

EESS 2006

Proceedings of the
**ACM SIGIR 2006 Workshop on
"Evaluating Exploratory Search Systems"**

<http://research.microsoft.com/~ryenw/eess>

Organizers:

Ryen W. White, Microsoft Research
Gheorge Muresan, Rutgers University
Gary Marchionini, University of North Carolina

Sponsored by:

Microsoft

In conjunction with:



Preface from the Workshop Co-Chairs

Welcome to the EESS 2006 Workshop! Our vision for this workshop is to bring together researchers from communities such as information retrieval, library and information sciences, and human-computer interaction for a discussion of the issues related to the formative and summative evaluation of Exploratory Search Systems (ESS). Exploratory search represents a shift from the analytic approach of query-document matching toward direct guidance at all stages of the information-seeking process. It can be seen as a specialization of information exploration – a broader class of activities where new information is sought in a defined conceptual area. Whilst search systems are expanding beyond supporting simple lookup into supporting complex information-seeking behaviors, there is no established framework for how to evaluate this genre of system. The focus in recent years has been on the development of new systems and interfaces, and the evaluation of these systems has been generally neglected. Given the range of technology now available we feel that the time has come to turn attention toward evaluating systems that support exploratory search. Our general aims for the workshop are to:

- Define metrics to evaluate ESS performance
- Establish what ESS should do well
- Influence ESS designers to think more about evaluation
- Discuss components for the non-interactive evaluation of ESS

Through paper and panel presentations, break-out sessions, and discussions, we hope to identify the issues pertinent to the evaluation of ESS in a way that will be beneficial to participants in their own endeavours, and to the research community through an opportunity to interact and explore the issues in this challenging area. We are very grateful to the special guest panelists: Andy Edmonds, Tom Landauer, Cathy Marshall, and Pete Pirolli, and the following people for taking the time to serve on the program committee, and review submissions for the workshop:

Amanda Spink, Queensland University of Technology, Australia

Anastasios Tombros, Queen Mary University of London, UK

Bill Kules, University of Maryland, USA

Birger Larsen, Royal School of LIS, Denmark

Daniela Petrelli, University of Sheffield, UK

Daqing He, University of Pittsburgh, USA

Diane Kelly, University of North Carolina, USA

Ed Cutrell, Microsoft Research, USA

Ed Fox, Virginia Tech, USA

Edie Rasmussen, University of British Columbia, Canada

Ian Ruthven, University of Strathclyde, UK

Jacek Gwizdka, Rutgers University, USA

Jaime Teevan, MIT, USA

Jim Jansen, Pennsylvania State University, USA

m. c. schraefel, University of Southampton, UK

Pia Borlund, Royal School of LIS, Denmark

Sherry Koshman, University of Pittsburgh, USA

Steven Drucker, Microsoft Research, USA

Xiangmin Zhang, Rutgers University, USA

In addition, we are grateful to Microsoft for their financial support. Participating in this event is an outstanding opportunity to influence the future of evaluating interactive systems. We hope you enjoy being part of the workshop, and are encouraged to join us on the journey that will follow!

Ryen W. White, Gheorghe Muresan, and Gary Marchionini

EESS Workshop Co-Chairs, June 2006

Table of Contents

Introduction

- **Evaluating Exploratory Search Systems** 1
R.W. White (*Microsoft Research, USA*), G. Muresan (*Rutgers University, USA*),
G. Marchionini (*University of North Carolina at Chapel Hill, USA*)

Panel papers

- **Building Models of Search Success with Experience Sampling and Event Logs** 3
A. Edmonds (*Microsoft, USA*)
- **Retrieval Evaluation sans Human Relevance Judgments** 5
T.K. Landauer (*University of Colorado / Pearson Knowledge Technologies, USA*)
- **Why a Corpus-Topics-Relevance Judgments Framework Isn't Enough: Two Simple Retrieval Challenges from the Field** 7
C. Marshall (*Microsoft, USA*)
- **Analysis of the Task Environment of Sense Making** 9
P. Pirolli (*Xerox PARC, USA*)

Presented papers

- **Layered Evaluation of Adaptive Search** 11
P. Brusilovsky, R. Farzan, JW. Ahn (*University of Pittsburgh, USA*)
- **Wrapper: An Application for Evaluating Exploratory Searching Outside of the Lab** 14
B. J. Jansen, R. Ramadoss, M. Zhang, N. Zang (*The Pennsylvania State University, USA*)
- **Exploratory Search Visualization: Identifying Factors Affecting Evaluation** 20
S. Koshman (*University of Pittsburgh, USA*)
- **Task-based Evaluation of Exploratory Search Systems** 24
W. Kraaij, W. Post (*TNO, The Netherlands*)
- **Methods for Evaluating Changes in Search Tactics Induced by Exploratory Search Systems** 28
B. Kules (*Takoma Software, USA*)
- **An Integrated Approach to Interaction Modeling and Analysis for Exploratory Information Retrieval** 30
G. Muresan (*Rutgers University, USA*)

Background papers

- **Exploratory Search in Wikipedia** 35
S. Fissaha and M. de Rijke (*University of Amsterdam, The Netherlands*)
- **A Pilot for Evaluating Exploratory Question Answering** 39
V. Jijkoun, M. de Rijke (*University of Amsterdam, The Netherlands*)
- **Impact of Relevance Intensity in Test Topics on IR Performance on Polyrepresentative Exploratory Search Systems** 42
B.R. Lund, P. Ingwersen (*Royal School of Library and Information Science, Denmark*)
- **From Question Answering to Visual Exploration** 47
D. McColgin, M. Gregory, E. Hetzler, A. Turner (*Pacific Northwest National Laboratory, USA*)
- **What do the Attributes of Exploratory Search Tell us about Evaluation** 51
Y. Qu, G.W. Furnas (*University of Michigan, USA*)

Evaluating Exploratory Search Systems

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
ryenw@microsoft.com

Gheorghe Muresan
School of Communication,
Information and Library Studies
Rutgers University
New Brunswick, NJ 08901 USA
muresan@scils.rutgers.edu

Gary Marchionini
School of Information and
Library Science
University of North Carolina
Chapel Hill, NC 27599 USA
march@ils.unc.edu

Online search has become an increasingly important part of the everyday lives of most computer users. Generally, popular search tools support users well, however, in situations where the search problem is poorly defined, or the information seeker is unfamiliar with the problem domain, or the search task requires some exploration or the consideration of multiple perspectives, such tools may not operate as effectively. To address situations where technology may not meet their needs, users have developed coping strategies involving the submission of multiple queries and the interactive exploration of the retrieved document space, selectively following links and passively obtaining cues about where their next steps lie. This is an example of *exploratory search* behavior, and comprises a mixture of serendipity, learning, and investigation [7].

Exploratory search can be seen as a specialization of information exploration – a broader class of activities where new information is sought in a defined conceptual area. It represents a shift from the analytic approach of query-document matching toward direct guidance at all stages of the information-seeking process. Through functionalities such as tabbed browsing and dynamic queries [10], Exploratory Search Systems (ESS) are helping users run multiple threads in parallel, and see the immediate impact of their decisions. By following hyperlinks, people can better define and refine their information problem, and bring it closer to resolution. Browsing is a serendipitous activity that can be attractive to users, who may benefit from the extraneous information if they have long-term interest in a particular topic, but is inefficient for fact-finding or known-item retrieval, so is therefore not appropriate for all circumstances [8].

Browsing a new document collection, beginning to gather information on a new topic, or trying to resolve an ill-defined problem, can be likened to the exploration of a maze in the physical world; the process is fraught with uncertainty, one is never able to see more than one step ahead at any given time, and the navigation of the maze comprises a series of on-the-fly selections that can impact the success of the journey. Analytical strategies that provide us with a ranked list of documents can be seen as providing a point or entry to the maze, or even dropping us in the middle. However, to find the prize at the center of the maze (or escape from it!) there is a need to provide tools to support navigation and decision-making. For example, finding one's way through the maze becomes much easier if a visual representation of the space being explored is provided (e.g., map with current location indicated). Now, imagine that the maze is multi-

dimensional, and that choices at each intersection are not limited to one out of two, three, or four possibilities, but rather tens and hundreds of possibilities, as is the case with exploring search results. The design of interfaces to help users navigate these complex environments is a crucial part of supporting exploratory search, and outweighs the analytic strategies prevalent in current search systems, which serve to parachute us into a starting point. As the articles in a recent issue of Communications of the ACM entitled "Supporting Exploratory Search" [11] demonstrate, research into the development of interfaces to support the understanding of information, rather than simply finding it, is gathering pace in communities such as human-computer interaction, information retrieval, library and information science, psychology, and beyond.

Exploratory search systems are capitalizing on new technological capabilities and interface paradigms that facilitate an increased level of interaction with information. However, evaluation of search systems has remained limited to those that support minimal human-machine interaction. Since the days of the Cranfield experiments some 40 years ago, the issue of evaluating retrieval systems has been considered highly important by the Information Retrieval (IR) community [2]. The annual NIST-sponsored Text Retrieval Conference (TREC) has provided a medium for the evaluation of algorithms underlying the analytic aspects of IR systems, yet struggled because the experimental methods of batch retrieval are not suited to studies of interactive IR. Since TREC-3, the conference has extended its mandate to recognize the importance of the user in information-seeking. The Interactive Track [3], and later the HARD track [1] have both attempted to bring the user into the loop. However, these tracks struggled to establish comparability between experimental sites, in terms of the experimental systems devised and the measures used. They were also adversely affected by the dependence on relevance judgments and interactions between users, tasks, and systems. Nonetheless, the Interactive Track was successful at highlighting the importance of users in information-seeking [5].

The more interactive options an application has, the greater the number of variables, and therefore the larger the likelihood for experimental confounds if compared against other systems. For example, a system with features A, B, C, D, and E should theoretically be compared against 119 other systems that vary the presence and absence of these five features. Even if the experimenters make pragmatic decisions about the number of experimental variations, it is still challenging to limit the number of comparator systems whilst maintaining control of the number of possible experimental confounds. This does not include the time required to complete the experiments, build the systems, and train the subjects using the systems.

Copyright is held by the author/owner(s).

SIGIR'06 Workshop, August 10, 2006, Seattle, Washington, USA.

Additionally, it may often be the case that the sum of the features in an ESS may lead to a different experience than the individual features in isolation (i.e., the interactions between features may be just as important as the features themselves). High levels of interaction, which are an integral part of exploratory search, pose a real evaluation challenge: there is potential for confounding effect of different exploration tools, the desired learning effect is difficult to measure, and the potential effect of fatigue limits the evaluation to a low number of topics, which makes it rather difficult to get the statistic significance required by a meaningful quantitative analysis.

The research community has focused for some time on how to develop novel interfaces to support users engaged in exploratory search. However, given the range of ESS now available, it is time to shift the focus of research toward understanding the behaviors and preferences of searchers engaged in exploratory searching, on tasks supported by such systems, and on measuring exploration success. For example, a key component of exploration is human learning (a topic studied extensively by cognitive psychologists [6]) yet this issue has not been explored in relation to ESS. Any evaluation of ESS should consider at least two factors: *metrics* (i.e., what is going to be measured?), and *methodologies* (i.e., how we are going to measure it?).

Metrics: The outcomes of the search and the search process itself can be used to evaluate the effectiveness of ESS. For example, assessments of relevance or utility by subjects during or after the search, structured or informal subjective evaluations, and examination of the resultant products or artifacts, all give insight into the effectiveness of the ESS. However, they give limited insight into how well systems support cognitive processes such as learning. One way to get access to such information is to look at users' interactions during their search. Behaviors can be seen as manifestations of internal information-seeking strategies. An examination of paths taken and decisions made during a search can allow us to make inferences about cognitive activity [8].

Methodologies: The approach taken to evaluate ESS is crucial. If possible, experiments should be longitudinal, and take place in a naturalistic setting. The task domain should contain a mixture of task types: some that relate closely to subjects regular activities, and some that are completely new. A challenge of ESS evaluation is to elicit exploration, and this can be more problematic if subjects are only engaged in tasks they are familiar with. Subjects should be classified based on familiarity with the topic or problem domain, expertise or frequency of using the retrieval system, and general range of computer experience. The setting and task domain should be controlled by the experimenter, to allow focus on the user and the system components of information-seeking. To counteract learning or order effects that may compromise the reliability of the experimental findings, there should be systematic variation of the independent variables in the experiment. Exploratory search is a cognitively intensive activity, and subjects should be allowed to conduct their searches with minimal interruptions. Techniques such as questionnaires and interview techniques can be valuable tools, but one must be careful to include them in the experiment in such a way as to not interfere with their exploration. If multiple sites are going to be involved in the experiment then care should be taken to coordinate planning and execution carefully.

Evaluating ESS is not substantially different from evaluating any other highly interactive system. Whilst of course we should be concerned with subjective measures such as user satisfaction and task outcomes, it is through the measurement of interaction behaviors, cognitive load, and learning that we can get a clear picture of how effective such systems can be. There are research opportunities to develop frameworks for the evaluation of ESS that incorporate such measures. The approach adopted at TREC has led to the rapid development of effective ranking algorithms for document retrieval. As a result of such research, search systems such as MSN Search and Yahoo! cope well with navigational requests (e.g., find a given person's homepage), and closed informational requests (e.g., answer to a question which has a single answer). However, none of these systems provides the explicit functionality to support exploration. It has been suggested that repositories of data and tasks (similar to TREC) could be used to evaluate ESS based on information visualization [9]. Our vision is of a framework for ESS evaluation that could validate the support these systems offer, and chart new courses toward improved search experiences for users.

REFERENCES

- [1] Allan, J. (2003). HARD Track Overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of the Text Retrieval Conference*, pp 24-37.
- [2] Cleverdon, C.W., Mills, J., and Keen, M. (1966). *Factors determining the performance of indexing systems*. ASLIB Cranfield project, Cranfield.
- [3] Dumais, S. and Belkin, N.J. (2005). The TREC Interactive Track: Putting the user into search. In Voorhees, E. and Harman, D. (Eds.) *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press.
- [4] Kelly, D. (2004). *Understanding implicit feedback and document preference: A naturalistic user study*. Unpublished doctoral dissertation, Rutgers University.
- [5] Lagergren, E. and Over, P. (2001). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 164-172.
- [6] Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The psychology of learning and motivation*, 41: 43-84.
- [7] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4): 41-46.
- [8] Marchionini, G. and Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1): 70-79.
- [9] Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced Visual Interfaces*, pp. 109-116.
- [10] Shneiderman, B. and Plaisant, C. (2005). *Designing the User Interface 4th Ed.*, Person/Addison-Wesley.
- [11] White, R.W., Kules, B., Drucker, S., and schraefel, m.c. (2006). Supporting exploratory search: Introduction. *Communications of the ACM*, 49(4): 36-39.

Building Models of Search Success with Experience Sampling and Event Logs

Andy Edmonds
One Microsoft Way
Redmond, WA 98052

aedmonds@microsoft.com

ABSTRACT

Exploratory search tasks stack additional challenges on the already difficult task of evaluating searching effectiveness. The experience sampling method has been used at MSN Search to assess the relationship between individual result and overall session satisfaction. Combining experience sampling with robust modeling of implicit variables has been shown to yield a productive model for general search. Methods for extending this approach to new systems designed to support exploratory search are proposed.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: Evaluation/Methodology

General Terms

Measurement, Human Factors.

Keywords

Exploratory search, experience sampling, event logging.

1. INTRODUCTION

With tens of millions of unique queries a day, understanding user success on a commercial search engine is a notable challenge. For the subset of searches that are exploratory, the difficulty increases. Fox, et al. [1] describe an experience sampling methodology [2] that models user feedback at critical points during the search experience to generate predictions from instrumented browser logs. They generate a model capable of predicting end-of-session ratings of overall satisfaction in addition to per result quality. This discussion will focus on the methodology for capturing this data stream and how it might be adapted to support evaluations of systems for exploratory search which support various aspects of the process.

Copyright is held by the author/owner(s).

SIGIR '06 Workshop, August 10, 2006, Seattle, Washington, USA.

2. MODELING USER SATISFACTION

2.1 Experience Sampling & Explicit Feedback

In the work reported in Fox, et al. (2006), 146 people participated over a 6-week period with an instrumented browser that captured events on MSN Search and Google. Explicit feedback was collected when a user left a search result site by returning to the result set with the back button, or via a discontinuous navigation like navigating from favorites or issuing a new query for a portion of user search clicks. Figure 1 shows a popup that collected the result evaluation.

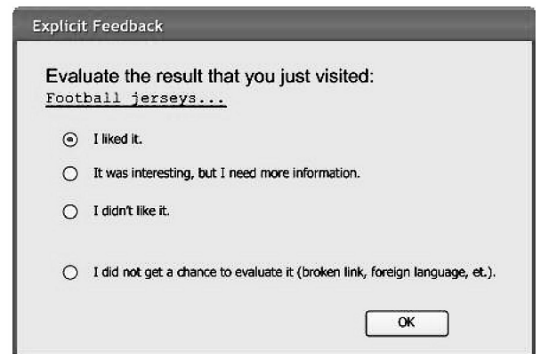


Figure 1 – Per Result Feedback

When the user transitioned from one search, a survey assessed continuation or new topic from the user. In the case of a new topic, or when the user ended the search session by explicit browser UI navigation or timeout, a survey captured feedback on session satisfaction. In some cases a timeout triggered the session end survey if the user continued to use the web but stopped using search. Figure 2 shows the post-session feedback popup.

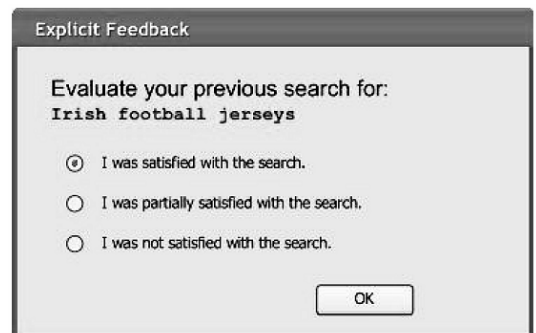


Figure 2 – Session Feedback

Recent work by Kellar & Watters [3] used an experience sampling method to model task changes and validated the use of browser UI events to indicate task change.

2.2 Predicting Session Satisfaction

Session satisfaction yielded a higher average value than result satisfaction, 59% versus 37%. Using per result satisfaction along with instrumented variables, Fox et al. were able to predict the outcome of the post-session survey at 70% accuracy, with higher accuracy given the ability to sub-select sessions for which the model generated high confidence predictions. The key instrumented values in the model included number of actions and session end operation. Individual result satisfaction was also highly predictable, largely from duration of visit and exit action.

3. EVALUATING NEW EXPLORATORY SEARCH SYSTEMS

The range of potential user interface and information retrieval operations that have and could be developed to support exploratory search is tremendous. A large history of work exists in supporting query refinement [4], browse by example and faceted browsing [5], and a variety of mechanisms for supporting memory of content engaged and revisitation [6].

3.1 Integrating Exploratory Search Features

Models of log data developed with explicit feedback are certain to be useful for enhancements to core information retrieval algorithms. User interface developments may contribute unique sources of user value above the retrieval of content to the user's current interest. In general, these improvements should lead to a greater number of engagements with quality content and the relationship between individual visit and overall session satisfaction should remain. However, reducing the cognitive effort in generating query refinements or in revisiting previous content may offer information synthesis and learning opportunities which go beyond individual result level satisfaction.

Introducing implicit variables which measure the level of engagement with features like query suggestions into a predictor model offers an opportunity to quantify the contribution of these features to overall user satisfaction. New task models and interaction designs for search do not make the methods described here intractable, but will require independent explicit feedback to create custom models.

Content familiarity, the learnability of content, availability of content in the search system, and the cognitive overhead of using the search system all contribute to end user success. In lab studies with controlled content, subject matter experts may be able to craft more sensitive content based assessments of the level of learning acquired or quality of the results selected. This approach is effortful and the generalizability of the findings less certain than in more naturalistic, user driven examples of exploratory search. Modeling based approaches may successfully capture the latter system quality and user workload attributes.

3.2 Targeting in the Task Model

One of the challenges in naturalistic usage scenarios is identifying user intent, and in this specific problem space, distinguishing informational and navigational from exploratory searches. Kellar, et al. [7] used a task bar with a categorical intent selection. The search result and session survey intercepts target the subtask and task-end states of exploratory search, while Kellar's approach captures task starts

Also noteworthy is that users were asked to monitor changes in task and generate feedback without prompting. Avoiding interruption is highly desirable if users have good insight into the events of interest and are motivated enough to be diligent in providing feedback at the right points.

4. CONCLUSIONS

Experience sampling involves choosing strategic points across a user experience to collect explicit end-user feedback. Modeling user state, as reported from explicit feedback, with event log data can yield models which have predictive value across other, uninterrupted user experiences. Combining implicit and explicit data sources can exceed the sum of the parts, complementing rich user intent and satisfaction data from smaller samples of explicit data with voluminous quantities of event data to find the most predictive event log patterns. In addition, these methods can quantify the contribution of user actions, and correspondingly system features, to overall satisfaction.

5. REFERENCES

- [1] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais and T. White (2005). Evaluating implicit measures to improve the search experience. *ACM:TOIS*, 23(2), 147-168.
- [2] Barrett, L. F. and Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Soc. Sci. Comput. Rev.* 19, 2 (Jun. 2001), 175-185.
- [3] Kellar, M., and Watters, C. (2006). [Using Web Browser Interactions to Predict Task](#). In the Proceedings of WWW 2006, Edinburgh, Scotland. 843-844. [[Poster \(.pdf\)](#)]
- [4] H. Daume, E. Brill (2004). [Web Search Intent Induction via Automatic Query Reformulation](#) 2004 *Proceedings of HLT 2004*
- [5] Elliott, A. 2001. Flamenco image browser: using metadata to improve image search during architectural design. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington, March 31).
- [6] Mayer, M. & Bederson, B. B.: Browsing Icons: A Task-Based Approach for a Visual Web History, HCIL-2001-19, CS-TR-4308, UMIACS-TR-2001-85, HCI Lab, University of Maryland, Maryland, USA, 2001.
- [7] Kellar, M., Watters, C., and Shepherd, M. (2006). The Impact of Task on the Usage of Web Browser Navigation Mechanisms. In the Proceedings of Graphics Interface 2006, Quebec City, Canada. 235-242.

Retrieval Evaluation sans Human Relevance Judgments

Thomas K Landauer
University of Colorado, Boulder and
Pearson Knowledge Technologies
4940 Pearl East Circle
Boulder CO, 80301
1-303-545-9092
Landauer@PearsonKT.com

ABSTRACT

A novel automatic retrieval evaluation method was developed for LSA-based cross-language retrieval of similar documents. Its chief properties of interest are: (a) it requires neither creation of example queries nor human relevance judgments, (b) It produces a quantitative distribution of expected performance over all possible queries within a corpus of data, not only ones of special interest to a pre-selected set of users. It thus can provide rapid feedback for iterations during IR system development and post hoc evaluations of operational systems. Variants of the approach should be useful for single language systems as well.

General Terms

Measurement, experimentation, theory, verification

Keywords

Search, retrieval, evaluation, testing, LSA, cross-language, multilingual, relevance

1. INTRODUCTION

For research, development and acceptance of new information retrieval methods or systems it necessary to test how well they fulfill their purpose of making it easy to find information. The standard method of evaluation from the beginning of automatic information retrieval has been to get potential users to provide queries or needs statements, and then to get human judgments of the relevance of returned information. It is widely recognized that this methodology has limitations. Obtaining trial users and queries is slow and expensive, as are exhaustive relevance judgments. Queries and need statements obtained from experimental users may not be representative of those that will occur in actual use. Neither are relevance judgments, whose reliability is less than perfect and themselves slow and expensive to evaluate. The methods are too cumbersome by far for optimum iterative design if new queries or relevance judgments need to be made. Most pertinent to this paper, they will rarely if ever test how well the system will perform across all possible queries.

The novel performance evaluation method described here avoids all of these limitations. However, its ability to do so at present can only be asserted with confidence for a special case, that of given one document finding other documents about the same topic. We will return to a brief discussion of this matter in section 3.

2. THE NEW APPROACH

In the LSA-based cross-language system, the similarity of two paragraphs (say an English and a Swahili newspaper story) is measured by their LSA cosine in a common semantic space. For evaluating the system's overall performance, we first find a baseline value for best possible cross-language retrieval for the database in question by obtaining a representative set of ca. 500 expertly translated paragraphs, which we call "mates", and calculating their cosines. We then compute the cosines between every paragraph in L1 and every paragraph in L2, that is, all non-mates. Finally, we compute statistics and plot the empirical distributions of cosines for mates and non-mates.

Ideal performance would separate the distributions for mates and non-mates perfectly. The area of overlap relative to the area for mates gives a single number for the system's ability to tell an average pair of translated (or merely comparable, if one wished) documents from other documents in the collection. By smoothing the plots we can visualize the manner in which the system spreads document similarities from most relevant to least relevant, and by numerical integration we can obtain its average precision, the average probability of finding a highly relevant document in the top n returns, and so forth. We can construct a precision/recall curve for any or all possible searches for similar cross-language documents in the collection and for all pairs of such documents believed to come from a source with the same distribution, for example a database of weather reports for Seattle and New Delhi on every day in 2000-2006 rather than for a random sample of 100 days.

To generalize the evaluation approach to single-language retrieval, it is necessary only to redefine what is meant by a "mate", the best possible degree of similarity for the system in question given the way it measures similarity. In most cases, within a single language collection, a query document will always have itself as the most similar. Therefore, the "mate" of a document is defined as itself. The quality of the overall system is then measured and described by how well it separates one document from all others. The same analytic properties accrue in the single-language case as in the cross-language.

Many readers will note that his approach closely resembles the signal-detection analysis long championed by John Swets [1], but

largely rejected by the IR community. The rejection has apparently been primarily on the grounds that given the customary way of evaluating accuracy the Swets method would require invocation of a theoretical, e.g. normal, distribution of similarities (despite Swets showing this not to be true.) The present method is not vulnerable to this criticism because the distributions involve are empirical distributions from the target corpus statistics, not theoretical.

3. APPLICABILITY

As mentioned at the beginning, the method has only been applied to finding similar documents within a collection of documents using LSA. How well will it work with other types of search engines? What about complex ad hoc need statements and the very common case of queries of a few words or phrases from outside the evaluated system? The straightforward answer at this time is that the answer is unclear. Are systems that are better at measuring the similarity of paragraphs always comparably better at measuring the similarity of other kinds of queries to paragraphs or to each other? It does not seem utterly unlikely, but it seems probable that the quantitative distributions they produce will be different. Further empirical exploration is needed to find out for what purposes the method is and is not how useful.

Nevertheless, is important to note that the case for which the method was developed and reported here, retrieval of similar documents across languages, is not the only one of its kind. The potential applicability of document-to-document retrieval is quite broad. For example, it appears widely in “like these” functions in search engines. We have used it in a system that cross-indexes every paragraph in a library so that a user can select any one, part of one or concatenated combination to use as a query to find the top *n* others with the nearest overall meaning as measured by LSA. And surely readers of online textbooks, and even digital novels, would profit from the ability to simply select a text segment and be taken to others that are topically related.

3. REFERENCE

- [1] Swets, J. A. Effectiveness of Information Retrieval Methods, *American Documentation*, 20,1 (1993), 72-99.

Why a corpus-topics-relevance judgments framework isn't enough: two simple retrieval challenges from the field

Catherine C. Marshall
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
cathymar@microsoft.com

ABSTRACT

In this paper, I use two challenges to illustrate how retrieval tasks can fall outside the current corpus-topics-relevance judgment evaluation framework. The two challenges that span both desktop search and standard information retrieval are: (1) encountering new information or re-encountering forgotten information and (2) retrieval of the appropriate version of semi-redundant information, either from a personal information space or from a public store of datasets. These challenges probe the assumptions underlying corpus construction, topic selection, and relevance judgment by suggesting some common activities violate them.

1. INTRODUCTION

Information retrieval researchers have developed a strong sense of community and an equally strong sense of the value of specific research contributions by focusing on a core set of metrics, standard evaluation methods, and reference corpora, queries, and relevance judgments. It is easy to measure progress against this backdrop – how different retrieval algorithms trade off against one another, whether tuning a particular algorithm produces better results, and how new strategies for filtering and retrieval measure up to old ones. Thusfar, this agreed-upon evaluation backdrop has paid off: the community has made significant progress on a variety of information retrieval problems, and by applying analogous evaluation techniques, has been able to assess progress in new areas (such as question answering) and retrieval methods that address new media types (for example, video).

We would thus expect challenges to the efficacy of these evaluation metrics and methods to arise when a search activity or corpus characteristics are at odds with the community's assumptions about information and how it is used. Of course, the fluidity, variability, and distribution of the information on the Web and the enormous range of information needs of the people who search it – not to mention the adversarial nature of Web search – pose a striking challenge to information retrieval business as usual; these challenges have been reflected in the addition of new TREC tracks as well as in changes to the strategies used by commercial search engines. But it is still helpful to identify specific types of situations in which the evaluation criteria do not quite apply.

In this workshop paper I focus on two such challenges, both of which seem to be characteristic of self-managed information spaces in addition to external information spaces such as corpora,

digital libraries, or the Web. One such challenge arises from retrieving material from what we might think of as “messy” information spaces, informal collections with semi-redundant items, such as multiple versions of the same document, perhaps with changes or revisions. The other such challenge arises from opportunistic information behavior such as clipping, in which people save or share items they have encountered in venues like magazines, newspapers, on the Web, or on broadcast media. Both of these challenges represent relatively ubiquitous everyday situations as people interact with information. Table 1 summarizes these challenges.

Table 1. Two overarching challenges

Challenges	desktop search	standard IR
<i>Encountering new or forgotten information</i>	Re-encounter of forgotten information in the searcher's file system or in the searcher's personal information space	Encounter of interesting, useful, or sharable information on the Web or in an archive or online publication
<i>Retrieval of appropriate version of semi-redundant items</i>	Locating the appropriate copy of an edited item from a personal digital store, given the searcher's information needs and expectations	Locating the appropriate copy of a revised item, given the searcher's information needs and the item's provenance

2. VERSIONS & DISTRIBUTION

Much information retrieval evaluation to-date assumes a “clean” information space that has been curated to remove items deemed to be duplicates. What happens when the information space is messy? At first blush, messiness doesn't seem like that much of a problem; there are many techniques to remove duplicate items from information spaces and it's easy to factor this kind of redundancy into evaluation. But what if the expected use of the item is central to the relevance judgment of which copy is the right one?

Field studies reveal that cleaning a collection to remove duplicates is not always as simple as it may seem; there's no straightforward heuristic that distinguishes among seemingly equivalent items. A recent field study of personal digital archiving practices revealed that consumers make copies of files as a hedge against storage catastrophes and accidental deletions [7]. This practice is not formal, but rather people welcome the opportunity to create additional versions of valued digital items. Table 2 shows the trajectory of an informant's photo of herself, one that she was very fond of. Normal use has resulted in 12 versions of a single

original; the photo is now in two different formats and the jpegs are in at least two resolutions. The file also has four different names and is stored on six file systems. (I use a photo because it's a real example; this could just as easily been a text document.)

Table 2. Tracking 12 versions of a single original photo

Description of photo file	Filename
Original on camera flash memory	126-2162_IMG.jpg
File copy on old desktop hard drive	126-2162_IMG.jpg
File copy edited in Photoshop	Eden20.psd
File copy in "sent" mail (sent to art partner who maintains web site)	Eden20.psd
File copy uploaded to web site (converted to jpeg and resolution reduced)	Eden20.jpg
File copies written to CD (as hard drive backup)	Eden20.psd & 126-2162.jpg
File copies restored from CD to new PC hard drive	Eden20.psd & 126-2162.jpg
File copy downloaded from website because psd files won't open	EB.jpg
File re-edited in photo-editing application	EB-4U.jpg
File in "sent" mail (emailed to "boys")	EB-4U.jpg

Suppose we're evaluating a very clever desktop image search algorithm. One information need she expressed (by browsing the filesystem, just to complicate matters) was to find this photo so she could attach it to an email message and send it to a prospective boyfriend she met via Match.com. Which copy of the photo counts as the right response to her query? How about the copy that's the appropriate resolution to send in an email message? How about the one that was edited in Photoshop and is now stored offline? How about the one that's the original version downloaded from the camera? Surely the nature of the task, the distribution of the data, and subtleties of the replication process should play into the presentation and evaluation of the results.

Interestingly, emerging e-science collections (especially those arising out of "little" science) yield similar types of examples [1]. For example, consider a situation in which a scientist curates a central collection of datasets. These datasets include contributions of comparable local datasets from other scientists worldwide. But each dataset is downloaded from the central site and used in many ways and many copies have been made along the way; gaps in the data are filled using different conventions and the data have been cleaned relative to different uses. For some uses, portions of the dataset are irrelevant. In other copies, measurements have been removed because the scientist using the data believes these measurements to be erroneous ("It's never 80 degrees in Greenland! This sensor must be collecting inaccurate values."). Which version is the right one? Without downloading all of them, the searcher can only tell through the use of visualization tools that run on the server side.

3. ENCOUNTERING

When we develop evaluation methods to assess algorithms, at the most basic level we assume that someone is looking for something, or – even if they are not – that they are engaged in some activity that would benefit from additional relevant

information, as they would in Implicit Query scenarios [3]. But an important component of our interaction with various types of media – newspapers, broadcasts, magazines, even conversations with our friends and colleagues – is encounter [4]. Serendipitous encounter with information is apt to spur exploration, discovery, and creativity. People also use encountered information socially: they share encountered material for a variety of reasons, and although the material that they share need not be central to a current activity, it does need to have connections that are meaningful to both the sender and the recipient [6].

Recently there have been a number of efforts to explore peoples' need to re-find items they have sought in the past or encountered serendipitously (e.g. [2]); however, this is only part of the problem. As we look into the longer term relationships that people have with information – their personal archives – we find that people may not search for these things again because they don't remember that they have them [6]. This holds especially true for encountered information, because it was saved outside of an information-needs context. Yet in several different studies, when our participants re-encountered certain kinds of particularly evocative information – things they'd saved to remind them of a place or an event in their lives – they appeared to derive great pleasure from coming upon these things again. It is difficult to think of these things as falling within current information retrieval evaluation paradigms.

These challenges are not intended to suggest that the IR community abandon the current mode of competitive evaluation that conforms to an established pattern of corpus-topics-relevance judgments. Instead these challenges – and indeed the proliferation of tracks and corpora at TREC – highlight a need to examine the assumptions that underlie the evaluation strategy. Because the cases described here are seen: (1) as outside the information retrieval rubric (e.g. information that is saved without a need/encounter); (2) as human failings (e.g. forgetting what is saved in personal archives/re-encounter); (3) as information sloppiness (e.g. uncontrolled replication of personal digital data/choice among similar copies); or (4) as part of an invisible process (e.g. scientific data cleaning/redundant datasets), and found with a human in the loop [5], they have a tendency to fade from sight. Yet they are all important – and very real – examples of how people claim and reclaim information.

4. REFERENCES

- [1] Borgman, C., Wallis, J., Enyedy, N. Building Digital Libraries for Scientific Data. to appear *Proc. ECDL 2006*.
- [2] Bruce, H., Jones, W., Dumais, S. Information behaviour that keeps found things found. *Info. Research*, 10, 1, 2004.
- [3] Cutrell, E., Dumais, S., and Teevan, J. Searching to Eliminate Personal Info. Management. *CACM*, 49, 1, 58-64.
- [4] Erdelez, S. Information Encountering: A conceptual framework. In *Proc. Information Needs, Seeking, and Use in Different Contexts*. Taylor Graham, 1997, 412-421.
- [5] Marchionini, G. Toward Human-Computer Information Retrieval. *ASIST Bulletin*, 22, 5, June/July 2006, 20-24.
- [6] Marshall, C.C. and Bly, S. Saving and Using Encountered Information. *Proc. CHI'05*, 111-120.
- [7] Marshall, C.C., Bly, S., and Brun-Cottan, F. The Long Term Fate of Our Personal Digital Belongings. *Proc. Archiving 2006* (Ottawa, Canada, May 23-26, 2006) 25-30.

Analysis of the Task Environment of Sense Making

Peter Pirolli
Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
pirolli@parc.com

ABSTRACT

Exploratory search technologies aim to provide users with means that go beyond current query-based search engines, to provide users with improved ways of understanding the topical and navigational landscape of available content, and to provide improved ways of making sense of that content to achieve users' goals. A measurement framework and theory remains to be specified for exploratory search. This paper sketches such a framework derived from the cognitive sciences, centered on the notion of task environments. A cognitive task analysis of expert intelligence analysis is presented as a concrete and complex form of exploratory search.

Categories and Subject Descriptors

H.5 [Information systems and presentation]: Human computer interaction

General Terms

Measurement, Human Factors.

Keywords

Exploratory search, information foraging, sensemaking.

1. INTRODUCTION

Exploratory search technologies aim to provide users with means that go beyond current query-based search engines, to provide users with improved ways of understanding the topical and navigational landscape of available content, and to provide improved ways of making sense of that content to achieve users' goals. A recent survey of such technologies [16], including browsing, clustering, and information visualization techniques, pointed out the need for ways of evaluation and measurement that go beyond the standard techniques employed with search engines (e.g., precision and recall). In this position statement I will attempt to point towards a path that may lead towards the achievement of a measurement framework and theory for exploratory search. The framework derives from the cognitive sciences, and was utilized in my own research on information foraging theory [10]. I present a cognitive task analysis (CTA) of expert intelligence analysis [4], which may be considered a

concrete and complex form of exploratory search [6]. This CTA provides suggestions for the shape of the analytic framework and for some things to measure within the framework.

2. A GOAL FOR MEASUREMENT: FITNESS IN TASK ENVIRONMENTS

I assume that the term "exploratory search" refers to tasks that drive human behavior for minutes, hours, days, and even years; i.e., longer than the 10 sec unit task level [2]. Typically, these tasks will be ill-structured problems, such as choosing a medical treatment or buying a house. These require additional knowledge from external sources in order to better understand the starting state of affairs, to better define a goal, or to specify the actions that are afforded at any given state [15]. In modern society, people interact with information technology that more or less helps them find and use the right knowledge at the right time. Increasing the rate at which people can find, make sense of, and use valuable information improves the human capacity to behave intelligently; increasing the rate of gain of valuable information increases fitness.

To measure such improvements in fitness requires analysis of users coupled to their information environments in the context of *task environments*. The classical definition of the task environment is that it "refers to an environment coupled with a goal, problem or task—the one for which the motivation of the subject is assumed. It is the task that defines a point of view about the environment, and that, in fact allows an environment to be delimited" [8, p. 55]. The task environment is the scientist's analysis of those aspects of the physical, social, virtual, and cognitive environments that drive human behavior. The information environment is a tributary of knowledge that permits people to more adaptively engage their task environments. From the standpoint of a psychological analysis, the information environment is delimited and defined in relation to the task environment. As argued elsewhere [13], the concept of task environment provides a way of developing true theories of measurement that assess the degree to which a person (and the technology they use) achieves perfect rational use of the knowledge accessible to them. To simplify somewhat, this involves analysis of (a) *what* goal or problem is being solved under what constraints, (b) *why* some solutions are more rational than others (or which are optimal), and (c) *how* solutions are actually achieved by users coupled to their technology. Even if we were interested in latent learning that occurs while exploring without a specific goal, the effects of that learning can best be measured in some task environment.

Copyright is held by the author/owner(s).

SIGIR'06 Workshop, August 10, 2006, Seattle, Washington, USA.

3. FORAGING AND SENSEMAKING AS DUAL SPACE SEARCH

The User Interface Research Area at PARC has been studying a broad class of tasks we call sensemaking [14], which is much like the class of investigative search discussed by Marchionini [6]. Such tasks involve finding and collecting information from large information collections, organizing and understanding that information, and producing some product, such as a briefing or actionable decision. One way to understand the structure of task environments and principles of psychological adaptation is to study extreme expert performance. Towards those ends, we have been studying experts in intelligence analysis

The initial phase of this research [12] involves cognitive task analysis (CTA: knowledge elicitation techniques derived from applied psychology that yield information about the knowledge, thought process, and goal structures that underlie observable task behavior [3]. Our studies of intelligence experts suggests that the overall process is organized into two major loops of activities: (1) a *foraging loop* that involves processes aimed at seeking information, searching and filtering it, and reading and extracting information [11], and (2) a *sense making loop* [14] that involves iterative development of a mental model (a conceptualization) that best fits the evidence. Information processing can be driven by *bottom-up* processes (from data to theory) or *top-down* (from theory to data). Top-down and bottom-up processes are invoked in an opportunistic mix. We also see many examples of *preparation-deliberation tradeoffs* [7] in which effort may be devoted to accumulating knowledge (e.g., through learning, or by developing caches of external ready-to-use sources) that may be rapidly used when new tasks come up (typically with deadlines).

The foraging loop is a tradeoff among three kinds of processes: exploration, enrichment, and exploitation (e.g., reading). These processes tradeoff against one another under deadline or data overload constraints. Typically, analysts cannot explore all of the space, and must forego coverage (recall) in order to actually enrich and exploit the information. We have found it useful to measure the effects of these tradeoffs by focusing on the overall *rate of gain of information value* [11], and learning effects such as *mental category formation* [9], and *foraging plans* [1].

The sensemaking loop involves problem structuring (the generation, exploration, and management of hypotheses), evidentiary reasoning (marshalling evidence to support or disconfirm hypotheses), and decision making (choosing a prediction or course of action from the set of alternatives). These processes are affected by many well-known cognitive limitations and biases, including limited span of attention and confirmation bias. The entire foraging + sensemaking process can be viewed as a variation of *dual space search* [5] observed in scientific reasoning. In dual space search, there is a problem space search process aimed at collecting evidence that will be relevant to testing a hypothesis, and a problem space search process around the generation of hypotheses. Techniques used in the study of dual space search can be applied to sensemaking.

4. ACKNOWLEDGMENTS

Portions of this research have been supported by Advanced Research and Development Activity, Novel Intelligence from

Massive Data Program Contract No. MDA904-03-C-0404 to S.K. Card and Peter Pirolli.

5. REFERENCES

1. Bhavnani, S.K., Domain-specific search strategies for the effective retrieval of healthcare and shopping information. in CHI 2002 Conference on Human Factors and Computing Systems, Extended Abstracts, (Minneapolis, MN, 2002), ACM Press, 610-611.
2. Card, S.K., Moran, T.P. and Newell, A. The psychology of human-computer interaction. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.
3. Chipman, S.F., Schraagen, J.M. and Shalin, V.L. Introduction to cognitive task analysis. in Schraagen, J.M., Chipman, S.F. and Shalin, V.L. eds. Cognitive task analysis, Lawrence Erlbaum Associates, Mahwah, NJ, 2000, 3-23.
4. Gersh, J., Lewis, B., Montemayor, J., Piatko, C. and Turner, R. Supporting insight-based information exploration in intelligence analysis. Communications of the ACM, 49 (4). 63-68.
5. Klahr, D. Exploring science: The cognition and development of discovery processes. Bradford Books, Cambridge, MA, 2000.
6. Marchionini, G. Exploratory search: from finding to understanding. Communications of the ACM, 49 (4). 41-46.
7. Newell, A. Unified theories of cognition. Harvard University Press, Cambridge, MA, 1990.
8. Newell, A. and Simon, H.A. Human problem solving. Prentice Hall, Englewood Cliffs, NJ, 1972.
9. Pirolli, P. The InfoCLASS model: Conceptual richness and inter-person conceptual consensus about information collections. Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 11 (3). 197-213.
10. Pirolli, P. Information foraging: A theory of adaptive interaction with information. Oxford University Press, Cambridge, UK, in press.
11. Pirolli, P. and Card, S.K. Information foraging. Psychological Review, 106. 643-675.
12. Pirolli, P., Lee, T. and Card, S.K. Leverage points for analyst technology identified through cognitive task analysis, PARC, Palo Alto, CA, 2004.
13. Pirolli, P. and Wilson, M. A theory of the measurement of knowledge content, access, and learning. Psychological Review, 105. 58-82.
14. Russell, D.M., Stefik, M.J., Pirolli, P. and Card, S.K., The cost structure of sensemaking. in INTERCHI '93 Conference on Human Factors in Computing Systems, (Amsterdam, 1993), Association for Computing Machinery, 269-276.
15. Simon, H.A. The structure of ill-structured problems. Artificial Intelligence, 4. 181-204.
16. White, R.W., Kules, B., Drucker, S.M. and schraefel, m.c. Supporting exploratory search: Introduction. Communications of the ACM, 49 (4). 36-39.

Layered Evaluation of Adaptive Search

Peter Brusilovsky
University of Pittsburgh
School of Information Sciences
Pittsburgh, PA, 15260 USA
peterb@pitt.edu

Rosta Farzan
University of Pittsburgh
Intelligent Systems Program
Pittsburgh, PA, 15260 USA
rosta@cs.pitt.edu

Jae-wook Ahn
University of Pittsburgh
School of Information Sciences
Pittsburgh, PA, 15260 USA
jaa38@pitt.edu

ABSTRACT

The goal of this paper is to discuss how adaptive search systems which embed exploratory options should be evaluated. We argue that a state-of-the-art evaluation of adaptive search systems should follow a “layered evaluation” approach. To support and explain this argument we describe how the layered approach was applied to the evaluation of the adaptive search component of Knowledge Sea II, a system that is powered by a social navigation support mechanism.

Categories and Subject Descriptors

H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness)

General Terms

Measurement, Design, Human Factors

Keywords

Social search, adaptive systems, exploratory search systems, layered evaluation

1. INTRODUCTION

The growing need for effective organization and maintenance of the increasing number of Web-based educational resources motivated us to construct a personalized information access system, Knowledge Sea II (KSII). KSII provides various types of information access methods, including two-level visualizations (a knowledge map plus a similarity-based visualization), hypertext browsing, recommendation, and social search. Personalization for all these access methods is provided by social navigation (SN) support [1], [5]. SN is a relatively well-known personalization approach for browsing-based and recommendation-based information access; however, its use for search personalization has been almost unexplored.

The adaptive search component of KSII combines a traditional vector search engine with SN support, allowing every user to benefit from the collective wisdom of the whole community. To stress it we will refer to it as “social search.” The results of the

search are adapted to the user by taking into account both the past interactions of the individual user and the user’s group. The SN support of KSII includes various information access methods that allow the user to do exploratory searching. She can start the exploration by browsing or by entering the map, then use her newly acquired knowledge about the domain’s terminology to choose better query terms. She can also modify her initial query after consulting SN information provided by the system. The main goal of this paper is to discuss how adaptive search systems with this exploratory nature should be evaluated, using KSII search as a model. We argue that state-of-the-art evaluation of adaptive search systems should follow a “layered evaluation” approach that is an active focus of research in the area of user-adaptive systems [2]. The core idea behind layered evaluation is that specific sub-components or layers of any user-adaptive system should be understood and evaluated independently. Layered evaluation can overcome shortcomings of the conventional methodologies, which try to test the adaptation process as a whole and can miss success or failure of critical sub-components. In our approach to layered evaluation, we divided the adaptation process into two parts: decision-making and interface adaptation and then evaluated each of them. In this paper, the nature of our adaptive social search system is presented (section 2) and our layered evaluation framework is discussed (section 3). The paper concludes in section 4 with a summary and brief discussion of the future direction of our research.

2. SOCIAL SEARCH IN KSII

Social navigation (SN) in KSII incorporates several information access methods, including social search. SN support is offered through by visually marking links with icons and color codes. Figure 1 shows an example of search results that have been annotated with SN cues. Standard information about each document in the list is given—such as rank (7), document source (Univ. of Leicester), title (Pointers), and a similarity score (0.4057)—while traffic- and annotation-based SN cues are on the right. The foreground and background colors of the human icon depict user and group traffic, associated with time spent reading this document [4]. The darker the color is, the higher the traffic. The background color of the annotation represents annotation density. The foreground icons represent the type and overall status of the annotation. For example, a “thumbs-up” icon represents positive individual annotation while the warm temperature shown on the “thermometer” represents positive group annotation. For example, the document “Pointers” shown on Figure 1 is ranked 8th in terms of its similarity score to the user query but is very popular among the community of the users. Thus the user might want to examine the contents of this document, despite its relatively low

Copyright is held by the author/owner(s).

SIGIR’06 Workshop, August 10, 2006, Seattle, Washington, USA.

score, to learn how to improve her query terms for the next stage of her exploration.

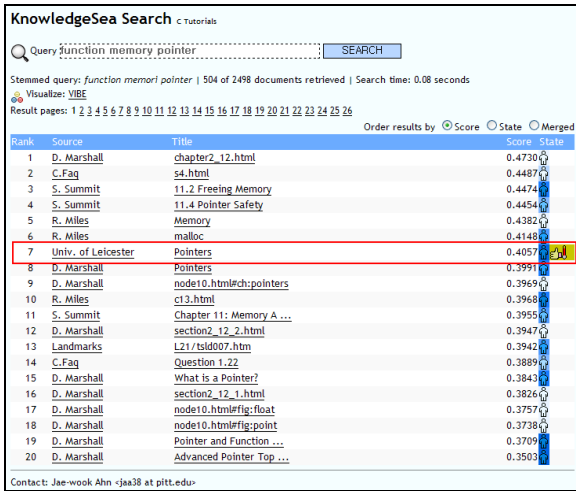


Figure 1 Social search with social navigation cues

3. LAYERED EVALUATION OF SOCIAL SEARCH

The need for the layered evaluation framework arose from the insight that conventional evaluation methods cannot pinpoint the effectiveness of critical layers of the adaptation process, which perform different tasks contributing to the final results. Current practices frequently attempt to evaluate adaptation as a whole by comparing the whole adaptive application to a baseline, an equivalent, non-adaptive application. However, even if the results turned out to be better than the baseline's, we cannot hastily conclude that all of its components perform well. Vice versa, if the adaptive system as a whole is lower than the baseline's, there is still the possibility that one of its layers was actually successful [3]. To address this problem, several authors have introduced layered evaluation frameworks. Brusilovsky et al defined user-modeling and adaptation evaluation layers in [2]. Weibelzahl introduced a 4-layer approach: the reliability and validity of input data, interface, adaptation decision, and the interaction [6].

To evaluate social search in KSII we adopted a 2-layer approach which considers the decision-making and adaptation layers separately. Based on the interaction history of the user's social group, the decision-making layer decides which pages should be useful and to what extent. The adaptation layer decides how to express to each user this calculation of the social importance of a specific page. In the current version of KSII, this layer generates icon-based annotations, as shown in Figure 1. However, this is only one possible way to express the social importance of documents.

3.1 Decision Making Layer

The goal of the decision-making layer is to predict how useful each document is to a user of a specific group. KSII uses two independent decision-making layers, based on traffic and annotation. Since the latter is rather straightforward, we focused on evaluating the traffic-based one. To argue that the traffic-based prediction works we needed to demonstrate that documents predicted as useful (those shown with darker blue backgrounds by

the adaptation layer) are really useful. Our gold standard for rating the importance of pages is that students find them good and important. Therefore, we focused on pages with student annotation.

For evaluation, we computed the normalized access rate for pages with and without annotation. As can be seen in Figure 2, "good and important" pages are accessed twice as often. Thus page traffic average is a good indicator of page quality. These pages will have a generally darker background, according to our traffic-based SN support algorithm.

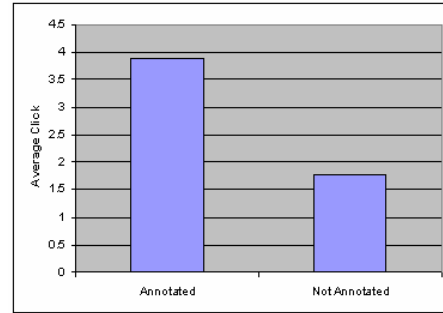


Figure 2 - Average click number over pages with and without annotations

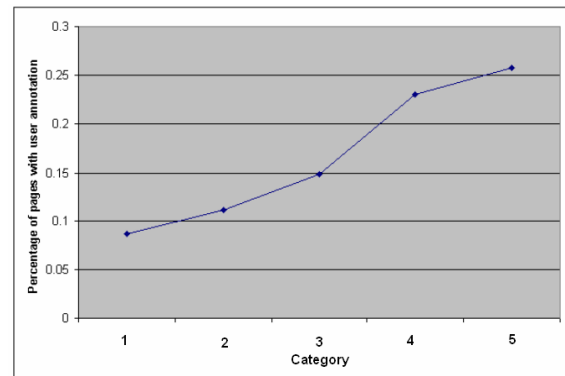


Figure 3 - Percentage of pages with user annotation for different levels of usage

To enhance the evaluation, we categorized accessed documents in five categories, based on the time spent on each page. The following table shows the details of this classification.

Category	Average Time Spent	Darkness of Background	Level of Recommendation
1	< 65 sec	No background	None
2	< 97 sec	Light blue	Slightly
3	< 152 sec	Blue	Recommended
4	< 217 sec	Dark blue	Considerably
5	> 282 sec	Very dark blue	Highly

For each category we computed the percentage of pages that were annotated by the students. To exclude the dependency of annotation and visit, we excluded annotations made by users of the target semesters while including annotations made by users of past and future semesters. As shown in figure 3, the pages with darker

backgrounds (higher usage) have a higher percentage of annotation. This data shows that important pages are being predicted as useful by our SN adaptation which means the important pages are augmented with darker background.

3.2 Evaluation of social search with social navigation cues

Once we established the positive correlation between quality and SN, it was important to evaluate the effect of SN cues. The goal of the cues is to attract user attention to socially important documents and to encourage them to examine them. In our context, we needed to evaluate how much the SN cues affect students' decision to choose links within search results. Moreover, since KSII social search separates the visualization of query relevance (document position in the search list) from visualization of social importance (intensity of background color in SN cues), we were interested in comparing the influence of positive SN cues to the influence of being a top ranking in the list.

To evaluate this layer, we decided to compare the *effective* and *random relative access rates* for links with high rankings (top of the list) and links with traffic-based cues. The *random relative access rate* tells which fraction of clicks would have been made if the user randomly selected specific links in the search results list. Basically, it shows how often the links with this property appear in the search results list. The *effective relative access rate* reports the actual proportion of target quality links, compared to total accessed links. If the effective relative access rate is higher than random, it means that the links with this quality successfully encourage users to access them.

The first question to answer is: "Do students prefer links with better rankings?" (considering the first three documents in the search results list to be *top ranked*). Since every results page shows 20 links, the random relative access rate for the top three ranked documents is $3/20 = 0.15$. Effectively, students accessed 53 documents from different search results lists, with 16 being top ranked. Therefore the effective relative access rate was $16/53 = 0.3$, which is twice the random (0.15). This is evidence that the students do take the document rank into account, preferring links on the top of the list.

The second question to answer is: "Do students prefer links with traffic-based SN cues?" To answer this question, we attempted to separately evaluate links with any visible past traffic (number of past clicks >1) from links with higher traffic (past clicks >2). The reason is that the links with two past click were annotated with a very light blue color, which, we afraid, some users might ignore. The links with 3 and more past clicks were annotated with reasonably dark blue color and were hard to ignore.

Computing the random relative access rate for links with group traffic was a complicated procedure. For each of the 53 cases of link access we had to re-create the group traffic accumulated at the time of access to understand how many social-cued links the user saw when making the selection. For each case, we calculated this rate by dividing the number of visible links with the target level of traffic by the total number of links. Then, we averaged the probabilities over all 53 cases and found that for pages with visible traffic the random relative access rate is equal to 0.08. Out of 53 cases, students choose 17 documents from the visible traffic category. Therefore the effective relative access rate for links with visible traffic is $17/53=0.32$, which is four times higher than the

random access rate (0.08). A similar ratio (0.05 to 0.19) was obtained for links with high traffic. This result shows that students do prefer links with visible group traffic. Moreover, the ratio of effective access rate to random is twice as high for pages with visible traffic than for pages with top rankings. This provides evidence that pages marked by visible group traffic do influence students. Moreover, the presence of "group traffic" gives the page an even higher chance to be visited.

4. CONCLUSIONS

In this study, we demonstrated how a 2-layered evaluation framework could be used for evaluating an adaptive search interface which enables exploratory searching by users. We divided the evaluation process into decision-making and adaptation layers, in order to better understand the effectiveness of each sub-component process. We were able to show a correlation between the predicted and effective social utility of a page (i.e., pages automatically predicted as important for the group by the decision-making component were actually rated as important by students). We also provided evidence that the specific interface adaptation approach used in KSII to attract the user's attention to socially important pages does influence user behavior in the expected direction. The proposed evaluation framework should be able to evolve by adopting more layers, such as user-to-system interaction and input data validation. In future research, we are planning to use the same layered framework to evaluate other kinds of adaptive information access methods, including information visualization.

5. REFERENCES

- [1] Brusilovsky, P., Chavan, G., and Farzan, R (2004). Social adaptive navigation support for open corpus electronic textbooks. In Proc. of Adaptive Hypermedia and Adaptive Web-Based Systems.
- [2] Brusilovsky, P., Karagiannidis, C., and Sampson, D. Layered evaluation of adaptive learning systems. *Int. J. Cont. Engineering Education and Lifelong Learning*, 14, 4, 2004, 402-420.
- [3] Brusilovsky, P., Karagiannidis, C., and Sampson, D. (2001). The Benefits of Layered Evaluation of Adaptive Applications and Services. In *Proceedings of Workshop on Empirical Evaluation of Adaptive Systems*.
- [4] Farzan, R. and Brusilovsky, P. (2005) Social navigation support in E-Learning: What are real footprints? In Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization.
- [5] Farzan, R. and Brusilovsky, P. (2005) Social navigation support through annotation-based group modeling. In Proc of User Modeling.
- [6] Weibelzahl, S., and Weber, G (2001). Advantages, Opportunities, and Limits of Empirical Evaluations: Evaluating Adaptive Systems. *Kunstliche Intelligenz*.
- [7] Lowrence, S., and Giles, C.L. Context and Page Analysis for Improved Web Search (1998). *IEEE Internet Computing*.

Wrapper: An Application for Evaluating Exploratory Searching Outside of the Lab

Bernard J. Jansen
College of Information Sciences
and Technology
The Pennsylvania State University
University Park PA 16802
jjansen@ist.psu.edu

Raghavan Ramadoss
Department of Computer Science
and Engineering
College of Engineering
The Pennsylvania State University
University Park PA 16802
ramadoss@cse.psu.edu

Mimi Zhang and Nan Zang
College of Information Sciences
and Technology
The Pennsylvania State University
University Park PA 16802
mzhang@ist.psu.edu,
nzz101@psu.edu

ABSTRACT

In this position paper, we assert that a focus on evaluating individual exploratory searching systems misses a critical aspect of assessing the exploratory searching process. Namely, that in complex information environments, searchers use multiple systems over an extended period marked by specific episodes of interaction with online systems. We argue that the focus of the evaluation should be on the process, not a single system. However, evaluating an exploratory searching process can be a difficult task to conduct in a naturalistic setting (i.e., outside of a laboratory). In response, we have developed a client-server application for use in the study and evaluation of exploratory searching processes. We describe the application and demonstrate the ability of the application in a pilot study. The results from our evaluation show that exploratory searching is indeed a chaotic process, demonstrated by the use of multiple information systems and repeated episodes of searching. The implications are that by using this tool one can successfully evaluate exploratory searching processes. Assessment of the entire process rather than a single exploratory searching system could significantly further the advancement of system design for this critical searching context.

Categories and Subject Descriptors:

H.3.3 [1] Information Search and Retrieval – relevance feedback.

General Terms:

Performance, Design, Experimentation, Human Factors

Additional Key Words and Phrases:

Implicit user feedback, exploratory search evaluation

1. INTRODUCTION

In exploratory search, the situational context in which the user performs individual searching episodes is critically important in the evaluation of the overall process. The user's searching episode (i.e., a distinct period of interaction with an online system) may

involve multiple queries related only at some high-level of information abstraction. There may be several searching episodes in close temporal proximity, or a considerable temporal span may separate the searching episodes. Additionally, these searching episodes may occur on multiple searching systems.

In such a complex situational environment, the evaluation of a single Exploratory Search System (ESS) could miss crucial elements of the user context, since the searching process may not occur on one ESS. In fact, our view is that once you have confined exploratory search to a single system, you have over simplified the problem. Our position is that the evaluation should not focus on a particular ESS but on the Exploratory Searching Process (ESP), which can span multiple searching episodes, multiple systems, and varied temporal spaces. However, evaluating ESPs has been nearly impossible or at least too costly in terms of effort due in part to the lack of automated methods of collecting ESP data.

In this paper, we present the Wrapper, a client-server application for the use in evaluating ESPs. The Wrapper is based on technology we developed for user evaluations on information retrieval (IR) systems [4]. We describe the Wrapper's design and value in terms of evaluating ESPs, and we show its value by discussing the results of a pilot study where we employed the Wrapper. With the Wrapper or similar client-server applications, one can conducted naturalistic studies of the entire ESP and not be limited to the study of a single ESS. We believe that such an approach provides much more realistic insight into the users' tasks, goals and behaviors.

Sections 2 present a brief literature review and the research objectives. We then discuss the structure of the proposed application, and research results to date in sections 3 and 4 respectively. Section 5 provides the concluding remarks, along with future aims and implications of the Wrapper for ESP evaluation.

2. LITERATURE REVIEW

Although there are open questions, the evaluation of a single ESS (or any single searching system) is relatively straightforward compared to evaluating ESPs. One can point to the series of Text REtrieval Conferences as an example. There are also good commercial applications for single system evaluation in labs, such as Morae 1.1¹. However, the evaluation of an ESP is much more difficult because the central actor is not the system but the user.

Copyright is held by the author/owner(s).

SIGIR'06 Workshop, August 10, 2006, Seattle, Washington, USA.

¹ <http://www.techsmith.com/products/morae/default.asp>

During an ESP, the user may access multiple systems. The search topic is difficult, and the period of searching is longer. All of these factors point to the need for naturalistic (i.e., outside of a laboratory) studies of the entire process rather than a system. However, there has not previously been an application to facilitate data collection and delivery available to the research community.

Researchers have relied on a variety of methods for data collection. Hancock-Beaulieu, Robertson, and Nielsen [2] used server-side transaction logs and online questionnaires. In their naturalistic and longitudinal study of professionals and their information seeking patterns, Choo, Betlor, and Turnbull [1] developed their own logging software but had to physically collection the logs. Kelly [5] used a spy software package and a proxy server. Spy software has inherent disadvantages including granularity of data capture, and privacy concerns. A proxy sever is limited to logging traffic only on one network. Toms, Freund, and Li [7] developed a system for conducting large-scale evaluations. However, the entire study must occur within the WiRE framework and is limited to one server.

To address the need for an application to study ESPs, we developed the Wrapper, a client-side application to collect and gather user data. The application is coded in Visual Basic 6, is easy to install, collects a wide range of user-systems interactions, and is not limited to a single server.

In the following sections, we present a description of the features and output of the application. We aim to provide this version of the Wrapper to the research community via a uniform resource locator (URL). Interested researchers can download the application from the URL for use in their research projects and studies.

3. DEVELOPMENT METHODOLOGY

One of the essential aims in developing the Wrapper is to facilitate the collection and gathering of data over an extended period and over multiple information systems.

3.1 Structure of the Application

The software consists of (1) the client-side module and (2) the server-side module. The client and server modules of the Wrapper communicate using sockets. The client module sends a string message containing the user details to the server in the form of a plain text document. This text file contains the computer's Internet Protocol (IP), time-stamped, implicit feedback action, and object of that action [3]. Implicit feedback interactions, including copy, bookmark, print, save, etc., indicate relevance.

A daemon, capable of simultaneous reception of files from multiple clients, runs at the server end and waits for incoming files from the clients. The server dumps the received file locally for further analysis on the collected data.

3.2 The Client Wrapper Application

The client-side module is a self-installing executable that can be downloaded and installed over the Internet. The executable is

generated from the Visual Basic programming environment. One can activate the application manually, via a batch file, or from a browser toolbar. The application has a Window interface (Figure 1) for real time observation that can be hidden to allow for unobtrusive monitoring. The application logs interactions with the IR system, along with other applications, using Dynamic Data Exchange (DDE). Output is to a text file, with a specifiable location and an automatically generated unique filename. Additionally, the client module also sends this output directly to the server-side module for data collection.

Referring to Figure 1, we numbered each of the functional aspects of the application, which we describe below.

1. Log filename (generated automatically using date and time)
2. Running text of log file.
3. List of all processing running.
4. The current value of the clipboard.
5. Text to be appended to log file.
6. Current system time.
7. Title and URL of current page.
8. Running list of URLs.

The dialog box in Figure 1 can be set to hide during studies, so the participant will never see it. An example of the application output is:

```
20:58:21 Clipboard Use
20:58:21 https://mail.ist.psu.edu/exchange/ View URL
20:58:43 http://search.yahoo.com/search?ei=utf-8&fr=slv1-
&p=successive+searching View URL
20:59:49 http://search.yahoo.com/search?ei=UTF-
8&p=successive+searching View URL
21:00:07 http://search.yahoo.com/search?ei=UTF-
8&p=successive+searching View URL
21:00:21
http://edc.techleaders.org/LNT99/presentations/05_Thu/strat
egies.htm View URL
21:00:40 View URL
21:00:40 Bookmark URL
```

In its current version, the application logs a wide range of user interactions, include interactions with the browser tool bars, interactions with the system clipboard, scrolling of results listing or documents, and numerous implicit feedback actions [6], such as bookmark, copy, print, save, and scroll. The user activates the Wrapper prior to performing a search and when the Web browser closes the Wrapper automatically terminates. Figure 2 illustrates how the Wrapper integrates with the browser and computer operating system.

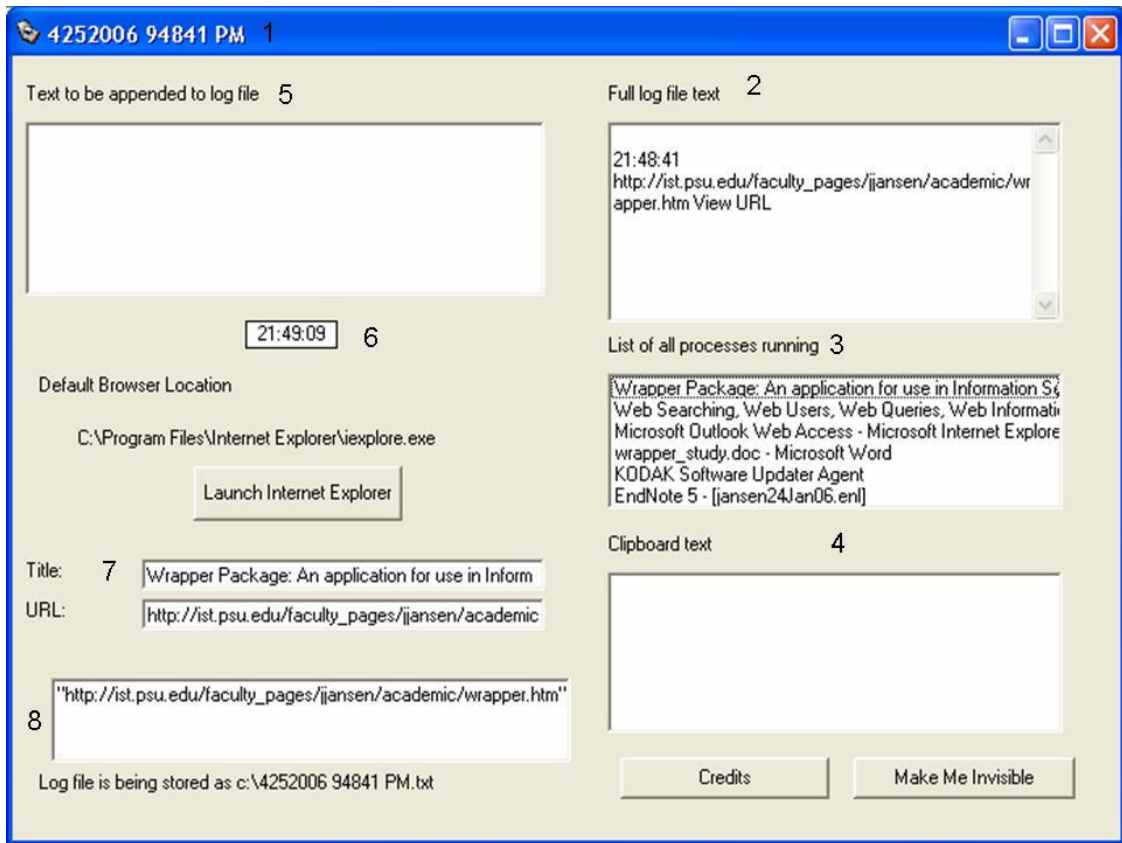


Figure 1. Visible Version of the Client-side Module of the Wrapper.

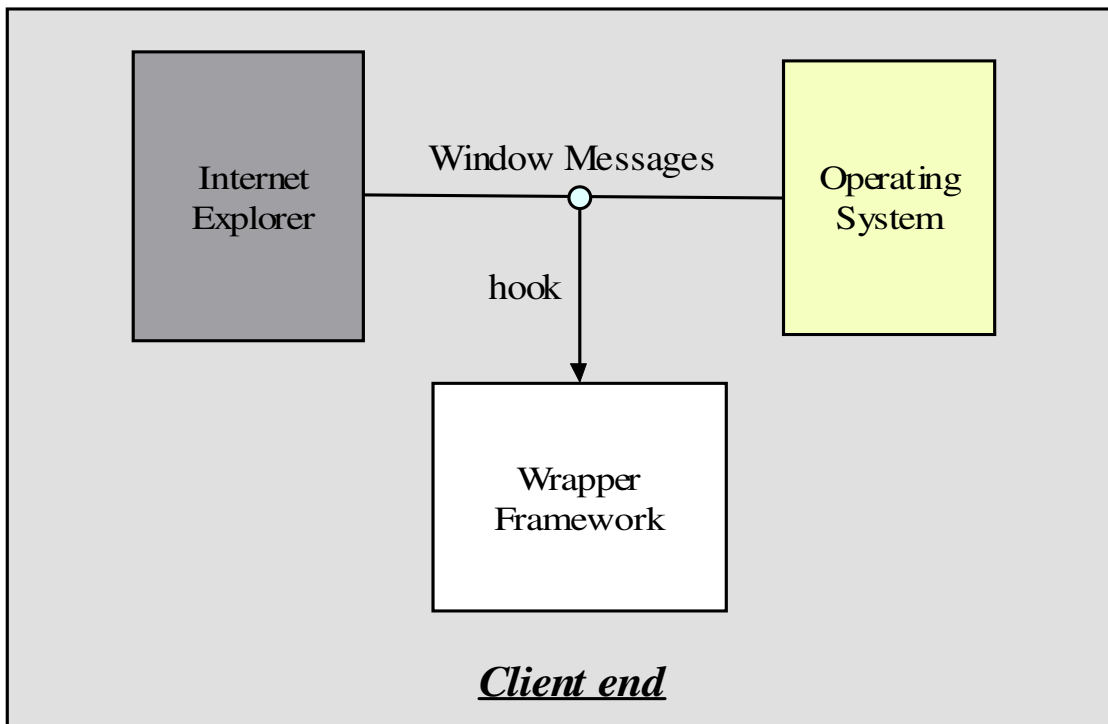


Figure 2. The Client-side Module Interfacing with the Browser and Search Engine.

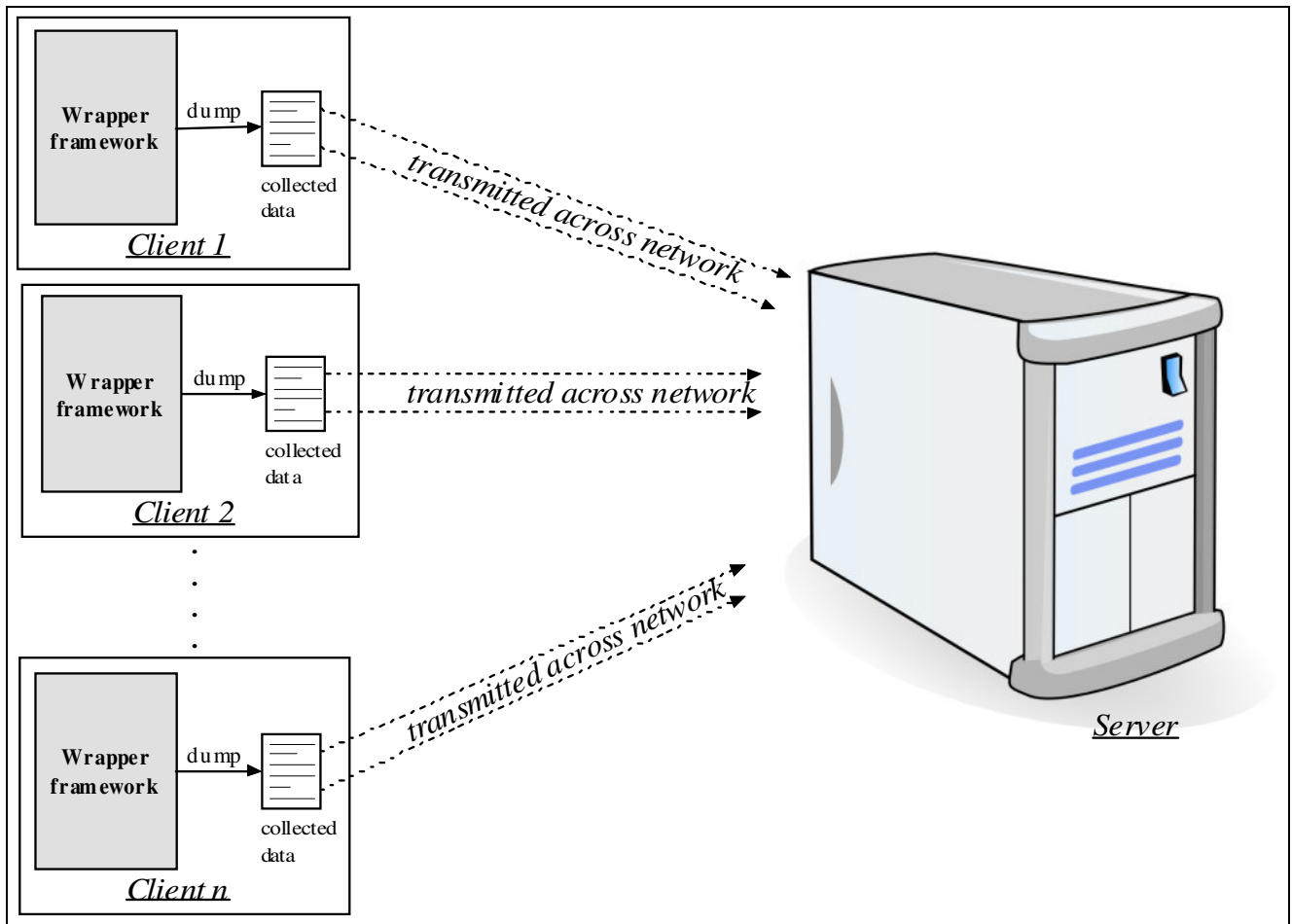


Figure 3. The Server-side Module Listening for Messages from the Client-side Module.

3.3 The Server Wrapper Application

The server-side module collects the information sent from one or more client-side applications, storing each information set into a file. The server-side module creates and names the file using the IP address of the computer on which the client-side module is running. The server-side module uses sockets to receive the messages from the client-side modules. The server-side application writes the received information values to the appropriate file at five-second intervals. These intervals can be adjusted with changing requirements by altering the server code. Figure 3 illustrates the communication among the server module and client modules.

4. WRAPPER EVALUATION

4.1 User Evaluation of Performance Tool

In order to gauge the effectiveness of the Wrapper, we conducted a pilot test of the application with 4 participants in order to determine the effectiveness of the software in light of users' differing searching experiences and searching needs. We conducted the study over a one-week period (i.e., 7 continuous 24-hour periods). The participants conducted their searching as they normally would, using whatever systems they deemed appropriate. The only additionally needed action was to activate

the Wrapper via a button on the browser toolbar at the start of each searching episode. The client-side module collected the data and sent it to the server-side module. The study did not tie the participants to any network, location, computer, or searching system.

4.2 Results from User Evaluation of Performance Tool

Aggregate results of the 7-day pilot study are presented in Table 1. As we can see from Table 1, searching is a disorganized task and does not conform to the logical sequence of events that one so often sees in the scenario approaches used in lab studies of IR systems. Instead, searchers employ a variety of searching and information sources and return to topics over multiple days. Users exhibited behaviors on certain topics that one would classify as being part of an ESP. In these ESP, users visited multiple search engines and searched on multiple Web sites. Some of these Web sites were directly off the search engine results listing. Others were browsed to from the search engine results or bookmarked Web sites.

Table 1. Aggregate Statistics Collected by the Wrapper During Pilot Study

User	Computers Used	Searching Episodes	Information Topics By Episode	Systems Used	Episode Duration
1	3	18	<ol style="list-style-type: none"> 1. Medical, 2. Entertainment, 3. Parenting, 4. Topic Research A, 5. Topic Research B, 6. Technology, 7. Technology, 8. Entertainment, 9. Topic Research, 10. Topic Research A, 11. Topic Research A, 12. Religion, 13. Topic Research A, 14. Topic Research C, 15. Ecommerce (Housing), 16. Ecommerce (Housing), 17. Ecommerce (Housing), 18. Ecommerce (Housing) 	5	min: 1 minute max: 31 minutes
2	2	4	<ol style="list-style-type: none"> 1. Ecommerce, 2. Hobby, 3. Technology, 4. Ecommerce 	12	min: 2 minute max: 147 minutes
3	1	4	<ol style="list-style-type: none"> 1. Topic Research D, 2. Topic Research E, 3. Topic Research D and Sports, 4. Topic Research D 	4	min: 1 minute max: 28 minutes
4	1	21	<ol style="list-style-type: none"> 1. Ecommerce, 2. Technology, 3. Topic Research F, 4. Topic Research G, 5. Ecommerce, 6. Topic Research H, 7. Topic Research I, 8. ECommerce, 9. Entertainment, 10. Topic Research J, 11. Topic Research K 12. Topic Research L 13. Work Requirement 14. People Search 15. Topic Research M 16. Topic Research N 	23	min: 2 minute max: 72 minutes

Table 1. Aggregate Statistics Collected by the Wrapper During Pilot Study

User	Computers Used	Searching Episodes	Information Topics By Episode	Systems Used	Episode Duration
			17. News, 18. Ecommerce, Technology, History, Art, ecommerce 19. Topic Research O, 20. Topic Research P, 21. Hobby		

During ESPs, users also employed multiple queries with few query terms in common, but the queries were related at a higher information abstraction. However, searchers who engaged in ESPs conducted their searching over multiple days.

5. CONCLUSION

The Wrapper is an open source application for use during ESP studies. It addresses a fundamental issue in exploratory search evaluation in that user may seek information over an extended period and on multiple information systems. The Wrapper collects and gathers, at a central location, the typical interactions of searchers from the client-side, thereby permitting studies of ESPs over extended durations and not limited to any one ESS. Therefore, the Wrapper directly supports the development of metrics to evaluate ESS performance and provides a focus on the searcher during evaluations. In future research, we aim to increase the number of user interactions the application logs.

6. REFERENCES

- [1] Choo, C. and Turnbull, D., Information Seeking on the Web: An Integrated Model of Browsing and Searching, *First Monday*, vol. 5, http://firstmonday.org/issues/issue5_2/choo/index.html, 2000.
- [2] Hancock-Beaulieu, M., Robertson, S., and Nielsen, C., Evaluation of online catalogues: an assessment of methods (BL Research Paper 78), The British Library Research and Development Department, London 1990.
- [3] Jansen, B. J., Designing Automated Help Using Searcher System Dialogues, in *Proceedings of the 2003 IEEE International Conference on Systems, Man & Cybernetics*, 2003. Washington, D.C., USA. 5-8 October. pp. 10 - 16.
- [4] Jansen, B. J., Search log analysis: What is it; what's been done; how to do it, *Library and Information Science Research*, forthcoming.
- [5] Kelly, D., Understanding Implicit Feedback and Document Preference: A Naturalistic User Study, Rutgers, The State University of New Jersey, New Brunswick January 2004.
- [6] Oard, D. and Kim, J., Modeling Information Content Using Observable Behavior, in *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, 2001. Washington, D.C., USA. 31 October - 4 November. pp. 38-45.
- [7] Toms, E. G., Freund, L., and Li, C., WiIRE: the Web interactive information retrieval experimentation system prototype, *Information Processing & Management*, vol. 40, pp. 655-675, 2004.

Exploratory search is facilitated by the user entering in a query and viewing the nodes and links in relation to the query node. If the user double-clicks on one of the nodes, then additional related items are retrieved and graphed around the selected node. The node is marked with a green “c” if there are no further links. Users can obtain additional information regarding an item by moving the mouse over a node and clicking on an “info button”. This brings up a small pop-up window with textual information possibly containing additional links that the user can navigate.

The focus of this research was to examine how topic level knowledge affects users browsing the display and making similarity-based selections made from the visualization display. Users’ selections were then compared to the system generated similarity selections as shown through the advanced radius feature. Interestingly, participants rated their topic knowledge as quite low for most tasks, however a high degree of participant-system item selection overlap was observed. There was a statistically significant relationship found between knowledge level and node use for half of the tasks and these tasks had a less cluttered visualization representing a more hub and spoke model than the others. The importance of these findings is that the visualization display aided user item selection on unfamiliar topics in a way which showed similar patterns to system generated results.

Exploratory search visualization for document retrieval can capitalize not only on the arrangement of items on the screen, but also the information conveyed within the document icons themselves. VIBE (Visual Information Browsing Environment) is a visualization system developed by researchers at Molde College, Norway and the University of Pittsburgh’s School of Information Sciences (Figure 2). Exploratory search is accommodated by the set of features VIBE contains to examine relationships among the data set. For example, the color feature applied to keywords allows the user to see the resulting intersecting set in red. Users may add, remove, or drag keywords to see their impact on the document icons on the screen.

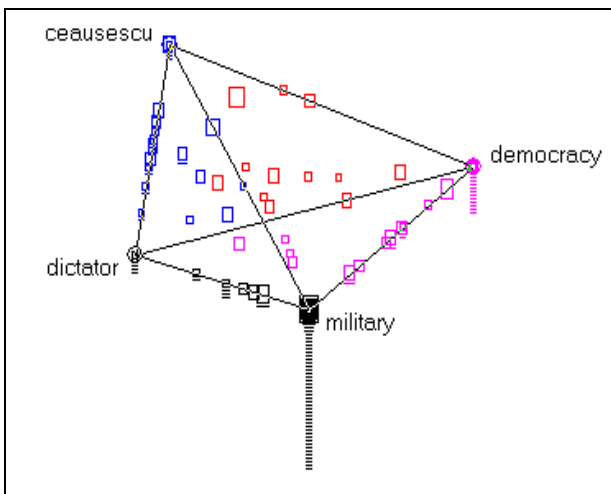


Figure 2. VIBE display.

Koshman [5] found that in examining user interaction with VIBE, visual tasks were solved more successfully than standard information retrieval tasks. Participants were asked to identify the largest document icon in the display and answer a question related

to the full text of the document retrieved. The use of pop-out icons against a set of distractor items can be used as visual cues generated by the system to aid user navigation. Repeated exposure saw an increase in the magnitude of correct tasks and task timings improved. These findings pointed toward VIBE being a learnable system. Other tasks drew upon VIBE’s interactive features such as “net”, “star”, and “lines” to aggregate, displace, and connect document icons to explore the resulting search set. One avenue for further research is that it may be possible to relieve the uncertainty surrounding exploratory search with interactive features which give the user more control over the manipulation of icons in a visualization display.

The notion of interactivity with the system’s display can contribute to the overall sense of user satisfaction with a visualization system for exploratory search. Thirty-two participants were tested with the web-based Missing Pieces tool by InfoSpace, Inc. [7]. Missing Pieces was developed to visualize the overlap of search engine results in comparison to the search results generated by the metasearch engine, Dogpile (Figure 3). The assumption behind the overlap study design is that web searchers typically consult only the first page of web search engine output to make selections.



Figure 3. Missing Pieces visualization.

The majority of participants were able to successfully identify results generated by the three search engines and Dogpile. The highest percentage of participants indicated that their primary criterion for selecting the most useful results was the URLs, followed by the proximity of items to the center of the display, paw print icons, and color. Overall, subjective satisfaction measures rated the visualization positively, however a frequently cited issue in the open-ended responses was the lack of interactive links in clicking on URLs in the display. While further investigation is warranted, this result points toward an obstacle in completing an exploratory search task since additional information from the actual web page could not be obtained easily by the user.

These studies and others offer a range of methodologies and tasks that extend user research with visualization systems. To enhance our understanding of using visualization tools for exploratory search, more consideration needs to be given to what type of

factors affect user interaction and which evaluative metrics can be implemented for testing users with these systems.

3. EVALUATION

The preceding examples demonstrate the application of visualization tools to facilitate exploratory search. These user studies and others are based on a foundation of information visualization user testing research. Investigating user interaction with the mechanics of various visualization systems has gained momentum, however the evaluation of visualization systems that can support exploratory search tasks is not well understood. Standard practice in information visualization user testing is to develop tasks that are not derived from the user in order to establish control in measuring factors such as task timings, task errors, and familiarity time [14]. Tasks are designed to accommodate the type of visualization that is implemented and taxonomies for visual tasks are used to structure the study design [8, 13]. Participant groups tend to be small and subjective satisfaction measures can play a critical role in determining the viability of a visualization system [6].

A visualization system's display and document glyphs add another layer of complexity when exploring search results. Part of a visualization system's appeal lies in the use of perception to analyze search results sets, however another aspect is the cognitive load imposed on the user who is required to interpret the document icons. While visualization systems appear to offer an intuitive solution to exploratory search, furthering our understanding of the learning curve and the mental models that users require is the key to user evaluation. Based on the research conducted and a review of user study literature, the following methodological factors and testing metrics need to be addressed:

1. A combination of field and laboratory testing may be conducted. Lab testing is the primary model for user research and this approach may be supplemented by observing user interaction in naturalistic environments such as library settings. User demographics will be more diverse and a broader spectrum of user interaction may be analyzed.

2. Longitudinal testing may be conducted along with briefer lab-based sessions. Typically, users do not have pre-established mental models of visualization systems in comparison to standard text-based search interfaces. Plaisant [10] supports long term studies as part of evaluating visualization technology. Repeated sessions may enhance knowledge levels and ease of use with the system and contribute to effective exploratory search.

3. Task design can incorporate investigator and user-derived tasks to determine if the effectiveness of the visualization system for exploratory search may be affected by task structure.

4. Standard metrics such as task timings, familiarity time, and task error rates can be used with increasing emphasis on the user to determine the completion of an exploratory search task. Successive and repeated searches may need to be taken into account for an exploratory search session on the web.

5. Subjective satisfaction measures play a significant role in understanding the users' perceptions of visualization systems for exploratory search. Factors affecting subjective satisfaction for visualization systems are discussed in detail by Koshman [6] and include:

- a. Training length and type (e.g. instructor led vs. training video.)

- b. Task type, timing, and purpose. Does the task facilitate exploration through visually identifiable features such as icon size or color? Is the completion of an exploratory search task defined by investigator task time limits or by the user? For example, Rivadeneira and Bederson [11] reported a ten minute task time limit for assigning factual information retrieval tasks with Grokker and Vivisimo interfaces.

- c. Number of visualization system features. Is the system operational or a prototype? How many of the features are being tested and how many do the users need to learn? One alternative approach de-features the visualization system in order to test the user interpretation of result presentation [9].

- d. Type of visualization (e.g. treemap, node-link diagram).

- e. The level of difficulty associated with decoding icons. Do the icons use arbitrary or sensory representations?

- f. Emotive factors in assessing how the system makes the user feel (e.g. confused, confident).

- g. User types (e.g. domain knowledge experts, system experts, novices) as found in [15, 12, 5].

- h. Comparative evaluation. Does the evaluation include a text-based system as a baseline for comparing the visualization system? Using a text-based system for testing has implications for task order and study design.

- i. Response type. Fixed vs. open-ended responses for post-task questionnaires.

- j. Procedures. Is subjective satisfaction measured after each task or after each session?

- k. Speed of system. Is there a perceptible lag time to generate the visualization display independent of the user's equipment?

Some of these factors have been applied in previous information visualization user studies. Other factors can be operationalized and used to improve user testing as well as lay the foundation for standardizing measures to evaluate visualization systems for exploratory search. Further discussion may reveal additional factors that will extend this list.

4. CONCLUSIONS

Visualization systems offer potential in supporting exploratory search and their role in becoming operational systems for web information retrieval and digital libraries will be dependent upon extensive user evaluation that identifies exploratory search as a definable and measurable goal. Fox et al. [3] recently cited visualization as a major component for exploration. Future research includes examining each of the factors presented in the previous section and designing studies to investigate their impact on user interaction and evaluation of the system. This will facilitate the growth of visualization from a novelty technology to a user-oriented pragmatic tool for exploratory search.

5. REFERENCES

- [1] Card, S. K., Mackinlay, J.D. & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann Publishers.
- [2] Chen, C., Chen, Y., & Maulitz, R. C. (2005). *Understanding the evolution of NSAID: A knowledge domain visualization approach to evidence-based medicine*. Paper presented at the 9th International Conference on Information Visualization (IV 2005), London, England, pp. 945-952
- [3] Fox, E. A., Das Neves, F., Yu, X., Shen, R., Kim, S., & Fan, W. (2006). Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM*, 49(4), 53-58.
- [4] Koshman, S. (2004). Web-based visualization interface testing: similarity judgments. *Journal of Web Engineering*, 3(3/4), 281-296.
- [5] Koshman, S. (2005). Testing user interaction with a prototype visualization-based information retrieval system. *Journal of the American Society for Information Science and Technology*, 56(8), 824-833.
- [6] Koshman, S. (Forthcoming). Exploring subjective satisfaction for information visualization evaluation.
- [7] Koshman, S., Spink, A., Jansen, B. J., Blakely, C., & Weber, J. (June 1-3, 2006). *Metasearch result visualization: an exploratory study*. Paper accepted at the Canadian Association for Information Science Conference, York University, Toronto, Ontario.
- [8] Morse, E., Lewis, M., & Olsen, K. A. (2000). Evaluating visualizations: using a taxonomic guide. *International Journal Human-Computer Studies*, 53, 637-662.
- [9] Morse, E., Lewis, M., & Olsen, K. (2002). Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical and "spring" displays. *Journal of the American Society for Information Science and Technology*, 53(1), 28-40.
- [10] Plaisant, C. (2004). The challenge of information visualization evaluation, *Advanced Visual Interfaces* (pp. 109-116). Gallipoli, Italy: ACM Press.
- [11] Rivadeneira, W., & Bederson, B. (2003). *A study of search result clustering interfaces: comparing textual and zoomable user interfaces*. Retrieved April 20, 2004, from <ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2003-36html/2003-36.htm>
- [12] Sebrechts, M. M., Vasilakis, J., Miller, M. S., Cugini, J. V., & Laskowski, S. (1999). Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces, *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3-10). Berkeley, CA: ACM Press.
- [13] Shneiderman, B. (2003). The eyes have it: a task by data type taxonomy for information visualizations. In B. Bedersen & B. Shneiderman (Eds.), *The Craft of Information Visualization: Readings and Reflections* (pp. 364-371). San Francisco: Morgan Kaufmann.
- [14] Shneiderman, B., & Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (4 ed.). Reading, Mass: Addison-Wesley Longman.
- [15] Swan, R. C., & Allan, J. (1998). Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems, *SIGIR '98, Conference on Research and Development in Information Retrieval* (pp. 171-181). Melbourne, Australia: ACM Press

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR 2006 workshop, Evaluating Exploratory Search Systems Seattle, USA
Copyright 2006 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

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-

¹Note 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 browser². 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 systems³. 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 ad-hoc 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 (2005) 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

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

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

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 discussion 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

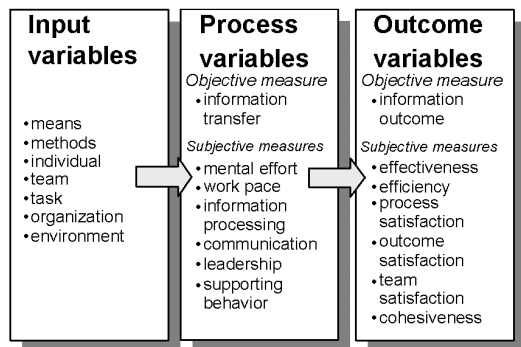


Figure 1: Meeting evaluation framework

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 [9], 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.

Methods for evaluating changes in search tactics induced by exploratory search systems

Bill Kules

Takoma Software, Inc.
Takoma Park, MD 20912

wmk@takomasoftware.com

ABSTRACT

Mixed research methods provide a way to understand how exploratory search systems change search tactics and strategies. Longitudinal studies may help understand how searchers adapt search strategies and tactics in support of creative, challenging, individualized activities.

1. INTRODUCTION

Evaluation of exploratory search systems is an exciting research challenge. The situated nature of exploratory search tasks can lead to many different task outcomes for different searchers. This can make it difficult to specify objective performance measures like time to completion, error rates, precision, or recall. Completing an exploratory task often involves developing and refining an information need that is specific to the individual. Mistakes, dead-ends, and back-tracking are part of the process as searchers learn concepts and vocabulary. Documents that have great utility or novelty to one person may have little value to another, because of variations in domain knowledge, interests, and previously encountered information, so establishing ground truth for a measure of relevance is problematic.

The strategies and tactics that searchers use are affected by the capabilities provided by the search system [1, 2]. Designers build interfaces to support specific strategies, based on intuition or analysis. The effect of new capabilities on search tactics may not be what designers anticipate. Unexpected problems may negate expected benefits. Serendipitous possibilities may present to searchers. In response, searchers may adapt their tactics and strategies as they become familiar with the capabilities. Our research seeks to understand how exploratory search systems with rich user interfaces change the way that searchers think about and pursue their searches. What strategies and tactics do exploratory search interfaces enable? And, ultimately, do they enable searchers to achieve their higher-level objectives?

Task-based evaluation of exploratory search systems using controlled experiments has been effective for showing subjective satisfaction differences between systems, but less effective at showing objective differences in task performance, particularly in task outcomes. [4, 13]. Evaluations have assessed and rated the quality of a task outcome to generate quantitative measures on lesson plan creation task [4] or measured incidental learning that occurred during a search session [7]. Exploratory tasks have been decomposed or narrowed to constrain the task [3]. A combination

of quantitative and qualitative evaluation methods have also been used [10, 13].

Controlled experiments and in-depth case studies are two approaches to evaluation of exploratory search systems. This position paper describes two approaches to evaluation of exploratory search systems. A mixed method approach was used to evaluate categorized overviews of search results. The longitudinal approach is proposed to extend the mixed methods.

2. MIXED METHODS

My dissertation research investigated the use of categorized overviews of web search results based on meaningful and stable categories to support user exploration and understanding of large sets of search result [6]. Web search engines are effective at generating extensive lists of results that are highly relevant to user-provided query terms. For known-item queries, users often find the site they are looking for in the first page of results. However, a list may not suffice for more sophisticated exploratory tasks, such as learning about a new topic or surveying the literature of an unfamiliar field of research, or when information needs are imprecise or evolving [12]. When searchers need to gather information from multiple perspectives or sources, categorized overviews can organize results from web or digital library searches. Categorized overviews can help searchers explore alternative sources, assess utility of results, and decide on next steps. When searchers' information needs are evolving or imprecise, categorized overviews help by stimulating relevant ideas, provoking illuminating questions, and guiding searchers to useful information they might not otherwise find.

Research prototypes and commercial search engines have incorporated categorized overviews, but there have been few user studies of categorized overviews for exploratory web search, and there is little research explaining whether they are effective, why, and under what circumstances. Research is needed to understand how categorized overviews change the way users conduct web searches, to guide the design of search engine interfaces, and to justify the entry and maintenance of category metadata.

To study this, we adopted a mixed methods approach, using an experimental design that counterbalanced two interface conditions and collecting qualitative data for analysis. The task was described in the context of a journalistic scenario, and the 24 subjects were recruited primarily from journalism students. During the two-hour session, subjects were provided training and practice, and then they conducted four searches using a think-aloud protocol, ending with a 30 minute semi-structured interview. Screen video and audio were recorded, and interactions (queries, clicks, scrolling, mouse movement, etc.) were logged using a custom JavaScript-based tool. Based on previous research (ours and other studies), we expected to observe quantifiable and significant differences relative to several behavioral measures,

including how deep in the search results subjects explored. A qualitative approach extended the hypothesis tests by looking for phenomena not modeled by the research variables. For example, we expected that the categorized overview interface would prompt tactical and cognitive changes, but there was no *a priori* list; that was developed from the data.

The study identified seven tactics that searchers began to adopt when the categorized overview was available. It highlighted two important considerations for future evaluation of exploratory search systems. First, it takes time for searchers to reflect on their searches and refine their tactics. The semi-structured interview fostered in-depth reflection by subjects. Analysis of their responses complemented the analysis of two questionnaires [5], in which users of a clustering search tool answered two questionnaires administered 6 weeks apart, reporting differences in their search tactics. Second, the strategies and tactics employed by exploratory searchers are individualized and varied.

3. LONGITUDINAL STUDIES

Longitudinal studies may be useful for addressing these two considerations. Longitudinal studies have been used, for example, to examine changes in tactics and query terms in relation to changes in searchers' information problem stage while developing a research proposal [11]. In-depth, longitudinal case studies have been used to evaluate information visualization interfaces and creativity support tools [8, 9]. These techniques integrate ethnographic and quantitative methods, using participant observation, surveys, interviews, and usage logs to study users performing complex tasks with individually defined goals. These techniques may be beneficial for investigating how searchers adapt their tactics when rich web search interfaces like interactive categorized overviews are available. They present the opportunity to observe changes as searchers become familiar with an exploratory search system and tactics mature. They also present challenges, because search is often a means to an end, and individual searches may be initiated to satisfy a higher level task. The search sessions may not be readily organized into blocks of time that can be scheduled with a researcher. We are tackling these challenges as we undertake a study using this methodology.

4. CONCLUSION

Understanding how exploratory search systems change searcher tactics and strategies is a necessary step toward designing better systems. Mixed research methods provide a way to understand how exploratory search systems change search tactics and strategies. Longitudinal studies may help understand how searchers adapt search strategies and tactics in support of creative, challenging, individualized activities.

5. ACKNOWLEDGEMENTS

This research was partially supported by an AOL Fellowship in Human-Computer Interaction and National Science Foundation Digital Government Initiative grant (EIA 0129978).

6. REFERENCES

[1] Bates, M. (1990). Where should the person stop and the information search interface start. *Information Processing and Management*, 26 (5). 575-591.

[2] Golovchinsky, G. Queries? Links? Is there a difference? in *Proceedings of the SIGCHI Conference on Human Factors in*

Computing Systems, Atlanta, GA, ACM Press, New York, 1997, 407-414.

[3] Janecek, P. and Pu, P. (2005). An evaluation of semantic fisheye views for opportunistic search in an annotated image collection. *Journal of Digital Libraries*, 5 (1). 42-56.

[4] Kabel, S., Hoog, R.d., Wielinga, B.J. and Anjewierden, A. (2004). The added value of task and ontology-based markup for information retrieval. *Journal of the American Society for Information Science and Technology*, 55 (4). 348-362.

[5] Käki, M. Findex: search result categories help users when document ranking fails. in *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, Portland, OR, ACM Press, New York, 2005, 131-140.

[6] Kules, B. Supporting Exploratory Web Search with Meaningful and Stable Categorized Overviews (Unpublished doctoral dissertation), University of Maryland, College Park, 2006. Retrieved June 29, 2006, from <http://hcil.cs.umd.edu/trs/2006-14/2006-14.pdf>.

[7] Pirolli, P., Schank, P., Hearst, M. and Diehl, C. Scatter/gather browsing communicates the topic structure of a very large text collection. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, Vancouver, British Columbia, Canada, ACM Press, New York, 1996, 213-220.

[8] Shneiderman, B., Fischer, G., Czerwinski, M., Resnick, M., Myers, B., Candy, L., Edmonds, E., Eisenberg, M., Giacardi, E., Hewett, T., Jennings, P., Kules, B., Nakakoji, K., Nunamaker, J., Pausch, R., Selker, T., Sylvan, E. and Terry, M. (2006). Creativity support tools: Report from a U.S. National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction*, 20 (2). 61-77.

[9] Shneiderman, B. and Plaisant, C. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies *Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV '06): A Workshop of the AVI 2006 International Working Conference*, Venezia, Italy, 2006. Retrieved June 29, 2006, from <http://hcil.cs.umd.edu/trs/2006-12/2006-12.pdf>.

[10] Toms, E.G., Freund, L., Kopak, R. and Bartlett, J.C. (2003). The effect of task domain on search. *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research*. 303-312.

[11] Vakkari, P. eCognition and changes of search terms and tactics during task performance: A longitudinal case study. in *Proceedings of the RIAO 2000 Conference*, 2000.

[12] White, R.W., Kules, B., Drucker, S.M. and schraefel, m.c. (2006). Supporting exploratory search. *Communications of the ACM*, 49 (4). 36-39.

[13] Yee, K.-P., Swearingen, K., Li, K. and Hearst, M. Faceted metadata for image search and browsing. in *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, Ft. Lauderdale, FL, ACM Press, New York, 2003, 401-408.

An Integrated Approach to Interaction Modeling and Analysis for Exploratory Information Retrieval

Gheorghe Muresan

Rutgers University

School of Communication, Information and Library Science

4 Huntington St., New Brunswick, NJ 08901, USA

+1-732-932-7500-x8228

muresan@scils.rutgers.edu

ABSTRACT

In this paper, we describe a methodology that integrates the conceptual design of user interfaces with the analysis of interaction logs. It is based on formalizing, via UML state diagrams, the functionality that is supported by a system and the interactions that can take place, on deriving XML schemas for capturing the interactions in activity logs, and on deriving log parsers that reveal the system states and the state transitions that took place during the interaction. While this approach is rather general and can be applied in studying a variety of interactive systems, it was devised and subsequently applied in research work on exploratory information retrieval, where the focus is on studying the interaction and on finding interaction patterns.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces – *evaluation/methodology*.

General Terms

Design, Experimentation, Human Factors, Theory.

Keywords

Interaction design, log analysis, UML, XML.

1. INTRODUCTION AND MOTIVATION

While much of the research work in **Information Retrieval** has focused on the systemic approach of developing and evaluating models and algorithms for identifying documents relevant to a well-defined information need, there is increasing consensus that such work should be placed in an **Information Seeking** framework, in which a searcher's context, task, and personal characteristics and preferences should be taken into account (Ingwersen and Jarvelin, 2005).

Since Robertson and Hancock-Beaulieu (1992) described the cognitive, relevance and interactive “revolutions” expected to take place in IR evaluation, the focus in interactive IR experimentation has shifted to exploring the dynamic information need that evolves during the search process, the situational context that influences the relevance judgments, and the strategies and tactics adopted by information seekers in satisfying their information need. This paradigm shift toward a cognitive approach to exploring search interactions and to studying **Human Information Behavior** has generated a large number of theories that attempt to model the search interaction and to predict the user's behavior in different contexts and at different stages of the interaction (Fisher et al, 2005).

Of particular interest to this author are models of the search interaction process and empirical work to validate such models by observing consistent patterns of user behavior (Ellis, 1989; Kuhlthau, 1991; Belkin et al, 1995; Saracevic, 1996; Xie, 2000; Vakkari, 1999, 2001; Olah, 2005). Of course, the interest is not simply in validating theoretical models, but also in designing systems that better respond to users' needs, that adapt to support various search strategies, and that offer different functionality in different stages of the information seeking process.

We are interested in methodologies for running such experiments. Although no systematic study has investigated the methodological details for this kind of experiments, there is plenty of anecdotal evidence to suggest that much of the investigation is manual: the researchers study interaction transcripts or videos, and code significant actions that take place and shifts between interaction stages. As any human activity, this process is slow, expensive, and error prone. Logs of interactions are sometimes employed to address this issue. However, in our experience, there is usually little or no formal process in designing the logs, the logging process, and the log analysis, in order for the states of the system and the stages of the interaction to be captured. What we are proposing in this paper is a semiformal procedure that supports logging and log analysis, so that the stages of the interaction and the states of the system are captured accurately, and can be analyzed in a systematic and at the same time flexible way.

A second motivation for the proposed methodology comes from observations of interactive IR experiments where the systems had clear usability issues: “Save”, “Bookmark” or “View” buttons active when no documents were selected, or even before a search was conducted, “Search” button active when no query

was specified, “Back” button active when no document was yet in the history stack, etc. Such situations are common and not at all surprising: these are experimental systems (as opposed to operational systems), built for studying certain aspects of the interaction, so little or no resources are available for high-quality design and usability testing. Unfortunately, this can potentially lead to compromised research results, as the usability of the interface can potentially affect the searchers’ behavior. Our proposed methodology, although imposing an initial design overhead, promises to alleviate this situation and to support an overall improvement in the quality of the experimental results.

2. SYSTEM STATE-BASED DESIGN OF INTERACTION AND LOGGING

2.1 General approach

Most often, the specification of an interactive system is in the designer’s natural language, such as English, accompanied by a set of the sketches of the interface at different stages of the interaction. Unfortunately, natural-language specifications tend to be lengthy, vague and ambiguous, and therefore are often difficult to prove complete, consistent and correct. Formal and semiformal languages, usually used in fields such as mathematics, physics or circuit design, have also proven their value in modeling command language systems (Shneiderman and Plaisant, 2004).

Our approach is based on statecharts (Harel, 1988) or, in the more modern UML (Unified Modeling Language)¹ speak, on state diagrams. These are extensions of finite state diagrams², in which the use of memory and of conditional transitions makes it practical to describe system behavior in reasonably compact diagrams. Such a model of a system describes; (i) a finite number of existence conditions, called **states**; (ii) the **events** accepted by the system in each state; (iii) the **transitions** from one state to another, triggered by an event; (iv) the **actions** associated with an event and/or state transition (Douglass, 1999; Fowler, 2004). Such diagrams have the advantage that they describe in detail the behavior of the system and, being relatively easy to learn and use, allow the participation of the entire research team in developing the conceptual model of the IR system to be employed in an experiment. It also makes it easier for the designated programmers to implement and test the system, as the logic is captured in the model.

While not widely used in designing IR or other interactive systems (according to our observations), the state diagrams are certainly not new. What is novel is our proposed integration of interface design with logging and log analysis and, at an implementation level, between UML and the XML (eXtensible Markup Language)³ family of languages. The general approach is described here, with details discussed in the following subsections.

From the state diagrams, an XML-based **Interaction Modeling Language (IML)** can be derived, which will capture in a DTD (Document Type Definition)⁴ or XML schema⁵ format the valid

states of the system, and the valid events and actions taking place during the interaction. Subsequently, two software modules can immediately be designed and implemented: (i) a logger that captures each valid event and action that takes place, and each state transition undergone by the system; (ii) a log analyzer that uses an XML parser and identifies events, actions and state transitions, and analyzes the data according to the research hypotheses being investigated. Note that, apart from a number of design decisions discussed below, these steps are straightforward, once the state diagrams and the IML are agreed upon. For example, if Java is the implementation language, then the standard logging package⁶ makes it extremely simple to output logs in XML. Also, open-source tools (such as NetBeans⁷) can automatically generate log parsers, given the DTD or XML schema adopted for the logs.

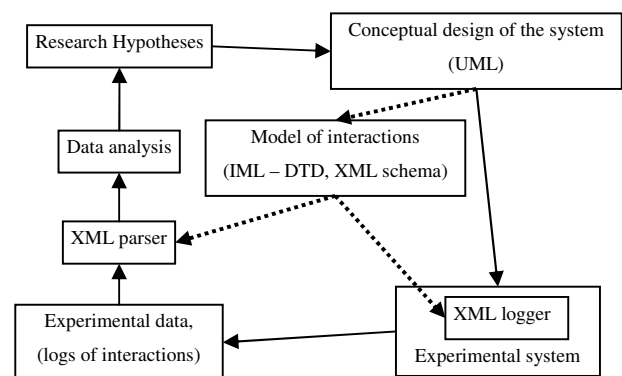


Figure 1. Integrated approach to design, logging and analysis

State diagrams allow different levels of granularity via sub-diagrams that can detail certain system states, showing any transitions between sub-states. It is up to the research team to decide the level of granularity and the precision of their model. On the one hand, a tight deadline may force the design to include just the actions and transitions relevant to the research questions investigated. On the other hand, a more detailed design upfront can produce a much richer log and support the exploration of un-anticipated research hypotheses. For example, in a current mediated retrieval project, we captured in the interaction model, and subsequently logged, all the query edits performed by the searcher (including backspace/del corrections, copying and pasting, etc). While this was not envisaged at the beginning of the project, we are now able to investigate additional research questions, such as whether the searcher’s familiarity with the search topic correlates with the number of query corrections, with the query length, or with the number of query terms typed in the query box (as opposed to copied and pasted from the description of the assigned search topic).

The following subsections discuss some of the decisions that need to be taken when applying this methodology, and some implementation details.

¹ <http://www.uml.org/>

² http://en.wikipedia.org/wiki/Finite_state_machine

³ <http://www.w3.org/XML/>

⁴ <http://www.w3.org/XML/1998/06/xmlspec-report.htm>

⁵ <http://www.w3.org/XML/Schema>

⁶ <http://java.sun.com/j2se/1.5.0/docs/guide/logging/index.html>

⁷ <http://www.netbeans.org/>

2.2 Design patterns in the log analyzer

Parsing XML has become routine due to the multitude of open-source parsers and parser generators available for a variety of programming languages. For extremely large logs, unlikely to fit in the computer memory for the analysis, a SAX (Simple API for XML)⁸ is needed. This type of parser identifies the beginning and end of various elements found in the log, and processes them based on the callback methods provided by the programmer/researcher. The more desirable approach, possible for logs of reasonable size, is to use a DOM (Document Object Model)⁹ parser, which builds in the computer memory a tree of the log, and allows the programmer to visit it in whatever order makes sense for a research hypotheses. For example, if the research hypothesis being investigated is related solely to the documents bookmarked by the searcher, it is possible and easy to visit just the nodes capturing document bookmarking.

It is common for XML parsers generated automatically based on DTD (such as the one produced by NetBeans) to implement the **Visitor** software design pattern, which allows flexibility in specifying which elements of the log tree should be visited and in what order, in order to collect, process and summarize information. From our experience, we suggest combining that with the **State** design pattern, where different classes correspond to states in the state diagram. This allows the state objects to accumulate, summarize and report information in a simple and flexible fashion (Gamma et al, 1995).

For simple systems, the implementation of the State design pattern is straightforward. For complex states, class inheritance is used to implement subclasses, and composition is used for concurrent orthogonal states.

2.3 Explicit vs. implicit logging of states

At first sight, explicitly logging the system states is natural, so that someone examining the logs can clearly see what happened while the system was in a certain state, and when a state transition occurred. However, logs are usually so large and contain so many details, that the researcher is unlikely to gain much knowledge from examining them visually. Rather, the logs should be processed and the information pertinent to a certain research question should be summarized, and possibly visualized, so that it can be interpreted by the researcher. Moreover, the log analyzer will be able to re-create or infer the states based on the events and actions captured in the logs.

Some arguments in favor of not capturing the states explicitly, and in having the log analyzer infer them, are compelling. First, complex systems such as the user interface of a search engine are likely to have complex states, with nested sub-states, and often have concurrent orthogonal states. For example, if the user edits text in an “answer panel”, based on information collected via searching and browsing in a “search hits panel”, then the states of the two panels are components of the overall system state, and the state transitions in the two panels may happen independent of each other. Attempting to log the parallel states and the transitions is likely to produce nesting that cannot be captured in a well-formed XML document.

⁸ <http://www.saxproject.org/>

⁹ <http://www.w3.org/DOM/>

Another advantage of capturing just events and actions in the log and re-creating the states via the log analyzer is that other interaction logs, obtained from previous experiments, or from experiments run by other researchers, can be analyzed based on the same approach., as long as these logs are converted from their native format into the XML format suggested by us.

2.4 Online vs. offline analysis

It is apparent that the State design pattern can be used both in designing the user interface, and in designing the log analyzer. A couple of related questions can be asked: (1.) is one set of classes sufficient, or should a set of classes be used in the user interface and a different one in the log analyzer ?; (2.) should the data be accumulated, summarized and analyzed online, while the experiment takes place, or should it be logged and analyzed at a later time ?

The second question is easier: we recommend logging all the events and actions, and doing the analysis offline. Here are some arguments: (i) occasionally, systems do crash during the experiment, in which case the information accumulated in the memory will be lost; (ii) for some of the statistical analysis, raw data rather than summaries or means are needed for between-subjects comparisons; (iii) previously not envisaged research questions may appear during the initial analysis, and these may be addressed if the entire raw data are available.

The answer to the first question depends on the complexity of the system, and on the designer’s preference. Our preference leans towards separating the software module for running the system from the software module for analyzing the data (in Java, these can be part of different packages) in the interest of increased cohesion and clarity.

3. CASE STUDY

We have started developing the proposed approach while running the Interactive TREC 2003 experiment¹⁰ and are continuing to refine it while applying it to a current project on Mediated Information Access. However, designing an IR system is a complex enterprise and the full state diagrams may be somewhat difficult to follow by an un-trained reader. Therefore, we are exemplifying here the first mock project on which we applied our methods: a JukeBox application adapted from sample code available with the Java SDK to demonstrate the use of audio and other media.

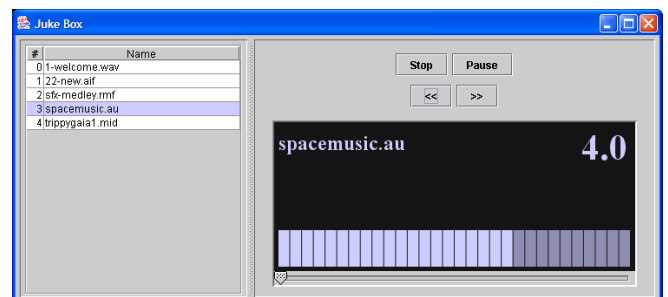


Figure 2. JukeBox user interface

¹⁰ <http://www.scils.rutgers.edu/~muresan/trec/inter2003.html>

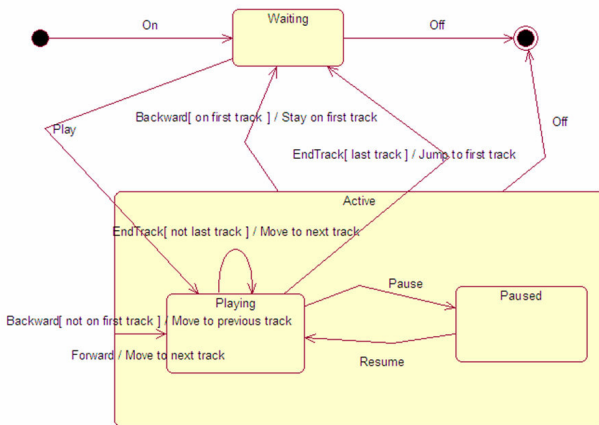


Figure 3. JukeBox state diagram

```
<?xml version='1.0' encoding='UTF-8'?>
<ELEMENT log (record)*>
<ELEMENT record (date, millis, message)>
<ELEMENT date (#PCDATA)>
<ELEMENT millis (#PCDATA)>
<ELEMENT message (StartSession|StopSession|Start|Pause|Forward|Backward|Stop|Resume)*>
<ELEMENT StartSession EMPTY>
<ELEMENT StopSession EMPTY>
<ELEMENT Start (#PCDATA)>
<ELEMENT Stop (#PCDATA)>
<ELEMENT Forward (#PCDATA)>
<ELEMENT Backward (#PCDATA)>
<ELEMENT Pause (#PCDATA)>
<ELEMENT Resume (#PCDATA)>
```

Figure 4. JukeBox interaction DTD

```
<?xml version="1.0" encoding="windows-1252"
standalone="no" ?>
<!DOCTYPE log (View Source for full doctype...)>
<log>
<record>
<date>2004-04-26T23:49:26</date>
<millis>1083037766266</millis>
<message>
<StartSession />
</message>
</record>
<record>
<date>2004-04-26T23:49:34</date>
<millis>1083037774378</millis>
<message>
<Start>1-welcome.wav</Start>
</message>
</record>
<record>
<date>2004-04-26T23:49:42</date>
<millis>1083037782429</millis>
<message>
<Pause>1-welcome.wav</Pause>
</message>
</record>
<record>
<date>2004-04-26T23:49:44</date>
<millis>1083037784042</millis>
<message>
<Forward>1-welcome.wav</Forward>
</message>
```

Figure 5. JukeBox log extract

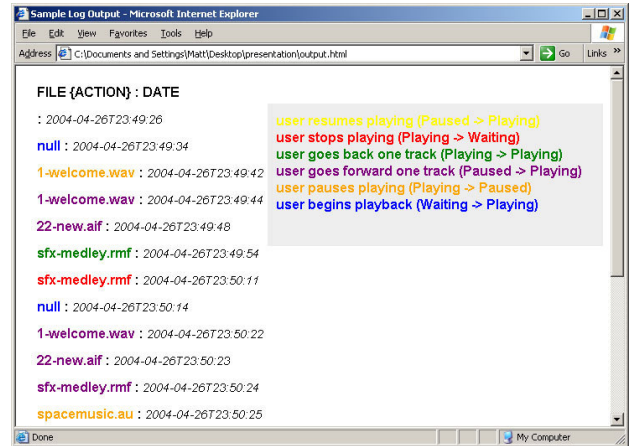


Figure 6. JukeBox activity summary, extracted from log

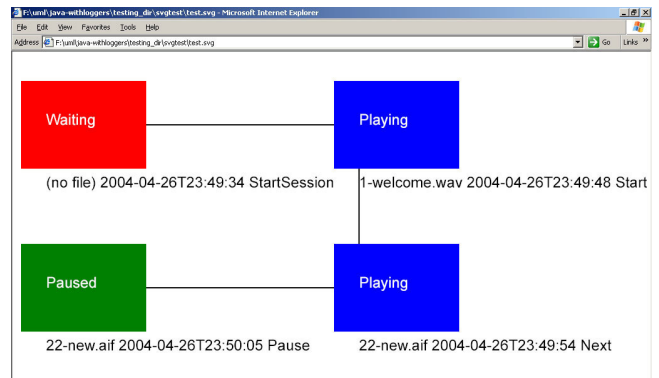


Figure 7. JukeBox state transition diagram

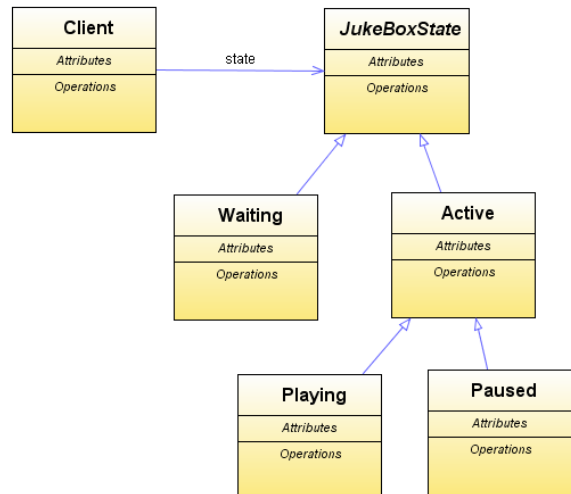


Figure 8. Classes for implementing the JukeBox states

Figures 2-7 depict the user interface, the state diagram, the Document Type Definition specifying the possible actions, an

extract of the use log obtained following a user's interaction with the system, and two versions of log analysis output: a summary of events and actions, in HTML format, and a diagram of states, in SVG (Scalable Vector Graphics)¹¹ format. Finally, Figure 8 shows the class diagrams corresponding to the states of the system. The attributes and methods of these classes are not specified and neither is the Client class, as two sets of such classes (with the same names) were used: one to run the application and the other to analyze the logs. The attributes and methods in the two sets correspond to different functionality, and are therefore different.

4. CONTRIBUTIONS AND FUTURE WORK

The proposed methodology is a novel and significant contribution to experimental Information Seeking and Retrieval. It is particularly suitable for studying exploratory searching, where the research questions are usually related to understanding patterns of behavior in different stages of the interaction. This approach has been successfully applied in a mock project (the JukeBox) and in a real IR project (Interactive TREC 2003) and is being refined while being applied on a new project.

One issue that we are currently investigating is the automatic generation of the XML schema or DTD describing the interaction, based on the UML state diagram. In our projects we used a variety of modeling tools, each with its own file format, and we generated the XML schema manually (or rather intellectually). These days most modeling tools allow the export of the diagrams in XMI (XML Metadata Interchange)¹² format, and we are looking into converting state diagrams from XMI into XML schemas with no or minimal human effort.

We are also investigating ways to automatically generate graphical diagrams that show the frequency of each state transition and thus give a visual display of user behavior (so far we have extracted transition frequencies with the log analyzer, but have built the diagrams manually).

Finally, we intend to investigate a number of IR user interfaces and to compare their state diagrams, trying to identify common patterns. This would allow us to provide support, in the form of reusable toolkits of frameworks, for researchers designing and evaluating user interfaces for Information Retrieval.

5. REFERENCES

- [1] Belkin, N.J., Cool, C., Stein, A., Thiel, U. (1995). Cases, scripts, and information-seeking strategies: on the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3), 379-395.
- [2] Douglass, B. P. (1999) *Doing Hard Time: Developing Real-Time Systems with UML, Objects, Frameworks, and Patterns*, Addison-Wesley, Reading, MA.
- [3] Ellis, D. (1989). A behavioral approach to information retrieval system design. *The Journal of Documentation*, 45(3), 171-212.
- [4] Fisher, K. E., Erdelez S. and McKechnie, L. (2005) *Theories of Information Behavior*, Information Today, Medford, NJ.
- [5] Fowler, Martin (2004) *UML distilled: A brief guide to the standard object modeling language*, 3rd ed, Addison-Wesley/Pearson Education.
- [6] Gamma, E., Helm, R., Johnson, R. and Vlissides, J. (1995) *Design Patterns – Elements of Reusable Object-Oriented Software*, Addison-Wesley, Reading, MA.
- [7] Harel, D. (1988) On visual formalisms, *Communications of the ACM*, 31 (5).
- [8] Ingwersen, P. and Jarvelin, K. (2005) *The Turn – Integration of Information Seeking and Retrieval in Context*. Springer.
- [9] Kuhlthau, C. (1991). Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5), 361-371.
- [10] Olah, J. (2005) Shifts Between Search Stages During Task-Performance in Mediated Information-Seeking Interaction, *Proceedings of the 68th Annual Meeting of the American Society for Information Science (ASIST)*, 42, Charlotte, NC.
- [11] Robertson, S.E., Hancock-Beaulieu, M.M. (1992). On the evaluation of IR systems. *Information Processing and Management*, 28(4), 457-466.
- [12] Saracevic, T. (1996). Interactive models in information retrieval (IR). A review and proposal. *Proceedings of the 59th Annual Meeting of the American Society for Information Science (ASIST)*, 33, 3-9.
- [13] Shneiderman, B. and Plaisant, C. (2005) Section 5.2: Specification Methods, in *Designing the User Interface*, Addison-Wesley / Pearson Education, p.175-183.
- [14] Vakkari, P. (1999). Task complexity, problem structure and information actions, integrated studies on information seeking and retrieval. *Information Processing and Management*, 35, 819-837.
- [15] Vakkari, P. (2001). Changes in search tactics and relevance judgments when preparing a research proposal: a summary and generalization of a longitudinal study. *Journal of Documentation*, 57(1), 44-60.
- [16] Xie, H. (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science*, 51(9), 841-857

¹¹ <http://www.w3.org/Graphics/SVG/>

¹² <http://en.wikipedia.org/wiki/XMI>

Exploratory Search in Wikipedia

Sisay Fissaha
ISLA, Informatics Institute
University of Amsterdam
sfissaha@science.uva.nl

Maarten de Rijke
ISLA, Informatics Institute
University of Amsterdam
mdr@science.uva.nl

ABSTRACT

We motivate the need for studying the search, discovery and retrieval requirements of Wikipedia users. Based on a sample from an experimental Wikipedia search engine, we hypothesize that the fraction of Wikipedia searches that are exploratory in nature is at least the same as that of general web searches. We also describe a questionnaire for eliciting search, discovery and retrieval requirements from Wikipedia users.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Design, Experimentation, Human Factors

Keywords

Wikipedia, interfaces, exploratory search

1. INTRODUCTION

In this paper, we describe work in progress that is aimed at eliciting the user requirements of Wikipedia users, with a special focus on so-called undirected (or exploratory) queries. Understanding what users search for, and how their information needs can best be met, is of increasing interest, both for the scientific community and for society at large, especially where it concerns valuable, and increasingly popular resources such as Wikipedia. Increasingly, the IR research community seeks to understand *why* users conduct searches since it is believed that knowing the user intentions or goals may help tailor output of search engines to the needs of a particular user. As a result, understanding what users search for and why users conduct searches, which we refer to as *user requirements*, has become an active area of research.

Broder [2] and Rose and Levinson [7] identified the follow-

ing broad classes of goals of web users: navigational and informational. Informational goals are further classified as directed, undirected, locate, advice, and list request. Directed requests may be open or close ended. Users issuing directional queries are typically looking for a focused answer whereas those issuing an undirected informational query about X want to learn something/everything about X. As the name implies, locate, advice and list requests are specialized requests that attempt to locate, get advice or list of items. Rose and Levinson [7] showed that over 60% of web search queries are informational in nature, and so-called undirected (or exploratory) queries form a significant subclass of these.

Do the above findings carry over from the general web to Wikipedia? For instance, are there navigational goals in Wikipedia searches? Is the distribution of directed and undirected queries different in the context of Wikipedia? We hypothesize that users of encyclopedias—electronic or otherwise—have relatively uniform information needs which may largely be characterized as *informational*. Is this correct?

Studies aimed at answering such questions range from developing generalized models, which identify the different factors that affect the information search behaviour, to devising specific strategies for collecting empirical data required for testing the different hypotheses [6, 4]. Most studies adopt different strategies such as user studies or query log analysis to identify typical user search requirements and behaviours. Recently, Pharo and Järvelin [6] indicated that such strategies are limited in their ability to provide the necessary data for studying the different factors and their relationships, and proposed an extensive data collection and analysis methods. It is, however, a very time consuming and expensive approach and, hence, it may not be feasible to conduct relatively large surveys. As a result, we adopt commonly used methods of data collections for the current study.

Specifically, we use a two-fold approach towards answering the above questions. First, we create pilot applications that implement novel ways of accessing the information provided by Wikipedia; see, e.g., [3] (link suggestions), [8] (focused access at the sub-document level), [11] (exploratory question answering using Wikipedia), and try to mine useful information from the query logs. Second, we are in the process of setting up an online questionnaire aimed at eliciting the requirements of Wikipedia users.

The remainder of the paper is organized as follows. In Section 2 we provide background on Wikipedia and on accessing Wikipedia. Then, in Section 3 we examine a sample from a Wikipedia search engine query log. After that we describe the design of our user survey, and we conclude in Section 5.

2. SEARCH, DISCOVERY AND RETRIEVAL IN WIKIPEDIA

2.1 About Wikipedia

Wikipedia possesses a number of special and useful characteristics which call for possibly different access methods and which make investigation of the information access problem especially challenging and interesting. Among these are:

- Wikipedia is the result of a collaborative content development effort regulated by group consensus, without strict guidelines and control. Therefore, individuals may want to have a flexible browsing and search facility which will enable them to have a global view of Wikipedia while editing locally.
- Wikipedia is an encyclopedia, hence contains different types of information which may call for different modes of access. Access to geographic information, for example, can be enhanced with a map-based explorative search interface.
- Wikipedia's content consists of both structured and unstructured textual data, and it also other data types such as images, video. This in turn provides a useful experimental setup to apply the methods developed for different data types.
- Wikipedia's content can be edited by anyone. The same user may be a reader with a specific information need, or an author who wants to create an entry. The only source of information for anyone who wants to create a page is the general guidelines provided in Wikipedia website. Authors are not normally trained. It should be possible for individuals to learn more about Wikipedia in the process of using it or contributing to it which in turn may call for a more explorative search interface.

2.2 Access to Wikipedia

Traditionally, access to (paper-based) reference works such as encyclopedias has relied on alphabetic listings of the titles of the entries, on cross-references, and on multiple indexes. Many of these strategies seem to have been carried over to their online counterparts—Wikipedia is no exception. Today, Wikipedia has become one of the primary reference sites; its main site (<http://wikipedia.org>) consistently ranks amongst the top 50 sites in terms of traffic [1]. Wikipedia's increasing popularity has gone hand in hand with its growth in size, which has been steady and exponential, going from 0 articles in 2001 to over 1,000,000 (for the English part alone) during the first half of 2006 [10]. This growth in size and popularity call for effective support methods; as the distinction between reader and author in the Wikipedia context is being blurred [5], such support methods are needed for both readers who want to locate information in Wikipedia and for authors who want to contribute to the growing number of stubs and articles.

Currently, Wikipedia provides a keyword-word based search facility which allows users to enter a set of keywords and get a ranked list of Wikipedia pages. There are also other search engines that provide efficient and focused access to the Wikipedia content though the basic search facilities remain more or less the same. In addition, Wikipedia has dense networks of hyperlinks which eases browsing through the content. Furthermore, each page is assigned to categories or lists that groups pages into some kind of semantic classes. These features allow for extra browsing and navigation facilities.

Though the facilities sketched above—and other ones not listed, such as the Wikipedia categories—ease the burden of searching and browsing through the Wikipedia content, we believe that the nature of Wikipedia and its development process may require more advanced search facilities that go beyond what is currently available. For example, one peculiar property of Wikipedia is that the distinction between readers and authors is blurred—for both it may be useful to be able to generate templates with slots for prototypical facts about entities falling within a particular category, and for authors/editors it may be useful to be able to automatically check for structural consistency such as link structure.

Results of a recent Wikipedia-based study [8] call for a proper investigation of the requirements of the new generation of users in order to better meet their information needs. Sigurbjörnsson et al. [8] conducted an experiment on the advantage of providing focused access (i.e., direct access to the sections: “go and read here”) to the content of Wikipedia over a full-document retrieval baseline. The result showed that “focused access allows users to solve their search task quicker, at least when the information need is specific.” But what if the information is not specific, and users need to *explore* Wikipedia's content, because they need to “find out” about a topic? In the following section we look at a sample from a Wikipedia search engine log file—the sample suggests that many users have such undirected information needs.

3. A WIKIPEDIA SEARCH ENGINE QUERY LOG

We analysed a random sample of 200 queries that are taken from an experimental Wikipedia search engine [9] that is publicly available. As we only have access to the query log—and not to the users submitting the queries—it is difficult to carry out a detailed classification of the user intention or goals. Hence, we could not use the classification used in [7]. For our classification, we adopted three classes: directed informational goal (“I want to learn something specific about my topic”; D), undirected or exploratory informational goal (“tell me about my topic”; X), and unknown (ones that we were unable to classify; UN).

For directed goals, we checked for the presence of the following properties in the queries:

- is the query in the form of a factoid question
- does the question have the following form: capital Iceland, population Netherland, inventor computer, Woody Allen married etc.—in short, queries that can typi-

cally be answered in terms of a specific named entity or clause.

For undirected informational goals, we checked the following properties in the queries:

- is the query a well-formed phrase and does it represent a well-defined concept or entity?
- does the query match the title of a Wikipedia page?
- does the query represent a person or name of a place or location?

All queries not covered by the above conditions were classified as unknown. The results of classifying a random sample of 200 queries from the log are given in Table 1. A significant portion of the queries are undirectional or exploratory queries. Of the 145 undirectional queries, 47 (32%) have a Wikipedia entry. This preliminary result shows that undirected search queries seem to be the dominant type of queries, as with general web queries. Since the above analysis is very limited and far from accurate, we plan to carry out a survey, to verify results obtained from the search engine query—setting up this survey is the topic of the next section.

Query Types	Frequency	
Directed	8	(4%)
Undirected	145	(72.5%)
Unknown	47	(23.5%)

Table 1: Classification results of the queries

4. SOLICITING REQUIREMENTS

We plan to extend our experimental Wikipedia search engine with a questionnaire. The questions are organized into four categories.

4.1 General

The first set of questions are used to identifying the type of user (reader or author or both), the frequency of access to Wikipedia, and purpose of use.

- How often do you use Wikipedia?
- Do you edit Wikipedia articles?
- What do you use Wikipedia for?

4.2 Type of Information

The next set of questions attempt to identify what sort of information people are looking for. This may roughly map to the user goals or intentions enumerated in the introduction, such as directed, undirected, etc.

- Did you have a specific question in mind?
- Would you prefer to express your queries in terms of natural language?

- Given the three types of information needs illustrated by the following questions, which one illustrates your information needs best?

- I want to know the capital city of Kenya. I want to know who invented the telephone.
- I want to know how to make pasta. I want to know why bears hibernate.
- I want to know something about lung cancer. I want to know who Micheal Jackson is.
- I want the list of countries in Latin America. I want the names of some programming languages.
- If none of the above, could you formulate a question expressing your information need?

4.3 Author-related Questions

Unlike the previous set of questions, which are targeted more to the readers of Wikipedia, the following set of questions is geared more to the requirements of authors of Wikipedia. Though the distinctions between the two may be blurred or non-existent from the user’s perspective, they may still pose different requirements when viewed from the system development perspective.

- I want to be able to automatically create hyperlinks.
- I want to be able to check the consistency of the hyperlink structure.
- I want to validate my sentences against snippets extracted from the web.
- I want to get automatic update support for a page’s content.
- Could you specify other information needs that you may think will fall under this category?

4.4 “Professional” users

So far the focus has been on the readers or authors of Wikipedia. As the size and popularity of Wikipedia increases, the intended use of it also varies a lot. Recently, its content has also become target of scientific enquiries. Though it might be very hard to characterize the type of users under this category, it might still be useful to include some possible questions in the survey.

- I want to retrieve sentences with a particular named-entities or describing a particular events.
- I want statistical summaries of the corpus. What sort of statistical summary do you need?
- I want to visualize Wikipedia content. What should the visualization should include?
- Do you have any particular requirements?

5. CONCLUSIONS

In this paper we motivated the need for studying the requirements of Wikipedia users. We hypothesized that the fraction of Wikipedia searches that are exploratory in nature is at least the same as that of general web searches. We described a questionnaire for eliciting search, discovery and retrieval requirements from Wikipedia users. We expect to be able to report on initial results from our questionnaire at the time of the EESS workshop.

6. ACKNOWLEDGMENTS

Sisay Fissaha Adafre was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.000.106, 612.066.-302, 612.069.006, 640.001.501, and 640.002.501.

7. REFERENCES

- [1] Alexa, 2005. Traffic ranking for Wikipedia. URL: <http://www.alexa.com/data/details/?url=wikipedia.org>, accessed October 2005.
- [2] A. Broder. A taxonomy of web search. In *SIGIR Forum*, pages 3–10, 2002.
- [3] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*, 2005.
- [4] P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50, 1996.
- [5] N. Miller. Wikipedia and the disappearing “Author”. *ETC: A Review of General Semantics*, 62(1):37–40, 2005.
- [6] N. Pharo and K. Järvelin. The SST method: a tool for analysing web information search processes. *Information Processing & Management*, 40(4):633–654, 2004.
- [7] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th intern. conf. on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.
- [8] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Focused access to wikipedia. In F. de Jong and W. Kraaij, editors, *6th Dutch-Belgian Information Retrieval Workshop (DIR 2006)*, pages 73–80, 2006.
- [9] Wikiii, 2006. Wikiii: A focused search engine for Wikipedia. URL: <http://berk.science.uva.nl:8080/wikiii/>, accessed May 2006.
- [10] Wikipedia, 2005. Size of Wikipedia. URL: http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia, accessed October 2005.
- [11] WiQA, 2006. Question Answering Using Wikipedia. URL: <http://ilps.science.uva.nl/WiQA/>, accessed May 2006.

A Pilot for Evaluating Exploratory Question Answering

Valentin Jijkoun
ISLA, Informatics Institute
University of Amsterdam
jijkoun@science.uva.nl

Maarten de Rijke
ISLA, Informatics Institute
University of Amsterdam
mdr@science.uva.nl

ABSTRACT

We describe a pilot on evaluating exploratory search in Wikipedia, the free online encyclopedia. The pilot will be held at CLEF 2006, and brings together both search and navigation, and reading and authoring.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software—*performance evaluation*

General Terms

Experimentation, Measurement

Keywords

Test collection formation, evaluation, question answering, Wikipedia

1. INTRODUCTION

Question Answering (QA) has attracted a great deal of attention, especially since the launch of the QA track at TREC in 1999. While significant progress has been made in technology for answering general factoids (e.g., *How fast does a cheetah run?* or *What is the German population?*), there is a real need to go beyond such factoids [5, 2]. At the TREC QA track this has been recognized through the introduction of definition questions and of so-called “other” questions that are far more exploratory in nature and ask for important information about a topic at hand that the user does not know enough about to ask.

In this paper we describe a pilot evaluation task that takes the “other” questions a step further. The task, called WiQA (Question Answering using Wikipedia [7]), will be organized as part of CLEF 2006. It involves answering undirected informational queries [3] against Wikipedia, the free online encyclopedia [6]. The purpose of the WiQA pilot is to develop novel question answering technologies, ones that go beyond

the traditional highly focused factoid questions to include more open and exploratory ones, using the rich structure and reliable content of the Wikipedia.

Below, we first describe our take on different ways of accessing Wikipedia; then we provide details of the WiQA pilot, including a detailed example. After that we briefly describe the assessment criteria and evaluation metrics to be used at WiQA.

2. ACCESSING WIKIPEDIA

We believe that natural ways of accessing the information in Wikipedia mix two types of things:

- search and navigation, and
- reading and authoring.

Given this assumption, there are many natural possible tasks, or aspects of tasks, that are of interest in the WiQA pilot. To start, these include (the usual) highly focused questions. For instance, when using Wikipedia as the source to answer factoid questions (such as *How big is Berlin?* or *Find tennis players born in Berlin*), QA systems can use layout, formatting and wording regularities to pinpoint answers. In addition, they can use explicit semantic annotation: lists (such as *List of male tennis players*), categories (e.g., *Andre Agassi* is categorized into *Las Vegans*, *American tennis players*, etc.), structured tables (so-called *templates* providing standard information about countries, people, etc.).

As to more exploratory types of questions, there are many scenarios that seem very natural in the Wikipedia setting as well as many research questions that such scenarios give rise to. Below we list some of these scenarios and research questions:

Summarizing the content of Wikipedia articles. This corresponds to answering non-factoid questions such as *Tell me important facts about Andre Agassi*. Addressing such information needs raises important research questions. Is the current structure of Wikipedia pages good enough: aren't the “leads” perfect summaries of single pages? Is some level of (user-dependent) summarization needed for very long articles?

Summarizing the structure of Wikipedia. This may allow us to recover relevant information that is not explicitly given on a page, but is rather distributed across entire encyclopedia. E.g., what are important articles that mention Andre Agassi, and what do they say about him?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

Alice Cooper's musical career begins (from the article *1969 in music*)

Relocating to London in the late '70s, they worked all over the United Kingdom and Europe to establish themselves, ... supporting the top hard-rock acts of the day including Alice Cooper (from the article *AC/DC*)

February 24: Alice Cooper announces that he is going to run for Governor of Arizona. (from the article *1988 in music*)

Throughout the 1960s and 1970s, Ann Arbor was home to many influential rock and roll bands, such as the MC5, Alice Cooper, Iggy Pop... (from the article *History of Ann Arbor, Michigan*)

Arthur Brown, a British rock and roll singer known for his flamboyant, theatrical style and significant influence on shock rockers such as Alice Cooper and Marilyn Manson (from the article *Arthur Brown (musician)*)

Table 1: Snippets to be added to the article on Alice Cooper

Automatic clustering of Wikipedia articles. Systems may use link structure, layout and semantic annotation to find similar or related pages. This is useful to identify potential missing list pages (such as *List of U.S. Open champions*). What techniques are appropriate here? Flat or hierarchical clustering, labelled or unlabelled clusters? Which of the many features of articles are useful for this task?

Handling navigational information needs. This group of tasks includes finding pages similar to a given one, as well as important or popular articles around a given topic. Addressing these needs may also involve generating (ranked) lists of different types of entities related to a given topic (the timeline of events, related locations and organizations), in context, as well as identifying and browsing multiple IS-A hierarchies provided by the category structure of Wikipedia.

Multi-lingual aspects. At present (June 2006) there are 10 languages with more than 100,000 articles, and 30 more have over 10,000 articles. Many pages are linked to their counterparts in other languages. Is it possible to automatically detect inconsistencies and missing article alignments? Can we compare existing cross-language alignments of Wikipedia pages, detect missing subgraphs for different languages? Is it possible to use machine translation to generate stubs for pages missing in one language?

As we will see below, the task defined for WiQA 2006 mixes some of these aspects.

3. WIQA 2006

The WiQA 2006 task that we envisage mixes search and navigation, and we are keen on exploring the reader-author inversion, building systems that help provide access to Wikipedia's content and that help author and edit its content. In the WiQA pilot, we will exploit the fact that, in Wikipedia, the distinction between author and reader has become blurred. Specifically, we aim to see how information retrieval and language technology can be effectively used to help readers and authors of articles get access to information spread throughout Wikipedia rather than stored locally on a single page.

Cryptonomicon is a 1999 novel by Neal Stephenson that concurrently follows the exploits of World War II-era cryptographers affiliated with Bletchley Park in their attempts to crack Axis codes... (from the article *Cryptonomicon*; assessed as novel, non-repeated and important)

A rare Abwehr Enigma machine, designated G312, was stolen from the Bletchley Park museum on 1 April 2000... (from the article *Enigma machine*; assessed as novel, non-repeated and important)

Together with the cryptographic efforts centered at Bletchley Park and also at Arlington Hall, the development of radar and computers in the UK and later in the USA, and the jet engine in the UK and Germany, the Manhattan Project represents one of the few massive, secret, and outstandingly successful technological efforts spawned by the conflict of World War II. (from the article *Manhattan Project*; assessed as not novel, non-repeated and important; the important information is actually contained in the original article "Bletchley Park": *The Bletchley Park effort was comparable in influence to other WWII-era technological efforts, such as... Manhattan Project...*)

Olivia's father, Brin Newton-John, originally from Wales, was an MI5 officer attached to the Enigma machine project at Bletchley Park... (from the article *Olivia Newton-John*; assessed as novel, non-repeated and not important)

Table 2: Snippets and their assessments for the Wikipedia article *Bletchley park*

3.1 Task description

As our user model we take the following scenario: a reader or author of a given Wikipedia article (the source page) is interested in collecting information about the topic of the page that is not yet included in the text, but is relevant and important for the topic, so that it can be used to update the content of the source article. Although the source page is in a specific language (the source language), the reader or author would also be interested in finding information in other languages (the target languages) that he explicitly specifies.

With this user scenario, the task of an automatic system is to locate information snippets in Wikipedia which are:

- outside the given source page,
- in one of the specified target languages,
- substantially new w.r.t. the information contained in the source page, and important for the topic of the source page, in other words, worth including in the content of (the future editions of) the page.

To illustrate these ideas, let us look at an example. Consider a user wishing to update the article for Alice Cooper. Table 1 lists snippets from other English articles that seem interesting and novel for the topic, thus, worth including in the page.

Participants of the WiQA 2006 pilot will be able to take part in two flavors of the task: a monolingual one (where the snippets to be returned are in the language of the source page) and multilingual (where the snippets to be returned can be in any of the languages of the Wikipedia corpus used at WiQA).

3.2 Corpus

The corpus to be used at WiQA 2006 consists of XML-ified dumps of Wikipedia in three language: Dutch, English, and Spanish. The dumps are based on the XML version of the Wikipedia collections [1] that include the annotation of the structure of the articles, links between articles, categories, cross-lingual links, etc. For the WiQA 2006 pilot the collections were enriched with annotations of sentences and classification of pages into named entity classes (person, location, organization).

3.3 Assessment of the systems' results

Given a source page, automatic systems return a list of short snippets, defined as sequences of at most two sentences from a Wikipedia page. The ranked list of snippets for the topic will be manually assessed using the following binary criteria, largely inspired by the TREC 2003 Novelty task [4]:

- *support*: the snippet does indeed come from the specified target Wikipedia article.
- *novelty*: the information content of the snippet is not subsumed by the information on the source page
- *non-repetition*: the information content of the snippet is not subsumed by the target snippets higher in the ranking for the given topic
- *importance*: the information of the snippet is relevant to the topic of the source Wikipedia article, is in one of the target languages as specified in the topic, and is already present on the page (directly or indirectly) or is interesting and important enough to be included in an updated version of the page.

Note that we distinguish between novelty (subsumption by the source page) and non-repetition (subsumption by the higher ranked snippets) in order for the results of the assessment to be re-usable for automatic system evaluation in future: novelty only takes the source page and the snippet into account, while non-repetition is defined on a ranked list of snippets.

To illustrate these ideas, Table 2 provides an example of assessments of snippets found for the target page *Bletchley Park*.

3.4 Evaluation metrics

One of the purposes of the WiQA pilot task is to experiment with different measures for evaluating performance of systems. WiQA will use the following simple principal measure for accessing the performance of the systems:

- *yield*: the average (per topic) number of supported, novel, non-repetitive, important target snippets.

We will also consider other simple measures:

- *success rate*: the number of topics with at least one supported, novel, important target snippet, and
- *overall precision*: the percentage of supported, novel, non-repetitive, important snippets among all submitted snippets.

These choices are considerably “simpler” than the evaluation set-up at today’s TREC QA track, where a type of

series-based scoring is used that involves requires to identify key information nuggets. For our pilot, we prefer the simpler measures listed above—both to keep the assessment load limited and because we believe they are more transparent.

4. CONCLUSIONS

In this paper we have motivated and described WiQA, a new pilot for evaluating exploratory question answering that will be launched at CLEF 2006. By the time of the EESS workshop we should be able to provide some initial results of the pilot.

5. ACKNOWLEDGMENTS

Valentin Jijkoun was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.-001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.-501, and 640.002.501.

6. REFERENCES

- [1] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.
- [2] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In *Proceedings of the Fourteenth ACM conference on Information and knowledge management (CIKM 2005)*. ACM Press, 2005.
- [3] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th intern. conf. on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.
- [4] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 38–53. NIST, 2003.
- [5] R. Soricut and E. Brill. Automatic question answering: Beyond the factoid. In *Proceedings HLT/NAACL*, 2004.
- [6] Wikipedia, 2006. Wikipedia. URL: <http://www.wikipedia.org>.
- [7] WiQA, 2006. Question Answering Using Wikipedia URL: <http://ilps.science.uva.nl/WiQA/>.

Impact of Relevance Intensity in Test Topics on IR Performance in Polyrepresentative Exploratory Search Systems

Berit Lund

Department of Information Studies
Royal School of LIS
Birketinget 6, 2300 Copenhagen S,
Denmark
+45 3258 6066

beritlund@gmail.com

Jesper W. Schneider

Department of Information Studies
Royal School of LIS, Aalborg Branch
Sohngaardsholmsvej 2 – 9000 Aalborg
Denmark
+45 9815 7922

jws@db.dk

Peter Ingwersen

Department of Information Studies
Royal School of LIS
Birketinget 6, 2300 Copenhagen S,
Denmark
+45 3258 6066

pi@db.dk

ABSTRACT

In this paper, we describe the experiments carried out on TREC 5 data concerning whether the number of relevant documents per search task 1) correlates with the number of relevant documents found in overlaps generated from several search engines based on the principle of polyrepresentation; and 2) influences on the performance of the involved systems, measured by Precision and Recall. The first research question investigates 50 TREC 5 topics by combining 12 different search engines involved in the TREC 5 evaluation. The second research issue studies how 30 TREC 5 topics containing as a minimum 45 relevant documents perform over the best performing 4 search engines.

Results show that a correlation indeed exist for the absolute numbers, most pointed when up to 5 engines are combined. Notwithstanding, no significance is detected when numbers of relevant documents per topic are sought correlated with proportions of relevant documents retrieved by any search engine combination. However, the number of relevant documents per search topic influences definitively on the retrieval performance measured by Precision and Recall.

Categories and Subject Descriptors

H.3.3. Information Search and Retrieval

General Terms

Experimentation; TREC topics.

Keywords

Information Retrieval, Polyrepresentation, Exploratory IR, IR Evaluation

1. INTRODUCTION

Exploratory IR systems build partly on the ideas by Bates [1] concerned with the ‘exploratory paradigm’ for information retrieval and information seeking, which led her to propose the search mode of ‘berry-picking [2], partly on the assumptions that not all searchers from the start of a session are capable of providing the system with well-defined representations of their information problem and work task situation [3]. Often, a searcher may simply submit tentative or exploratory keys that evolve throughout the search process. It becomes thus important at the start of a retrieval session that the system retrieves and presents as many potentially relevant documents and perspectives as possible; and – when the searcher has formed a more coherent view of the problem – the system captures those documents that are pertinent to that problem or task. The former initial search mode is largely recall-based while the latter is precision-oriented.

In real-life situations the number of relevant documents per search task is unknown. However, in laboratory tests one should ensure that a sufficient number of relevant documents exist per search task. The foremost reason is that the application of traditional performance measures, like Precision, Recall or Mean Average Precision (MAP) requires a substantial number of relevant items in order to ensure valid performance results [4; 5].

The present paper investigates whether the number of relevant documents per search task 1) correlates with the number of relevant documents found in overlaps generated from several search engines based on the principle of polyrepresentation [6; 7]; and 2) influences on the performance of the involved systems, measured by Precision and Recall. The first research question investigates 50 TREC topics from TREC 5 [8] by combining 12 different search engines involved in the TREC 5 evaluation. The second research issue studies how 30 TREC 5 topics containing as a minimum 45 relevant documents perform over the best performing 4 search engines [9; 10].

The paper is organized as follows. First, the polyrepresentation principle is briefly described and associated to the exploratory search perspective of IR. The methodological aspects of the tests and their results according to the two research questions follow this. A discussion section ends the paper.

Copyright is held by the author/owner(s).

SIGIR'06 Workshop, August 10, 2006, Seattle, Washington, USA.

2. THE POLYREPRESENTATION PRINCIPLE

According to Ingwersen [6] and Ingwersen & Järvelin [3, p. 206] a *principle of polyrepresentation* can be developed as one of several consequences of a cognitive perspective for Interactive Information Retrieval (IIR). Polyrepresentation encompasses cognitively different representations deriving from the interpretations by different actors and functionally different representations that derive from the same actor, such as, author generated text structures, image features, diagram captions, and references or out-links (anchors) [3; 10].

The principle of polyrepresentation for IR is regarded highly precision-oriented. It is based on the following hypothesis: "...the more interpretations of different cognitive and functional nature, based on an IS&R situation, that point to a set of objects in so-called cognitive overlaps, and the more intensely they do so, the higher the probability that such objects are *relevant* (pertinent, useful) to a perceived work task/interest to be solved, the information (need) situation at hand, the topic required, or/and the influencing context of that situation." [3, p. 208]. Such interpretations are commonly taking the form of *representations*, e.g., various ways of indexing the documents in a collection, or different ways of representing a searcher's information situation, e.g., by a request, a problem formulation or a work task description. In the present experiments the overlaps of sets of objects consist of documents retrieved simultaneously by up to 12 different TREC-5 IR engines. Each engine corresponds to a representation of its designer(s)' retrieval ideas and is, in a cognitive sense, thus cognitively different from other engines.

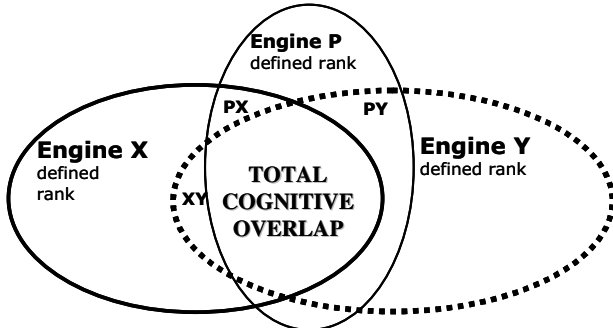


Figure 1. Polyrepresentation of three different search engine's retrieval results in the form of overlapping documents. Variation of [3, p. 347].

The more engines that retrieve a document the higher relevance weight should be assigned that document according to the principle [9; 10]. Fig. 1 illustrates the experimental situation with three different search engines, X; Y; P – and their pair wise overlaps as well as the potential 'total cognitive overlap' formed by all three engines. Polyrepresentation by engines implies more than simplistic data fusion techniques.

Some search engines are more different from each other than others – see for instance the four engines applied in experiment 2, Section 3.2, where two engines are from the same vector space family (SMART-based) whilst one is extended by natural language processing (NLP) features and the fourth engine displays unique

retrieval features. Hence, in experiments not reported here the idea is to give higher weights to documents found in overlaps by very different engines, than to documents in overlaps made from systems of the same family.

In principle one may regard each search engine as providing a *different perspective* to the retrieval result. The so-called 'total cognitive overlap' signifies the documents that comply with many perspectives whereas documents found in overlaps from fewer engines each represents fewer perspectives. In exploratory search systems the more diversified the perspectives of a topic the better the chance that a searcher may encounter a relevant one. Following the polyrepresentation principle the 'total cognitive overlap' may thus on the one hand retrieve potentially interesting documents but – on the other hand – they will be retrieved in small numbers. Consequently, one may wish to supplement the search result by documents from overlaps of fewer engines – hence potentially from fewer perspectives – see for instance overlaps XY, PX or PY – Figure 1. Such experiments are not presented in this paper but are discussed in [9; 10].

With the possibility of quite few documents retrieved in the inner overlaps provided by several search engines, the number of relevant documents per request or topic becomes a cardinal issue for evaluating exploratory search systems based on polyrepresentation.

3. EXPERIMENTAL SETUP

TREC-5 was selected as test bed because of its variance of relevance intensity over its ad-hoc topics. Already from TREC-6 topics having a high frequency of relevant documents, or too few, were discharged from the experiments by NIST [5]. One may thus assume that the TREC-5 topic distribution of relevant documents is rather realistic and may mirror what might happen when translating to natural environments where relevance is not known in advance.

To observe if a correlation exists between the number of documents in the 50 TREC-5 topics assessed relevant and the number of relevant documents retrieved by an increasing number of search engines, the experimental setup was as follows (note that we are not looking into IR performance in this experiment, only correlations). From the TREC-5 performance competition the 12 best performing engines were selected. The top-100 document ID numbers retrieved by each engine were captured from the NIST website for each of the 50 topics. Correspondingly, the ID numbers for all relevant documents per topic were imported. The lists were processed in Access and Excel software.

By means of pivot tables it was then possible for each topic to determine which documents that were retrieved by the search engines. In this way the 12 engines became divided into three groups per topic: a) one group where documents were retrieved exclusively by 2-5 engines. The number of times this took place was aggregated; and groups b) and c): the cases where exclusively 6-9 or 10-12 search engines retrieved the same document, respectively. To be found in an overlap of 6-9 (or 10-12) engines does not imply that a document also counts as found by 2-5 search engines. Indeed, some topics like topic no. 265, Table 1, are having such characteristics that a large number of search engines often simultaneously find identical documents – that also are relevant – but less often documents by combining fewer engines.

The 50 topics were sorted by frequency of relevant documents – see extraction Table 1. The table also includes the corresponding number of retrieved relevant documents per group of search engine overlaps.

3.1 The Second Experiment over Four Engines and 30 TREC-5 Topics

In a second test the four best performing search engines and the 30 topics with the highest number of relevant documents (> 44) were selected. These engines consisted of two versions of the SMART system (based on the vector space model) from Cornell University (Cor5M2rf) and Swiss Federal Institute of Technology (ETH), the former using human relevance feedback for query expansion, a third one mixing natural language processing and vector space with query expansion from an US laboratory group (genrl3), and a fourth engine running on very different principles applying GCL (structured) query language from University of Waterloo (uwgcx1). The former two systems [11; 12] are hence from the same family of retrieval principles, but regarded functionally different, whilst the third [13] is cognitively different from the former two search engines. The fourth engine [14] is cognitively very different from the three other ones.

The topics became divided into three groups, each consisting of 10 topics according to the number of relevant documents per topic. For each group the four retrieval engines were run individually and in all their combinations over its ten topics. Recall and Precision performance measures were applied (Document Cut-off Value, DCV = 100) in order to observe the potential influence of number of relevant documents per topic on IR performance. For each engine the retrieved documents were assigned an artificial ranking weight, independent from the involved search engines’ own output values; in the case of document overlaps additional weights were added to the documents depending on the number of overlaps in which they were found [9; 10]. So, the higher the retrieval intensity (that is, the number of engines retrieving a document) the higher the weight added.

4. RESULTS

With respect to the first experiment Table 1 displays the top-25 TREC-5 topics sorted by highest frequency of relevant documents.

4.1 The First Experiment – Absolute Relevance Intensity

Figure 2 illustrates the scatter of the rank distribution of the number of relevant documents per topic compared to the OL 2-5 rank distribution of retrieved documents over all 50 topics. The X-axis represents the ranks of the relevant documents per TREC-5 topic, based on the *absolute numbers* (Column 2, Table 1) and the Y-axis corresponds to the ranks from column 3.

Evidently, there exists a *strong significant correlation* (Spearman non-parametric, two-tailed $r = .929$, at $\alpha = .050$; $r^2 = .86$) $CV = .34$) between the two distributions – Figure 2. The results demonstrate that the correlation drops for the OL 6-9, although the Spearman coefficient is still significant and far above the critical value (CV) ($r = .614$; $\alpha = .050$). At OL 10-12 the correlation using absolute numbers becomes statistically insignificant ($r = .209$). These scattergrams are not shown.

Table 1. Number of retrieved relevant documents from TREC-5 ad-hoc track over 25 topics in 2-5, 6-9 and 10-12 overlapping engines (OL), respectively – sorted by column 2.

Topic	# Rel. doc. per topic	OL 2-5	OL 6-9	OL 10-12
269	594	100	3	0
251	579	93	4	0
273	513	100	31	19
291	407	59	0	0
264	281	38	4	1
285	261	84	48	14
294	160	52	16	2
265	147	21	20	70
286	142	36	40	13
289	141	22	19	5
266	139	40	17	7
257	135	34	14	5
282	131	21	9	9
274	119	32	23	33
290	119	36	2	0
270	116	11	17	43
258	115	41	20	7
255	109	21	7	1
288	92	40	17	18
298	91	20	24	6
261	87	23	11	20
271	86	20	15	14
297	86	25	11	33
254	85	14	13	16
283	84	30	20	3

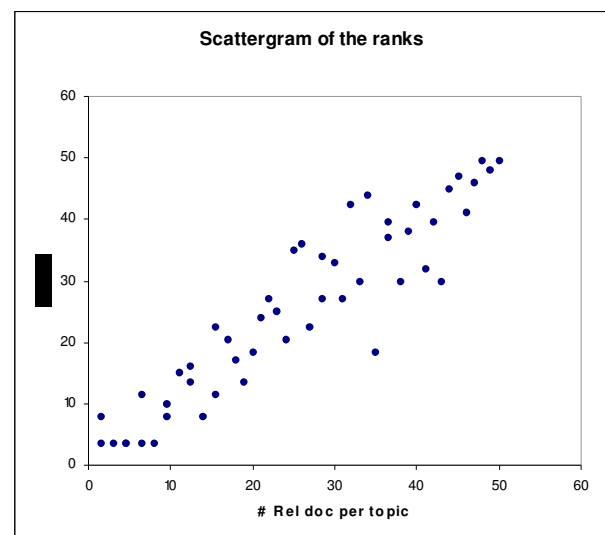


Figure 2. Scattergram of the rank distribution of number of relevant documents and the ranks of the retrieved relevant documents per TREC-5 topic retrieved by 2-5 search engines.

4.2 The First Experiment – Proportional Relevance

The reason for doing this analysis is that one would like the strong linear correlation, Figure 2, also to be observed for the proportion of relevant documents retrieved: the hypothesis would be that with an increasing number of relevant documents per topic, one obtains not only 'more' retrieved relevant documents (in absolute numbers) – but that the proportion of relevant retrieved documents is the *same* over all topics – irrespective of the actual number of relevant documents per topic.

However, the scatter and correlation coefficients shift dramatically when the proportion of the retrieved relevant documents over relevant documents per topic is taken into account. Figure 3 demonstrates this kind of scattergram, which rejects the hypothesis.

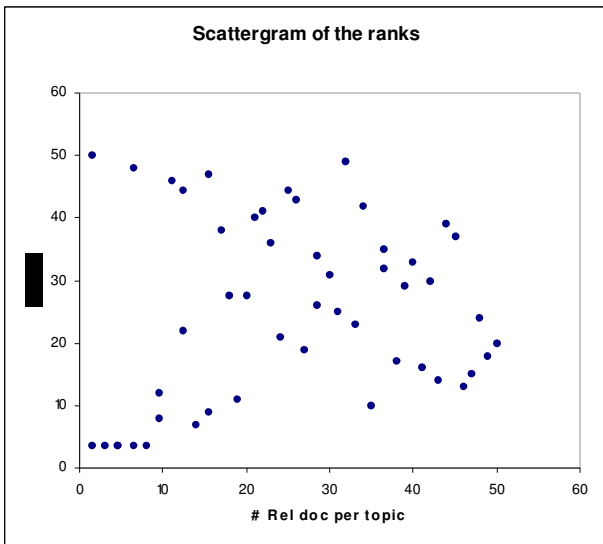


Figure 3. Scattergram of the rank distribution of number of relevant documents and the ranks of the proportion of retrieved relevant documents per TREC-5 topic retrieved by 2-5 search engines.

This phenomenon can also be observed on Table 1, Topic 269 etc. Dividing 100 retrieved relevant documents by 594 relevant ones (proportion = .168) does not provide a value close to those of the preceding topics with decreasing frequency of relevant documents: topics 251, 273 or 291, etc. (= .160; .194; or .144; etc.). The Spearman coefficient for the OL 2-5 is .142 (not significant at $\alpha = .050$), and with increased number of combined engines the correlation becomes increasingly negative.

4.3 The Second Experiment

The second experiment is carried out on the retrieval performance of the 30 TREC-5 topics with most relevant documents. It shows that Precision and Recall depend on how many relevant documents potentially exist in the search tasks. Table 2 demonstrates that Precision increases with increased number of relevant documents in the tested topic groups over the four engines in all their combinations. Correspondingly Recall decreases. The Middle intensity group of 10 topics approximates the Recall and Precision values for all 30 TREC-5 topics.

Table 2. Precision and Recall value ranges for the high intensity group of topics (> 140 rel. docs./topic), the middle intensity group (> 90 rel. docs./topic), low intensity group (> 44 rel. docs./topic) and for all 30 TREC-5 topics (DCV = 100). Each group consists of 10 topics.

Topic Intensity	High	Middle	Low	All 30 topics
Precision	.51 - .65	.36 - .50	.28 - .34	.38 - .48
Recall	.19 - .23	.31 - .38	.36 - .43	.29 - .34

5. DISCUSSION

The results from the correlation tests imply that indeed 1) there exist a correlation between the number of documents assessed relevant in test topics and the number retrieved by different search engines combined: by combining an increasing number of search engines the correlation becomes increasingly non-linear and unpredictable owing to interference from unknown variables, such as the nature of the topics themselves. By combining few search engines (2-5) the resulting overlaps produce absolute numbers of relevant documents *in accordance* with the known number of relevant documents (87 % of the correlation accounted for).

However, it is *not* possible to *predict* the proportion of relevant documents retrieved per topic. This is a different facet of the nature of topics in collections. The two contradictory results for the same types of polyrepresentative 'cognitive' overlaps indicate that in investigative practice we may be able to measuring performance realistically without knowing in advance the relevance frequencies. On the other hand we are *not in control* of the proportion of documents used for such measurements. Predictions cannot be made. In polyrepresentative (laboratory) evaluation experiments the number of documents judged relevant (the relevance intensity) becomes thus critical – as does the number of topics – since we commonly require a substantial number of topics as a minimum (> 25) in order to ensure statistical validity in such experiments. This is not always realistically possible in non-laboratory field investigations or experiments with human test persons [3]. Instead such investigations should employ a substantial number of test persons and as many 'open assignments' or 'simulated IR or work task situations' [15] as ergonomically possible. The interpretations (or explorations of the situation) made by the persons may then each count as a different 'topic' in the sense of TREC laboratory tests. In short: many search task situations are required in order to ensure enough relevance intensity associated with the tasks.

The results from the second experiment on retrieval performance imply that when applying polyrepresentation of topical perspectives for exploratory searching via the best performing engines in concert the IR performance results depend on the number of documents assessed relevant in the involved search tasks. The more relevant documents in search tasks the higher the performance. This outcome may not seem surprising since polyrepresentation is favoured by the existence of a substantial amount of relevant documents, owing to its precision-oriented nature. However, in case of comparisons between retrieval made by polyrepresentation principles and other IR techniques, search task characteristics – such as number of relevant documents – may heavily influence the evaluation.

6. ACKNOWLEDGMENTS

We wish to thank Donna Harman and NIST for allowing us to make use of the TREC-5 data, and Tove Faber Frandsen for her conceptual contribution to research question 1. The NORdic Research School of Library and Information Science (NORSLIS) support this research by a travel grant.

7. REFERENCES

- [1] Bates, M.J. (1986). An exploratory paradigm for online information retrieval. In: Brookes, B.C. (Ed.) *Intelligent Information Systems for the Information Society: Proceedings of the IRFIS 6 Conference*, Frascati, Italy, 1985. Amsterdam, NL: North-Holland: 91-99.
- [2] Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5): 407-424.
- [3] Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.
- [4] Sormunen, E. (2002). Liberal relevance criteria of TREC – counting negligible documents? In: *Proceeding of the 24th ACM-SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press: 324-330.
- [5] Voorhees, E. & Harman, D. (1998). Overview of the Sixth Text Retrieval Conference (TREC-6). In: *Proceedings of the Sixth Text Retrieval Conference*. NIST Special Publication 500-240. Available on: http://trec.nist.gov/pubs/trec6/t6_proceedings.html
- [6] Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1): 3-50.
- [7] Larsen, B. & Ingwersen, P. (2005). Cognitive overlaps along the Polyrepresentation Continuum. In: Spink, A. & Cole C. (eds.), *New Directions in Cognitive Information Retrieval*. Springer: 43-60.
- [8] Voorhees, E. & Harman, D. (1997). Overview of the Fifth Text Retrieval Conference (TREC-5). In: *Proceedings of the Fifth Text Retrieval Conference*. NIST Special Publication 500-238. Available on: http://trec.nist.gov/pubs/trec5/t5_proceedings.html
- [9] Lund, B. (2005). Polyrepræsentation og Datafusion: Test af Teorien om Polyrepræsentation gennem forsøg med fusion af TREC-5 Resultater. Danmarks Biblioteksskole, 2005 (MSc Thesis).
- [10] Larsen, B., Ingwersen, P. Kekalainen, J. (2006). The Polyrepresentation Continuum in IR. Paper submitted to the First IiX Symposium, Copenhagen, October 18-20, 2006.
- [11] Buckley, C., Singhal, A. & Cormack, G.V. (1997). Using query zoning and correlation within SMART: TREC-5. In: *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*: 105-118. Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html (April 6, 2006).
- [12] Ballerini, J.P., Büchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P. & Wechsler, M. SPIDER retrieval system at TREC-5. In: *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*. Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html (April 6, 2006).
- [13] Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin Fang, Perez-Carballo, J. Straszheim, T., Wang Jin & Wilding, J. Natural language information retrieval: TREC-5 report. *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*: 291-314. Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html (April 6, 2006).
- [14] Clarke, C.L.A. & Cormack, G.V. (1996). Interactive substring retrieval (Multitext retrieval for TREC-5). In: *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*: 105-118. Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html (April 6, 2006).
- [15] Borlund, P. (2003b). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), paper no. 152. Available at: <http://informationr.net/ir/8-3/paper152.html>. (May 13, 2006).

From Question Answering to Visual Exploration

David McColgin
Pacific Northwest
National Laboratory
902 Battelle Blvd
Richland, WA. 99354
509-375-2148

Michelle Gregory
Pacific Northwest
National Laboratory
902 Battelle Blvd
Richland, WA. 99354
509-375-2824

Elizabeth Hetzler
Pacific Northwest
National Laboratory
902 Battelle Blvd
Richland, WA. 99354
509-375-6698

Alan Turner
Pacific Northwest
National Laboratory
902 Battelle Blvd
Richland, WA. 99354
509-375-6670

{david.mccolgin;michelle.gregory;beth.hetzler;alan.turner}@pnl.gov

ABSTRACT

Success in Question Answering has been traditionally measured by precision and recall, which are good metrics for identifying specific best answer(s) that might be obtained by a lookup type of search. These metrics do not address the many information gathering techniques in exploratory interactions. In this paper, we present an integrated Question Answering environment that combines a visual analytics tool with state-of-the-art query expansion, and complements the cognitive processes associated with an information analyst's work flow. In our system, questions result in a comprehensive answer space that allows users to explore the variety within the answers and spot related information in the rest of the data. The exploratory nature of the dialog between the user and this system requires tailored evaluation methods that better address the evolving user goals and counter cognitive biases inherent to exploratory search tasks.

Categories and Subject Descriptors

I.3.6 [Computer Graphics]: Methodology and Techniques – Interaction techniques, I.6.9 [Visualization] – Information Visualization, Visualization Techniques and Methodologies, I.7.5 [Document Capture] – Document analysis, H.5.2 [User Interfaces] – Evaluation/methodology, H.3.3 [Information Search and Retrieval] – Search process

General Terms

Design, Human Factors.

Keywords

Information Visualization, User interaction design, exploratory search, evaluation.

1. INTRODUCTION

Marchionini defines three types of search: lookup, search to learn, and investigative [6]. While returning factoid answers satisfies many search needs, the information needs of an information analyst require an investigative approach. In this paper we present an integrated Question Answering (QA) system that combines state-of-the-art query expansion [2] with a document visualization tool, INSPIRE [3]. In this system, users query a document space with a natural language question that is expanded and optionally edited by

the user. Queries result in the identification of relevant passages and the selection of matching documents within the context of the whole document set. This approach leads to a sophisticated dialog in which the user can explore the QA results and maximize understanding of the data before reading individual documents and without relying solely on retrieved passages. The advantage of this analysis environment lies not in the power of any one visualization or tool, but in the process supported by using them in concert. With improved understanding of the answer space, users can better form new questions, detect answer patterns, or select the most interesting documents to read in detail. The system we present here supports analysts' goals by helping to identify the presence of conflicting data, data from other sources, answer patterns (e.g. geographical or temporal), and even information on other topics not returned by the query but potentially relevant. The evolving information needs of the analysts require system evaluation metrics that go beyond precision. In this paper, we discuss the information needs of analysts and use a work flow scenario to present our exploratory system. We report on initial formative evaluation of the system and conclude with a discussion of formal, summative evaluation metrics.

2. THE ANALYST'S TASK

Information analysts spend much of their time foraging complex and contradictory bodies of information in support of their ultimate reasoning goals. They are seeking detailed knowledge of specific facts that can 1. support or refute candidate positions on the subject they are investigating; 2. allow them to credibly identify and bridge the gaps in their knowledge, and; 3. discover previously unknown evidence and relationships. As domain experts on the topics that they are exploring, their goal is not to simply isolate "the best" facts, but rather to explore new dimensions of the data and arrive at reasoned and supportable conclusions [9]. They perform these tasks under significant pressures and constraints including time limits, the required form of their output (e.g., a verbal briefing, a written report, the length of the report, etc.), often unfamiliar topics and great uncertainty, and information sources of variable accuracy.

An exploratory system for information analysts must maximize data understanding within the level of domain knowledge and multiple constraints on the working conditions. In addition, such a system also needs to help overcome the potential cognitive pitfalls of analytical work under pressure such as satisficing, anchoring, vividness, and oversensitivity to consistency [1,4].

QA and interactive query expansion within a visual analytics environment offer the chance to counteract such biases. Instead of querying, interpreting limited results, and querying again, the analyst is presented with a comprehensive visual answer space that can be interactively explored. Variations in the extracted

Copyright is held by the author/owner(s).

SIGIR'06 Workshop, August 10, 2006, Seattle, Washington, USA.

answer passages, contextual information about other documents in the collection, and patterns in the answers across time, source, or theme reveal alternative explanations and unanticipated influential factors.

3. IN-SPIRE

IN-SPIRE is a visual analytics tool developed by Pacific Northwest National Laboratory to facilitate rapid understanding of large textual corpora [3]. IN-SPIRE generates mathematical signatures for each document in a set. Document signatures are clustered according to common themes to enable information exploration and visualizations. Information is presented to the user using several visual metaphors to expose different facets of the textual data. The central visual metaphor is a Galaxy view of the documents as clustered dots that allows users to intuitively interact with thousands of documents, examining them by theme (Figure 1). The Galaxy has been shown to provide value beyond traditional retrieval systems [3]. While the concept of cluster projection is not new, the current line of research is exploring its value within a larger visual QA process. Additional analytic tools allow exploration of temporal trends, thematic distribution by source or other metadata, and query relationships and overlaps.

QA functionality is being integrated within IN-SPIRE so that users can explore specific questions within the massive data collections. Query expansion is provided by Language Computer Corporation’s FERRET application [2]. The interface is incorporated into IN-SPIRE within its query tool. Users can ask questions in natural language and FERRET finds answer passages as well as returning expanded queries in Boolean and Query by Example (QBE) syntax for use in IN-SPIRE. Users have direct access to the output and can edit or add terms to the query. Queries can search the whole dataset or only the currently selected documents to help refine an information need. In the next section, we demonstrate these capabilities through a sample scenario.

4. WORK FLOW SCENARIO

An analyst is given the task to determine the largest environmental threats posed by nuclear power. Given the plethora of avenues one can use to form hypotheses on this topic, simple searches with factoid answers are not adequate to explore all relevant details from the data. The first step in the work flow is exploring the document collection within which the analyst will work. The dataset can be opened in the Galaxy view (see Figure 1), which allows her to assess the size and thematic coverage of the collection.

The task could be approached in a variety of ways but she would first like an historical perspective. She asks “Which nuclear reactors had the worst safety incidents?” in IN-SPIRE’s Query tool. FERRET returns documents with answer passages, giving her concise facts about her question and helping to guide her subsequent investigation (Table 1).

A malfunctioning control rod caused the shut down of Zaporizhzhya-4 on 13 April 1997. [1, 2] According to a **plant** spokeswoman, one of the 61 control rods used to moderate **nuclear** activity failed to descend into the **reactor** core within the time allowed by regulations.

Table 1. Sample answer.

Expanded queries are also returned below the original question in both Boolean and vector-based forms (Table 2). Together, the two expansions provide syntactic expansion, top-down semantic expansion based on external sources like WordNet, and bottom-up instances of related terms from the data itself.

(reactors OR "nuclear reactor" OR reactor) AND (worst OR defective OR risky) AND (incidents OR matter OR event OR incident) AND (safety OR safe OR guard) AND (nuclear OR atomic)

Table 2. Sample Boolean expansion.

Analytically, the query expansions help to start a dialog that provides additional insight and context. The analyst reviews the expansions and decides to use the Boolean query. She has the chance to remove undesired terms, change the Boolean logic, and add concepts of her own before executing.

4.1 Galaxy View

Her results appear ordered by relevance in the Document Viewer where she can access the title and full text. Results are also displayed in the Galaxy in the context of the entire collection. The clusters and labels help her gain valuable insight into the content of the query results without having to read each document individually or rely solely on the documents that match her query. The thematic view helps her to identify and eliminate irrelevant results, refine her information need, or find a new facet of the answers worth exploring. Nearby labels describe the related topics and nearby documents contain related material that the query results alone would not have provided. In this case, the analyst decides to visualize the result documents alone (Figure 1).

When the Galaxy is recalculated to show only the results, she begins to investigate the clusters. Recognizing the name Chernobyl she first scans documents in clusters with that label and finds information about several specific incidents, their effects, and the international response. Investigating the clusters to the right labeled “integrum, mayak, ctr” and “launchers, ss, india”, she gains insight into an unexpected risk; these documents contain information on accidents and radiation leaks from military vehicles such as ships and submarines, the threat posed to marine environments, and the remediation efforts and methods in use. While they do not directly answer her question, these documents provide complementary information about environmental safety. When she examines the documents clustered at the bottom of the visualization, another theme emerges. In contrast to the clusters above, these documents are primarily about new reactors and development programs with much discussion on new safety technologies and protocols that could mitigate the risks and effects of future incidents.

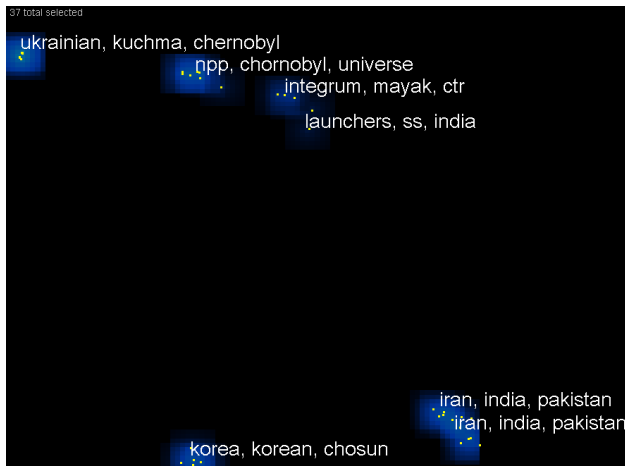


Figure 1: Galaxy view recalculated to include only the 37 query results.

4.2 Correlation Tool

The analyst can group any set of selected documents, whether they are selected manually or by virtue of matching a query. In this example, she has made many groups based on the answers, Boolean query, selections from the other tools, and independently determined groups such as countries. The Correlation tool allows her to explore the overlap between the groups she has created. Figure 2 shows her query results (y-axis) distributed over countries (x-axis). As with all of the tools, Correlation is linked to the visualizations, so that clicking a column here results in a selection in the other tools.

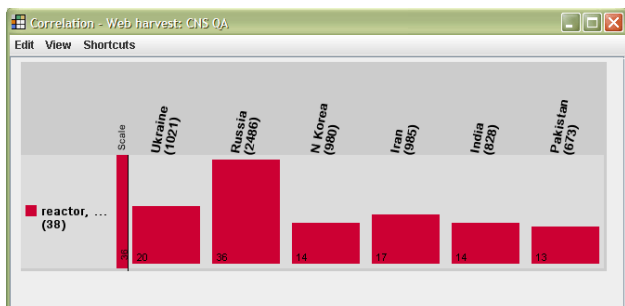


Figure 2: Distribution of expanded query results by country.

4.3 Additional Tools

There are a variety of other interactive tools and visualizations that can help analysts investigate and gain insight in IN-SPIRE. Affect, trends and salient events in time, and other data attributes can be measured, portrayed, and used interactively. Hypothesis tracking and outlier evolution are also explicitly supported.

4.4 Review

By now, the analyst has an overall sense of the data from the visualization, which helped her to formulate an initial question. She was presented with extracted answers that helped her refine the system's query expansions to her interests. Portraying query results in context helped her find useful non-hits that could be important. Seeing variability within her results allowed her to find unanticipated relevant information: an unexpected type of risk and information on modern safety improvements. She was able to see

the country with the most content about safety incidents using the Correlation tool, and see differences over time. Certainly, the process helps the analyst find facts and discover evidence and relationships. But it also exposes facets of the answer space that influence her subsequent interactions and helps to further define her information need. The analyst may continue exploring by refining her question, starting a new line of inquiry based on her acquired knowledge, or performing deeper analysis with the other tools.

5. EVALUATION

The challenge of evaluating exploratory search systems shares many of the challenges of evaluating visual analytics systems in general. Certainly, *usability* is one part of the solution, including quantitative measures, such as time and errors, and qualitative measures, such as user satisfaction. In the case of our tool, formative usability evaluation with analysts has helped reinforce our main direction while suggesting specific improvements, such as additional kinds of user interaction.

Several approaches to evaluating *utility* also provide merit, although exploratory utility is harder to assess given the lack of solid "correct answers." NIST used quality of users' written analysis reports as one metric to evaluate the system used in creating the report [5]. Contests such as those run by the InfoVis Symposium judge systems based on the ability of the tool to interactively reveal insights into the data. [8] We have also found that having a tool developer or designer work together with a user to carry out an analysis task can be an excellent way not only to assess the potential utility of the system, but also to sharpen the perception of user needs.

We propose that a good exploratory system should encourage *sound usage strategies*, and are researching an approach where this goal serves as the basis of evaluation. For example, an experiment conducted by Patterson et al. identified searching behaviors that led to exclusion of key documents, correlating very well with errors in users' verbal reports [7]. Typically, users started with a broad search and then progressively narrowed it to reduce the number of hits to a reasonable level, often excluding key documents without realizing it. In addition to the obvious metric of how many key documents were found, several complementary metrics could provide insight:

- 1) How many search paths did the user try? It's routinely easy for users to add terms to a previous query, often to narrow the results; fewer systems make it easy to try a new tack or combine multiple strategies.
- 2) Of the key documents found, how many were recognized as important? This is a subtle question, aimed at assessing a system's capabilities to quickly help users assess the value of documents. Many systems provide metadata, such as year and source, or fragments of text to help with this assessment. Still, the daunting task of skimming tens or hundreds of such fragments may lead users to quickly resort to a new smaller search. What kinds of clues are needed to ensure that once a key document is located (e.g., by a search), it is actually recognized and not discarded?

- 3) How many of the key documents were actually considered in the user's decisions? In the midst of information overload, users may easily forget details of specific documents. Exploratory systems provide a challenge in this regard, as the information tasks often follow unexpected paths. This metric is aimed at evaluating how well a system supports retention and use of important discovered information.

Another facet of sound usage is the ability of a system to help *counter user bias*. Exploratory search systems are inherently a partnership between user and system, and ideally should utilize the strengths of each to compensate for the weaknesses. As discussed in Section 2, users carrying out investigative search are vulnerable to a number of cognitive biases. In contrast to a lookup search task for which there may be a single best answer, an investigative task involves identification and consideration of multiple alternative answers. One example bias is anchoring, where a user's initial judgment or estimate of the answer unduly influences evaluation of subsequent evidence [4]. Metrics related to this bias might include:

- 1) How many alternatives did the user explore? This question tries to go beyond the simple identification of alternatives to assess whether the user spent time actually investigating more than one explanation. A system's ability to support and track multiple alternatives can make this task easier, hopefully leading to more in-depth investigations by users.
- 2) How much credence did the user give to counter evidence? This question aims at one of the aspects of anchoring, that users will discount evidence contrary to a chosen explanation.

While enticing, time as a metric can be misleading in this context. A system that helps users more quickly come to a conclusion might also contribute to anchoring or satisficing rather than helping to counter them.

Exploratory systems can provide great value to users in many fields. While recognizing the value of usability and utility measures, we propose that metrics be developed based on sound usage strategies and the combination of user/system capabilities to counter weaknesses in each.

6. ACKNOWLEDGMENTS

IN-SPIRE developers Jon McCall and TJ Hoeft were key contributors to the technical implementation. Our thanks also to the Language Computer Corporation and especially Sanda

Harabagiu, John Lehman, and Patrick Wang for their collaboration on integrating FERRET.

7. REFERENCES

- [1] Connaway, L.S., Prabha, C., & Dervin B. An overview of the IMLS Project "Sense-making the information confluence: The whys and hows of college and university user satisficing of information needs". Presented at Library of Congress Forum, American Library Association Midwinter Conference, Boston, MA, Jan 16, 2005.
- [2] Harabagiu, S., Hickl, A., Lehmann, J., & Moldovan, D.. Experiments with interactive question-answering. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05) (Ann Arbor, June 25-30, 2005). Association for Computational Linguistics, New Brunswick, NJ, 2005, 205-215.
- [3] Hetzler, E. and Turner A. Analysis experiences using information visualization. IEEE Computer Graphics and Applications, 24:5, 2004.
- [4] Heuer, R.A. *Psychology of Intelligence Analysis*. Government Printing Office, Pittsburgh, PA, 1999.
- [5] Kelly, D., Kantor, P.B., Morse, E.L., Scholtz, J., Sun, Y. User-centered evaluation of interactive question answering systems. In Proceedings of the Interactive Question Answering Workshop at HLT-NAACL (New York City, NY, June 8-9, 2006). Association for Computational Linguistics, Stroudsburg, PA, 2006 49-56.
- [6] Marchionini, G. Exploratory search: from finding to understanding. Communications of the ACM. 49:4 2006, 41-46.
- [7] Patterson, E.S., Woods, D.D., Tinapple, D., Roth, E.M., Finley, J.M., and Kuperman, G.G. Aiding the Intelligence Analyst in Situations of Data Overload: From Problem Definition to Design Concept Exploration. Technical Report ERGO-CSEL 01-TR-01, Inst. for Ergonomics/Cognitive Systems Engineering Lab, Columbus, OH, 2001.
- [8] Plaisant, C. The challenge of information visualization evaluation. In Proceedings of the Working Conference on Advanced Visual Interfaces. ACM Press, New York. 2004, 109-116.
- [9] Thomas, J. and Cook, K.A. (Eds.). *Illuminating the Path--Research and Development Agenda for Visual Analytics*. IEEE Press, Los Alamitos, CA, 2005.

What do the Attributes of Exploratory Search Tell us about Evaluation

Yan Qu, George W. Furnas
School of Information
University of Michigan
Ann Arbor, MI, 48109, USA
yqu, furnas@umich.edu

ABSTRACT

Evaluation of exploratory search systems could be informed by the examination of special attributes of exploratory search. In this paper, two attributes of exploratory search are described: 1) the tight coupling between search and other information activities, and 2) the gradual growth of task structure representation. Their implications for designing, and therefore evaluating, exploratory search systems are discussed.

1. INTRODUCTION

Search is an important way to access the proliferating information on the Web. Recently, research interest has increased on studying “exploratory search” [4][6], where the user is not looking up specific information, but instead tries to learn and investigate more broadly. In exploratory search, a person may have a general information need (e.g., I want to know more about classic music, I want to know why my computer has a blue screen). However she does not have enough knowledge to form a specific query or to navigate through an information space. In many cases, the person pursues information she knows little about yet. Search is used to help explore the information space.

Researchers from various disciplines such as Information Retrieval, Library and Information Science, and Human-Computer Interaction have started laying out the design space for systems that support exploratory search [6]. Evaluation plays an important role in this process, guiding the iterative design that improves the systems we create. We believe that both the rich structure of exploratory search activity, and the early stage of the field’s understanding of it, argue against any simple evaluation metrics (like precision and recall), and in favor of more qualitatively rich approaches. These more formative evaluations, rather than narrow summative ones, will help us understand aspects of the interaction of the users’ behavior and system design more richly. Attention should be given to comprehensive qualitative evaluation methods that reveal and appraise components of the exploratory search process and how different factors in system design affect the process. This should be much more valuable than some generic, performance metrics with one or two numbers as the result.

Our discussion on evaluating exploratory search systems follows this concern, as we examine several attributes of exploratory search and discuss their implications in evaluation.

We first explain how people face basic uncertainties common to exploratory search situations, such as unclear information needs and lack of knowledge about the information space. These in turn reveal two important attributes of exploratory search activities. Then we explain why system level evaluation is more appropriate for exploratory search system and how the special attributes of exploratory search activity give insights in both system design and evaluation.

2. EXPLORATORY SEARCH

Most search activity entails some degree of uncertainty about the user’s information need and the information space. We believe that as these uncertainties become greater, search must become more exploratory. The complex, iterative, interactive and situated behavior of exploratory search is required to reduce these uncertainties.

In exploratory search, the unclear information need and the lack of knowledge about the information space prevents people from formulating specific queries that could retrieve useful information directly. Querying still plays a role, it is used to probe the information space to see what useful information is available, it can serve as a starting point for explorations, but is only part of a process. To illustrate, we examine two strategies used in exploratory search, *Query Initiated Browsing* and *Query Initiated Analysis*. The first of these is familiar in most use of modern search engines on the web, the second is being explored in various prototype information environments.

Query Initiated Browsing [3]. Although we often ignore the distinction, it is important to note that, unlike classical IR systems, Web search engines do not return *pages* relevant to the query, they return *pointers* to those *pages embedded in the web*. As such the pointers are gateways for extended exploration. From the nominal “returned page”, people move on to explore the adjacent area using the navigation structures built within the information space (labeled hyperlinks). For example, the person diagnosing her frozen PC queries Google with “blue screen”, then follows one search result to the Wikipedia website where she reads about different types of blue screens, and later follows various links to computer troubleshooting websites to learn more about this issue. That the search results are not pages but gateways not only increases the probability of encountering useful information, but allows the user to see how the relevant topic is organized. This in turn helps them understand both their own problem and the information resources that are available.

Query Initiated Analysis. In another sort of exploratory search, users employ search to filter out a set of data for further investigation with various analysis tools. For example, techniques such as clustering, classification and visualization are often used on query result to reveal useful structure or patterns in the data

[1][7]. In our blue screen diagnosis scenario, the person could search “blue screen”, and use clustering and visualization techniques to analyze the information set filtered out by the query. She may find a cluster on “Blue Screen of Death” which is worth further investigation.

Additionally, although people often start exploratory search without clear information needs and full fledged task plans (subtasks, steps, etc.), their information needs are gradually clarified and their task plans are gradually formed by improvisation during exploratory search. A person without a computer science background who tries to diagnose a sudden blue screen may not have a clear information need other than asking “what’s wrong with my computer”. She also does not have a clear idea about how many steps the diagnosis will consist of, or how many different things (Hardware? Software?) she needs to check. Her first foray into the Web leads her to the wikipedia webpage on blue screen, where she learns about types of blue screen and how to interpret the error message. With that knowledge, she decides on the first step of the diagnosis: search and learn about the specific error message on her blue screen. As the exploration continues, the information need will become more clear to enable more specific queries and the diagnostic plan will gradually emerge.

This discussion has revealed two attributes of exploratory search activity that may lead to interesting implications in design and evaluation of systems for exploratory search: 1) the tight coupling between search and other information exploration activities, namely browsing and analysis, and 2) the gradual growth of task structure representation (e.g., the diagnostic plan in the blue screen scenario). In the next section, we will show how these attributes give insights into the kind of evaluation needed.

3. EVALUATION OF EXPLORATORY SEARCH

When evaluating an exploratory search system, the standard precision+recall (P&R) metric for Information Retrieval algorithms are not sufficient for several reasons. First, reasonable P&R is a necessary but not sufficient condition for the success of querying in exploration. For example, in *Query Initiated Browsing*, reasonable precision and recall may help get the user to a good starting place, however that is only beginning of the process, and the evaluation has to reflect the efficacy of the larger activity. Second, the success of an exploratory search system depends on the overall design of the system rather than the search algorithm alone. The efficiencies of different information activities and how well the system supports the integration of the activities must also be considered.

There have been decades of work on information system evaluation. These gives us a general guidelines that apply to evaluating exploratory search system. However, specific attributes of exploratory search may require special emphasis or particular methods in evaluation. In this section, we explore this issue by inspecting the two attributes of exploratory search mentioned in the last section and discuss how they provide special implications for system design and evaluation.

Attribute 1: Tight coupling between search and other information activities

When search is used in exploration, it is tightly coupled with other information activities, such as browsing, organization,

visualization, etc. In *Query Initiated Browsing*, querying precedes browsing. In *Query Initiated Analysis*, querying precedes clustering/classification, comparison, visualization, etc. on query results. At the same time, search could follow any activities that create/refine/clarify information needs. We draw two implications for design and evaluation from this attribute.

The first implication is fairly generic – support the activity shifts. That is, the shifts between querying and other information activities should be emphasized in design and evaluation. To integrate search with other information activities, designers should understand how search is used in the exploratory tasks, such as the *Query Initiated Browsing* and *Query Initiated Analysis* strategies mentioned above. Which activities could proceed or follow querying? Which transitions between activities are most likely. Designers’ understanding of the information flow in the task will greatly influence their design decisions. For example, if users always need to draw concept maps of search results for a certain task, then tools can be designed to streamline the move to such mapping, e.g., showing the search results in clusters or building a pipeline that sends the search result for automatic entity or topic analysis. On the other hand, if users embrace a broad variety of post-query activities, the system should show different choices to the user and make easy and smooth the shift to whichever activity is chosen next.

In the implication for evaluation is that the shift between different information activities should be analyzed, and the ease of the transitions should be measured. Transition diagrams could be drawn to show the dynamic relationships between various activities. The cost of the shifts should be measured, for example by the time and number of operations needed, and the cognitive load required. The overall ease of the transition not only depends on the cost of each individual shifts, but also on how well the cost structure of the system matches the cost structure of the task, more frequent transitions should be lower cost.

Second, beyond supporting smooth transitions between components, designers and evaluators need to rethink the components themselves to provide greater synergy. For example, in order to support *Query Initiated Browsing*, the search functionality should care not only about the relevance of the returned information, but also whether the results point to places in the information space that supporting browsing activities. Understanding a retrieved page as a gateway and the beginning of a subsequent navigation process argues that a more complex suite of P&R measures should be calculated, reflecting not just individual pages, but the navigable neighborhoods around them. A page is usually more useful if the relevant topic continues throughout the neighborhood. The success of the subsequent exploration also depends on other factors, including the traversability and navigability of that neighborhood [2].

Attribute 2: Gradual growth of task structure representation in exploratory search

Instead of having a pre-existing representation of task structure (sub-tasks, steps, etc. in a user’s mind or externalized in computer systems) and using that representation to direct actions throughout the whole task, users gradually grow and change their task structure representation in exploratory search. Imagine how the person confronting the blue screen learns about steps and tests in diagnosis during the exploration. The evolution of the task structure representation could be illustrated by Russell et al’s

sensemaking model [5] (Figure 1) (exploration is often a sensemaking process). A user starts with a preliminary representation of the task structure which is a rough idea about the problem they are facing and actions to take. Unless the user is very lucky, their first representation is not sufficient to finish the task. The failings of the representation (called “Residue”, stuff that does not fit) motivate her to seek another representation better suited to support the accomplishment of the task (Search for good representation). The representation continues to grow in the exploration process as the user gradually learns an appropriate decomposition of the task or the steps to take in the task. In the blue screen diagnosis example, the user begins with only a rough representation of the task structure (“search for information on how to diagnose blue screen”). During the exploration process, she learns about steps and tests to take to accomplish the task. Sub-tasks, steps and tests are gradually added to her increasingly elaborate representation of the task structure.

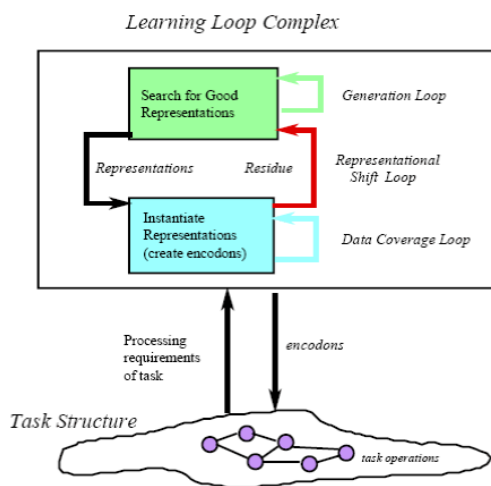


Figure 1. Evolvement of user’s representation on task structure in sensemaking processes ([Russell et al, 1993])

This sensemaking aspect of exploratory search suggests that an exploratory system should support the growth and change of a user’s representation of task structure. More concretely, the system should not have a rigid task or process model. Instead, improvised change of task structure representation should be allowed, such as the change in process steps or task decomposition.

In evaluation, we first need new methods to analyze the shift of task structure representation which may not be directly observable (when the representation is in people’s mind). Think aloud protocols, interview methods, and system logs could be used to extract task structure representation at different stages in the exploration. The goal of the analysis is to understand when and how the shifts happen and obtain insight in design perspectives that influence the representation shift. After that, new metrics are

needed to measure how well the system supports the change of task structure representation. One possible approach could be the cost analysis of the representation change from both cognitive and operational perspective, e.g. what is the cognitive cost of learning about a new diagnostic test using the system; what is the operational cost of changing the diagnostic plan in the system.

4. CONCLUSION

This paper has discussed how, in exploratory search, search is interwoven with other exploratory processes, including browsing, analysis and representation evolution. The richness of these users’ activities creates a rich design challenge. A fundamental purpose of evaluation is to guide design. It follows that the design for rich exploratory search activity must, in turn, be guided by adequately rich evaluation methods that reveal how well all the various aspects of the users’ activity are being supported.

REFERENCES

- [1] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. 1992. Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Copenhagen, Denmark, June 21 - 24, 1992). N. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. SIGIR '92. ACM Press, New York, NY, 318-329.
- [2] Furnas, G. W. 1997. Effective view navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, United States, March 22 - 27, 1997). S. Pemberton, Ed. CHI '97. ACM Press, New York, NY, 367-374.
- [3] Furnas, G. W. and Rauch, S. J. 1998. Considerations for information environments and the NaviQue workspace. In *Proceedings of the Third ACM Conference on Digital Libraries* (Pittsburgh, Pennsylvania, United States, June 23 - 26, 1998). I. Witten, R. Akscyn, and F. M. Shipman, Eds. DL '98. ACM Press, New York, NY, 79-88.
- [4] Marchionini, G. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (Apr. 2006), 41-46.
- [5] Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands, April 24 - 29, 1993).
- [6] White, R. W., Kules, B., Drucker, S. M., and schraefel, m. 2006. Introduction. *Commun. ACM* 49, 4 (Apr. 2006), 36-39.
- [7] Turetken, O. and Sharda, R. 2005. Clustering-Based Visual Interfaces for Presentation of Web Search Results: An Empirical Investigation. *Information Systems Frontiers* 7, 3 (Jul. 2005)