become more mature (clustering, faceted browsing etc.). These exploratory search systems pose new challenges to the IR community. The traditional batch style experiments (Cranfield/TREC) have been attractive for IR researchers (and even inspired evaluations in other communities such as natural language processing), since experiments were easy to conduct, and well controlled because humans were excluded from the loop. Still many researchers felt that these studies were limited, since they failed to model a real search process.

Evaluation types. The component based evaluation which is the model for TREC is sometimes referred to as intrinsic evaluation in contrast to an evaluation where the component's performance is measured in the user context (extrinsic). When evaluating a complete system, intrinsic evaluation approximates *performance* evaluation and extrinsic evaluation is related to *adequacy* measurement[6] [1]. Performance measurements are usually aimed at comparing systems, whereas adequacy measurements focus more on the usability for an end user. But also cost-effectiveness could be an important factor determining adequacy. Performance is most probably one of the contributing factors to adequacy, if the system is doing something useful. In practice, "adequacy" is the most important aspect for the "acceptance" of a system by end-users. However task based evaluations are not so often reported in literature. This is strange since it is well known that there is a strong link between task complexity and search behaviour [1].

In section 2, several examples of evaluations of interactive will be discussed, to illustrate that the focus of the evaluation is sometimes on performance, sometimes on adequacy. In section 2.5 in particular, we will outline an extrinsic evaluation framework that is currently applied for the evalua-

[1]Note that intrinsic system evaluation is not necessarily synonymous to system centered evaluation, since a system could contain a user model in the form of personalization. On the other hand, an extrinsic evaluation can be rather system oriented if it is mostly concerned with system performance.

tion of a meeting browser[2]. The paper is concluded with a discussion about the strengths and weaknesses of the different approaches to the evaluation of interactive information systems.

2. SHORT CASE STUDIES OF INTERACTIVE SYSTEM EVALUATIONS

In the following subsections we will discuss some case studies of research projects and evaluation programs which have shaped our ideas concerning the evaluation of interactive search systems[3]. We will discuss the different evaluations in terms of user centered (adequacy) vs. system centered (performance) evaluations.

2.1 Interactive track at TREC

For nine years an interactive task was included at TREC. The task evolved from interactive query modification for adhoc and routing, via aspectual retrieval and a factoid QA task, to a Web task [4]. Over the years, various experimental designs were tried, an experiment with cross-site comparisons was discontinued, since the additional overhead involved did not pay off in terms of results. In later years, the track focused on within site experiments, applying a 2 year schedule, giving room for user centered observational studies and more system oriented experiments.

2.2 Video retrieval (TRECVID)

At TRECVID, the annual benchmark conference for video indexing and retrieval, a search task has been studied for five years now. In the automatic task, a query has to be constructed automatically from a topic description, interaction is not allowed. For manual runs, the query can be constructed by the experimenter. Interactive runs allow in addition to refine queries and modify the ranked result list. In the beginning, interactive or manual search was a pure necessity, since automatic query construction in terms of constraints on low level image features resulted in very poor performance. In the mean time, automatic search results have reached almost the same level as manual search, but still interactive search (where users are allowed to interact with the system *after* processing the initial query) performs significantly better[9]. Recent years of TRECVID search have consistently showed that a two step paradigm consisting of iterative query refinement in combination with manual cleaning of the result list provided highly competitive results. For both tasks a well-designed GUI is a must. Last TRECVID (20050 showed an experiment pushing human perceptual limits by applying the Rapid Serial Visual Presentation method for selecting shots from a list[5]. Other sites (e.g. [12]) experimented with advanced visual browsers in order to optimize local browsing within a shot and between adjacent shots.

2.3 Broadcast news analysis system

Novalist is a system for the analysis of various news sources including newspaper, websites and TV programs [3] . The system applies temporally biased document clustering, followed by automatic metadata extraction and has its roots in prototypes that were built for the TDT and DUC evaluations. Novalist has been conceived as an exploratory search system combining search with browsing structured result sets, catalogue search, browsing through individual issues of newspapers, magazines or TV programs, timeline based browsing and a standard keyword search pane. The system was piloted by a government organization interested in financial activities. The extrinsic evaluation of the system consisted of two components: a qualitative questionnaire and interview based evaluation and a quantitative task based performance evaluation. The latter evaluation consisted of re-running an analysis task (creating a dossier on a specific entity). Quantitative results could be measured since timesheets for the original investigation were on file and the search result (in terms of retrieved relevant documents) could be compared with the result of the original search (using the existing working method). The qualitative method also yielded interesting results, since many useful system improvements could be distilled from the answers. While the individual components of the system performed well in intrinsic evaluations [13, 7], the task based (extrinsic) evaluation shows several important areas for improving the adequacy of the system for operational tasks e.g. the wish for having a better integration of the pilot system into the work task of the individual investigator (persistence of search result context).

2.4 Browser for meeting recordings archive

Meetings are an object of active research in the area of multimodal analysis. In the context of the EU project AMI (Augmented Multiparty Interaction) a collection of 100 hours of meetings has been recorded and annotated [2]. The majority of the meetings are based on a scenario (i.e. they are more or less controlled, acted meetings). The scenario is based on a design team working on a new remote control. Each of the 4 team members has a distinct role: project manager, UI designer, technical designer or marketing expert. Each design project consists of 4 meetings, reflecting distinct stages in the project. Approximately 30 series of design meetings have been recorded at three different labs in Europe using multiple sensors (overview and close-up cameras, far-field and close talking microphones, smart pens etc.), resulting in a multimedia meeting archive. The multimedia data has subsequently been manually and automatically annotated for various semantic features, such as transcripts, movements and disussion topics.

Several meeting browsers have been developed to access the archive. These browsers serve two purposes: either as an analysis instrument for the researchers, but more importantly as an access tool for a multimedia archive, to be used by end users. It is the latter function that is of interest for the scope of this paper. Currently two types of browser evaluation methodologies have been developed within AMI for the end-user test. The first method: BET (Browser evaluation test) is modeled as an efficiency test [14]. Test subjects are asked to answer questions, which require browsing the meeting archive. Questions are based on a random sample from a pool of "observations of interest" that have been annotated by assessors. The second method [10] focuses on team effectiveness as a whole and is based on a procedure involving questionnaires and a model based evaluation. A meeting browser has the potential to substantially increase the effectiveness/efficiency of a team, but its contribution

[2]Full details of the framework are described in [10].

[3]We do not claim that the selection of these case studies is a representative sample of interactive IR studies.

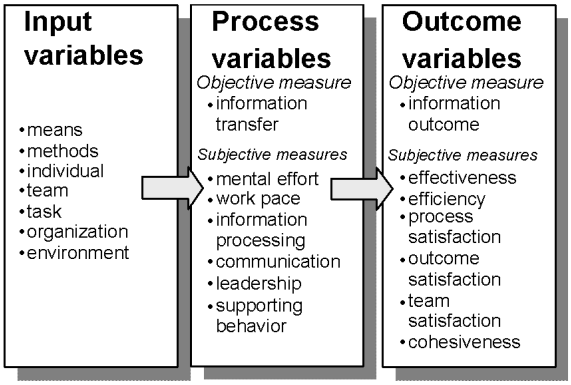

Figure 1: Meeting evaluation framework

is measured rather implicitly in comparison with the BET procedure.

2.5 Proposed evaluation of a task oriented meeting browser

The focus on a task-oriented setting has inspired a complete re-design of the meeting browser. The new meeting browser will be optimized for end-users instead of researchers. Central data structure in the GUI of the meeting browser will be the project plan structure, with hyperlinks into relevant meeting segments in the archive. Evaluation of the meeting browser will be based on a specific scenario, where subjects are instructed to replace an existing team and resume their activities. Design team members will use the meeting archive in order to get ready for their new task. Evaluation will be based on the method described in , consisting of objective and subjective measures (questionnaires)

The evaluation method will be based on a framework in which various factors for successful meetings are related (see fig. 1 and [10]). The task oriented meeting browser - a particular meeting means - should be regarded as an input factor. Together with other input factors, such as the particular meeting method used, characteristics of individuals and the team (including roles), the particular task type (here design), and specifics of the organization (such as culture) and its environment (e.g., market demands), the factors determine how well a meeting process takes place, and consequently how well meeting outcomes are reached. Three core process factors are distinguished: the transfer of necessary information between the participants, the workload of the participants, and team behaviour (such as communication, leadership and supportive behaviour). Four basic outcome factors are distinguished as well: information outcome (are the exchanged information indeed used to make the right decision, or to solve a problem), effectiveness (were the right decisions taken and the problems solved), efficiency (was this done with minimal time and effort), and satisfaction. In this evaluation method, the objective process and outcome factors are determined by analysing the information flow. The subjective process and outcome factors are determined by means of questionnaires and rating scales before and after each meeting.

The large set of factors illustrates the relatively small contribution of the factor "means" on performance outcome. The impact of a means should be seen in a broader context of all other factors. Our task-oriented meeting browser takes several input factors into account at once. It is a particular means (such as a meeting browser) for a particular method (well defined design meetings within the context of a design project), and makes use of individual and team characteristics (retrieval will be based on individual history and role description) and deeper knowledge of a particular task (design). We therefore expect that the browser will have a broader impact on performance outcome.

3. DISCUSSION AND CONCLUSIONS

The various cases of evaluations of interactive systems show quite a diversity in task-setup and focus. The system oriented "TREC-style" evaluation focuses on a well defined uniform task. A system is tested by a number of instances of this task, in order to control for variations in query difficulty (an important determinant of system performance). Such an experimental set-up improves the generalizability, but has the danger to zoom in on just a single quality aspect. A user oriented (HCI) evaluation measures the outcome of the user's task as a whole and tries to gauge the influence of the system on the user's performance in the task. It is clear that compromises have to be made here with respect to the goal to test many "topics" in order to maintain a good generalizability. But since a task based evaluation comprises a more complete model of a user's task, such a method might very well detect important determinants for adequacy that would be overlooked in a system centered evaluation.

User centered evaluations are costly. The question is whether that's a reason to neglect extrinsic evaluations. We have shown that task based evaluations spawn interesting research on the cross-roads of HCI and IR. Examples of interesting topics include personalized systems and GUI's optimized for a certain task. A disadvantage of scenario based task oriented evaluations is that the setting is rather specific, it's therefore not clear whether results generalize well.

On the other hand, this specificity can lead to new, unforeseen IR improvement. In the example of the task-oriented meeting browser, search behaviour of one team member may lead to automatic IR improvement for another team member. Moreover, interpreting the information needs of a team member may also lead to identifying another type of information source: a colleague team member, who you can consult for the information (which is a quite common team feature). Or even on an organizational level, another team. It is exactly these new types of retrieval solutions that will not be found only with a system oriented IR approach.

IR researchers can learn a lot from the experimental traditions that are commonplace in social sciences, such as a comparative study of the factors that have an impact on the adequacy/performance of a system. On the other hand HCI researchers can benefit from research on search behaviour, e.g. [11]. An important research question requires expertise from both fields: "what are the determinants for system adequacy, what is their relative importance and can we identify dependencies between these factors.

4. ACKNOWLEDGMENTS

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-178).

5. REFERENCES

[1] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Information Processing and Management*, 31(2):191–213, 1995.

[2] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005.

[3] F. de Jong and W. Kraaij. Novalist: Content reduction for cross-media browsing. In *RANLP workshop Crossing Barriers in Text Summarization Research*, 2005.

[4] S. T. Dumais and N. J. Belkin. *TREC Experiment and Evaluation in Information Retrieval*, chapter The TREC Interactive Track: Putting the User Into Search, pages 123–152. MIT Press, 2005.

[5] A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 skirmishes. In *Proceedings of TRECVID 2005*, 2005.

[6] L. Hirschman and H. S. Thompson. *Survey of the State of the Art in Human Language Technology*, chapter 13.1 Overview of Evaluation in Speech and Natural Language Processing. 1996.

[7] W. Kraaij, M. Spitters, and A. Hulth. Headline extraction based on a combination of uni- and multidocument summarization techniques. In *Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002)*. ACL, June 2002.

[8] G. Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 2006.

[9] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton. TRECVID 2005 an overview. In *Proceedings of TRECVID 2005*. NIST, 2005.

[10] W. M. Post, M. H. in 't Veld, and S. van den Boogaard. Evaluating meeting support tools. *Personal and Ubiquitous computing*. submitted.

[11] T. Saracevic, P. Kantor, A. Y. Chamis, and D. Trivison. A study of information seeking and retrieving. i. background and methodology. *Journal of the American Society for Information Science*, 39.

[12] C. G. M. Snoek, J. C. van Gemert, J. M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. D. Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, and M. Worring. The Mediamill TRECVID 2005 semantic video search engine. In *Proceedings of TRECVID 2005*, 2005.

[13] M. Spitters and W. Kraaij. Unsupervised event clustering in multilingual news streams. *Proceedings of the LREC2002 Workshop on Event Modeling for Multilingual Document Linking*, pages 42–46, 2002.

[14] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker. A meeting browser evaluation test. In *CHI Extended Abstracts 2005*, 2005.