Bich-Liên Doan, Joemon Jose, Massimo Melucci and Lynda Tamine-Lechani (Eds.)

Proceedings of the

2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation

CIRSE 2010

Milton Keynes, UK

March 28, 2010

http://www.irit.fr/CIRSE/

in conjunction with ECIR 2010
http://kmi.open.ac.uk/events/ecir2010/

Contents

Preface

Since the 1990s, the interest in the notion of context in Information Access, Seeking and Retrieval increased. Many researchers have been concerning with the use of context in adaptive, interactive, personalized or collaborative systems, the design of explicit and implicit feedback techniques, the investigation of relevance, the application of a notion of context to problems like advertising or mobile search.

These proceedings include the contributions to the 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE), held in conjunction with ECIR-2010, Milton Keynes, UK, March 28th, 2010. The submitted papers were peer-reviewed by the members of the Programme Committee. Selection was based on originality, clarity, and technical quality. The abstracts of two keynotes are also included. The keynotes are by:

- **Stephen Robertson**
  Microsoft Research, Cambridge, and City University, London, UK
- **Ian Ruthven**
  University of Strathclyde, Glasgow, UK

The previous edition of this workshop held in Toulouse (CIRSE 2009) and other workshops and conferences, i.e. IR in Context (IRiX, 2005), Adaptive IR (AIR, 2006, 2008), Context-based IR (CIR, 2005, 2007) and Information Interaction in Context (IIiX, 2006, 2008) gathered researchers exploring theoretical frameworks and applications which have focussed on contextual IR systems. An important issue which gave raise to discussion has been Evaluation. It is commonly accepted that the traditional evaluation methodologies used in TREC, CLEF, NTCIR and INEX campaigns are not always suitable for considering the contextual dimensions in the information seeking/access process. Indeed, laboratory-based or system oriented evaluation is challenged by the presence of contextual dimensions such as user interaction, profile or environment which significantly impact on the relevance judgments or usefulness ratings made by the end user. Therefore, new research is needed to understand how to overcome the challenge of user-oriented evaluation and to design novel evaluation methodologies and criteria for contextual information retrieval evaluation.

The CIRSE workshop series aims to have a major impact on future research by bringing together IR researchers working on or interested in the evaluation of approaches to contextual information access, seeking and retrieval to foster discussion, exchange ideas on the related issues. The main purpose is to bring together IR researchers, to promote discussion on the future directions of evaluation.

The Workshop Organisers

Bich-Liên Doan, Supélec, France, Bich-Lien.Doan@supelec.fr

Joemon Jose, University of Glasgow, UK, jj@dcs.gla.ac.uk

Massimo Melucci, University of Padua, Italy, melo@dei.unipd.it

Lynda Tamine-Lechani, IRIT, France, Lynda.Lechani@irit.fr

Organization

Organizing Committee

- Bich-Liên Doan (Supélec, France)
- Joemon Jose (University of Glasgow, United Kingdom)
- Massimo Melucci (University of Padua, Italy)
- Lynda Tamine-Lechani (IRIT, France)

Program Committee

- Birger Larsen (Royal School of Library and Information Science, Denmark)
- Emanuele Di Buccio (University of Padua, Italy)
- Gilles Falquet (University of Geneva, Switzerland)
- Nicola Ferro (University of Padua, Italy)
- Martin Halvey (University of Glasgow, United Kingdom)
- Hideo Joho (University of Glasgow, United Kingdom)
- Gareth Jones (Dublin City University, Ireland)
- Peter Ingwersen (Royal School of Library and Information Science, Denmark)
- Diane Kelly (University of North Carolina, USA)
- Claude de Loupy (University of Paris X, France)
- Maarten de Rijke (University of Amsterdam, The Netherlands)
- Mathieu Roche (University of Montpellier, France)
- Ian Ruthven (University of Strathclyde, United Kingdom)
- Tassos Tombros (Queen Mary, University of London, United Kingdom)
- Robert Villa (University of Glasgow, United Kingdom)

Proceedings

- Emanuele Di Buccio (University of Padua)

Acknowledgements

The Organising Committee thank the Organisers of the 32nd European Conference on Information Retrieval (ECIR 2010) for their support.

On queries and other messages

Stephen Robertson
Microsoft Research Cambridge
7 J J Thomson Avenue
CB3 0FB Cambridge, UK
ser@microsoft.com

ABSTRACT

There are three parts to this talk – related in rather tangential ways. First, I will give a recap of an argument developed in a couple of earlier talks – at IIiX in 2008 and at the SIGIR evaluation workshop in 2009. The gist of the argument is about thinking about IR as a science, and the consequences for both theory and experimentation in the field. The second makes use of some ideas from general systems theory and the notion of open systems, and applies these ideas to information retrieval and the task context of search. The third discusses the status of queries in the current search world, and suggests two new ways to think about queries.

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

But where do we go from here?

Ian Ruthven
University of Strathclyde
Department of Computer and Information Sciences
ian.ruthven@cis.strath.ac.uk

ABSTRACT

Contextual information retrieval has promised much and has produced interesting ideas and new systems developments. Although there has been much discussion on the evaluation of context-free information retrieval systems, there has been far less discussion on the challenges of evaluating contextual systems.

In this presentation I will step back from the debates about specific methodologies to argue for approaches that acknowledge the context of use of IR systems and the context of information itself. I shall pose some challenges for contextual IR evaluation based on the context in which our systems are deployed in real-life situations and give a personal view on some useful evaluation directions.

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

An Experiment and Analysis System Framework for the Evaluation of Contextual Relationships

Ralf Bierig
Rutgers University
bierig@rci.rutgers.edu

Michael Cole
Rutgers University
m.cole@rutgers.edu

Jacek Gwizdka
Rutgers University
cirse10@gwizdka.com

Nicholas J. Belkin
Rutgers University
belkin@rutgers.edu

Jingjing Liu
Rutgers University
jingjing@eden.rutgers.edu

Chang Liu
Rutgers University
imliuc@gmail.com

Jun Zhang
Rutgers University
zhangj@eden.rutgers.edu

Xiangmin Zhang
Rutgers University
xiangminz@gmail.com

ABSTRACT

This paper presents an experiment and analysis system framework that allows researchers to design and conduct interactive experiments and analyze data for the evaluation of contextual relationships.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Context Evaluation, Information System, Data Analysis

1. INTRODUCTION AND BACKGROUND

In the last decade, context-aware computing has made much effort to formalize context[3], describe general context models[7] and develop systems that apply such models in different application domains [5] – such as mobile computing (e.g. tourism and recreation [16, 11]). There is, however, only limited research about the experimental evaluation of context, particularly about the effects of various contextual attributes and their interaction. This gap is beginning to be addressed with several workshops and conferences [9, 8, 12, 2].

Rigorous experimentation in this domain presents challenges in that such experiments are generally difficult to administer and demanding in resources [6, 10]. Although software frameworks for contextual enrichment of applications exist [13, 4] there is generally little system-related support for comprehensive evaluation of context attributes and models. This paper presents a system framework that provides researchers with a tool to: 1) *design and conduct experiments* for the evaluation of particular contextual attributes and 2) *integrate data and analyze results* to better understand contextual relationships. The framework promotes an interactive and task-oriented viewpoint that is supported by a wide range of logging tools.

Section 2 reviews the system architecture consisting of the experiment and the analysis system. Section 3 describes how the architecture supports researchers to investigate and evaluate contextual relationships. Section 4 discusses the current state of the system and future plans for its dissemination.

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

2. SYSTEM FRAMEWORK

The system framework is part of a project deliverable[1] that aims to investigate ways to improve users' ability to find information in search environments such as digital libraries. In particular we analyze various interacting contextual factors that are involved in such online search activities. Despite our focus, results are expected to contribute to a much wider range of application environments such as mobile search and recommender systems.

2.1 Overview

The overall aim of our framework is to reduce the complexity of designing and conducting experiments and integrating and analysing results from experiments for the evaluation of contextual relationships as usually expressed in user and context models. Such experiments usually require a complex arrangement of system components (e.g. GUI, user management and persistent data storage). Our framework enables researchers to focus on research related issues (e.g. task and questionnaire design and the selection of experiment variables) rather than the creation of the experiment logic and the transformation, integration and the processing of data and results after the experiment has been completed. This helps to reduce the overall time and effort that is needed to design and conduct experiments and to get valuable results about contextual relationships from experiment data. As shown in figure 1, the system framework consists of two parts

[1] http://comminfo.rutgers.edu/imls/poodle/

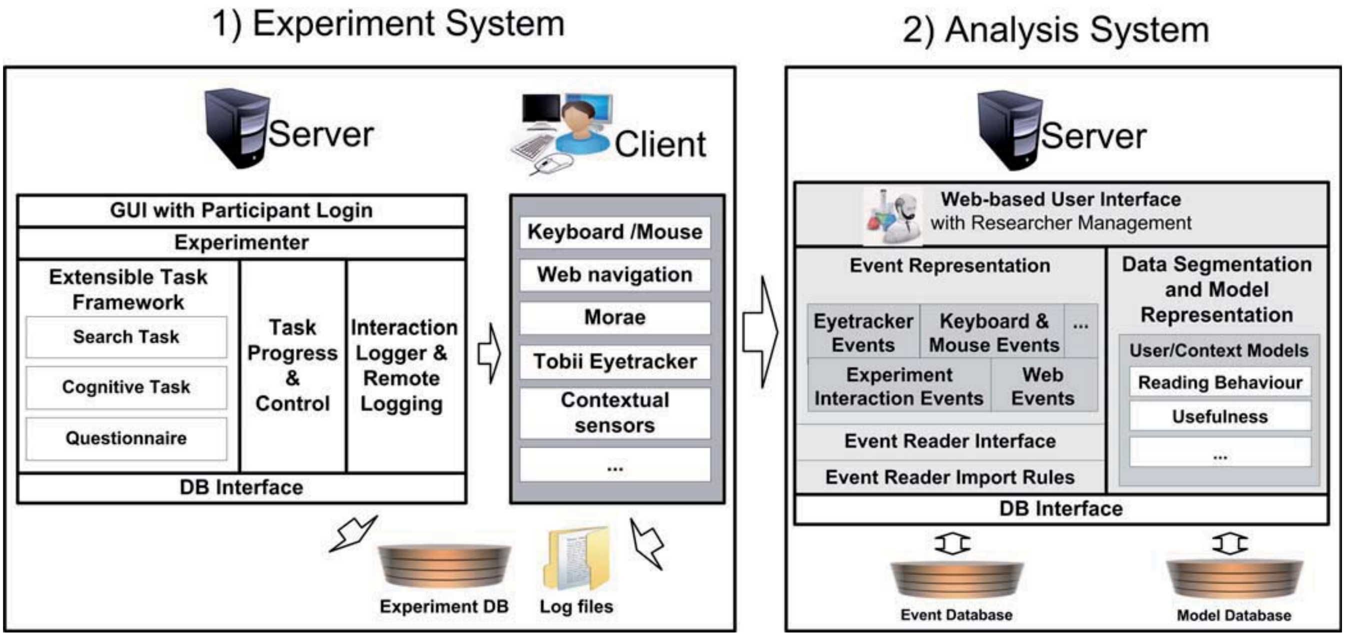

Figure 1: Components of the experiment and analysis system framework

– 1) an *experiment system* that allows researchers to design and conduct interactive experiments in close-to-operational application environments and 2) an *analysis system* that enables them to integrate and analyze results obtained from such experiments.

2.1.1 Experiment System

The experiment system, described in more detail in [1], includes a number of components.

The *GUI* provides authenticated login for participants, their assignment to one or more experiments and basic navigational support during an experiment. The *Experimenter* controls and coordinates an *Extensible Task Framework* that offers researchers a set of reusable tasks that can be used for creating experiments (e.g. standard open web search tasks). Own tasks can also be added to this collection. Furthermore, the Experimenter manages *Task Progress and Control* that balances task sequences, monitors the progress of participants including the safe recovery of interrupted sessions. In addition, the *Interaction Logger with Remote Logging* provides a mechanism for tasks to log contextual information internally at specific points during an experiment task and to call external logging applications on the client. This allows creating more effective experiments that may include different kinds of contextual data logging on both server and the client side. Whereas the server has a central logging facility, the client consists of a flexible and expandable array of independent loggers. Currently, these loggers observe the most commonly known user behaviours – keyboard and mouse activities, web navigation, usability information from Morae[2] and eye-tracking data from Tobii[3]. This list can easily be expanded with other (existing or new) logging tools that cover additional contextual information from the user or the user's environment. Examples may include location information (e.g. geographic position or proximity to points of interest) or physiological states of the user (e.g. heart rate or Galvanic skin response). Logging information is either stored in an experiment database through the *DB Interface* or in application-specific log files.

2.1.2 Analysis System

The analysis system serves as an extension of the experiment system with additional features to integrate experiment data into a unified data structure. Researchers can inspect and explore these data sets and segment and model results to gain a better understanding of contextual relationships. The analysis system consists of the following components:

- The *Event Representation* integrates experiment data through the *Event Reader Interface* into a unified event data structure. This data structure is extensible and the collection of event readers mirror the logging tools provided with the experiment system as described in the previous section. An extensible set of event types ensures that researchers can adapt and extend the analysis framework to process data from a variety of experiments under a single platform. This ensures that additional logging tools can be introduced through the experiment system to capture additional types of user context either through the logging of high-level user behaviour or through the application of low-level sensors as described in [14].
- *Event Reader Import Rules* can be used to configure event readers and therefore adapt the data import process. Such rules can for example be applied to add additional filters for event readers (e.g. excluding web

[2] http://www.techsmith.com
[3] http://www.tobii.com

events with certain URLs) or providing standard validation (e.g. tagging certain events as problematic thus flagging results for manual inspection).

- *Data Segmentation* divides experiment data into semantic units guided by research hypotheses. The system framework provides a standard minimal segmentation by distinguishing data based on experiments, users and tasks. The research can add additional levels of data segmentation to structure data in smaller logical units. A segmentation can for example differentiate interaction data based on users' current stage in a search task (e.g. distinguishing users' task stages of query formulation, result page inspection and content page viewing) or, more generally, data can be segmented along low-level decision points (e.g. mouse clicks and/or key strokes).

- *Model Representation* processes (segmented) event sequences to test specific research hypotheses i.e. verifying effects of context attributes and relationships between them (e.g. identifying users' perceived usefulness of content and determining reading behaviour). Other data segmentations and model representations can be added by researchers to further specialize the system framework for particular types of analysis.

- The *Web-based User Interface* extends the system to an online service where researchers can generate, inspect and share event representations, data segmentations and models within one or across multiple experiment data sets. These are stored through a *DB Interface* that persists both event and model representations into separate databases for later reuse. The user interface supports authenticated login to allow the system to be used as part of a collaborative research platform.

3. CONTEXT EVALUATION WITH THE SYSTEM FRAMEWORK - BENEFITS AND LIMITATIONS

The system design incorporates many aspects useful for the evaluation of contextual relationships from data obtained in interactive and task-based experiments. This section summarizes these aspects, shows how they relate to the system framework, points out how they can help researchers to evaluate context, and expresses limitations that should be considered.

- *Modularity:* Context models may cover a wide range of attributes based on dimensions such as the application environment (e.g. library or mobile environment) and the intended user group (e.g. professional journalists or online web searchers) as well as others. The system framework supports this requirement in a number of ways. First, a modular and multi-dimensional logging framework within the experiment system can record behavioural data from the user and sensory data from the user's environment. Second, these multi-dimensional data streams can be integrated into a unified stream of events within the analysis system. Third, this event stream can be treated holistically through segmentation, as a tool for data categorization and conditioning, and through modelling to investigate and discover contextual relationships.

- *Extensibility:* As an extensible framework with respect to contextual logging tools (in the experiment system) and readers, rules, segmentations and models (in the analysis system) the framework offers researchers ways to adapt and extend it to their own requirements and research agendas. These extensions however require additional, customizing implementation work by the user of the system framework; for example adding another logging tool to measure a new contextual aspect from the user also requires implementing the corresponding event representation and an additional reader to import the new data log. Such procedures, however, are guided through the application of programming interfaces and supported with examples that are available in open source as part of the project. This is not much different from other extensible software frameworks such as WEKA [15].

- *Separation between data and modelling:* Data (in the form of low-level event representations) is separated from its interpretation (in the form of high-level segmentations and models). Thus, it is possible to generate multiple, alternative context models from the same underlying events that can each be evaluated in isolation. This also allows user and context models to be reused for different data segments from one or across multiple experiments.

- *Collaboration* is central to the design and has been supported in both parts of the system framework. The experiment system allows researchers to implement and share experiment tasks thus building a collaborative repository (e.g. internet search tasks, tag cloud search, standard questionnaires for language understanding and various cognitive tests). Likewise, configurations for behavioural and contextual logging tools can be created and reused across different experiments and shared between researchers. The analysis system offers a meeting platform through its web-based user interface. Data, segmentations and models can be configured, integrated and shared between researchers allowing collaborating with data and ideas and forming virtual research groups. Researchers can create and exchange integrated event data sets from experiments specific to the needs of individuals or groups (e.g. event data limited to a subset of experiment participants, experiment tasks or types of context such as web activity or eye movement). Shared data sets can then be applied for further data segmentation (e.g. selecting only particular user activities or contextual states, such as query input or reading behaviour). An extensible pool of models can be applied to such segments and accessed collaboratively. Basic summary visualizations are available and findings can be exported allowing researchers to further process data with third-party tools and apply results (e.g. integrating a learned context model in a personalized desktop search application).

4. CURRENT STATE AND FUTURE PLANS

A prototype of the experiment system has has been designed and developed with active work on improving logging comprehensiveness (especially for contextual, sensor-based logging) and scalability. The experiment system has already been applied to design and conduct four experiments each with distinctive design and goals for our research project. In those experiments we have collected rich contextual information for the basic investigation of relationships between use behaviour and various user context attributes such as cognitive abilities and individual differences, reading and scanning behaviour and perception of usefulness during online search. The analysis system has been designed and the modelling and user interface is in active development. The experiment system framework has been released as open source[4]. The analysis system will be released as open source when it is feature complete and stable. Both of these systems can benefit the research community by allowing for collaboration between researchers and enabling additional improvements and extensions to better serve the needs of context researchers.

Acknowledgements: This work is supported by IMLS grant LM-06-07-0105-07.

5. REFERENCES

[1] R. Bierig, J. Gwizdka, and M. Cole. A user-centered experiment and logging framework for interactive information retrieval. In N. J. Belkin, R. Bierig, G. Buscher, L. v. Elst, J. Gwizdka, J. Jose, and J. Teevan, editors, *SIGIR 2009 Workshop on Understanding the user - Logging and interpreting user interactions in IR*, Boston, MA, 2009.

[2] P. Borlund, J. W. Schneider, M. Lalmas, A. Tombros, J. Feather, D. Kelly, A. de Vries, and L. Azzopardi. Second symposium on information interaction in context (iiix). London, UK, 2008. ACM Press.

[3] A. K. Dey, G. Kortuem, D. Morse, and A. Schmidt. Special issue on situatuated interaction and context-aware computing. *Personal and Ubiquitous Computing*, 5(1), 2001.

[4] P. Fahy and S. Clarke. Cass - middleware for mobile context-aware applications. In *2nd International Conference on Mobile Systems, Applications, and Services (MobiSys 2004), Workshop on Context-Awareness*, Bosten, MA, USA, 2004.

[5] A. Göker, H. I. Myrhaug, and R. Bierig. Context and information retrieval. In A. Göker and J. Davies, editors, *Information Retrieval: Searching in the 21st Century*. John Wiley and Sons, Ltd, Chichester, UK, 2009.

[6] J. Goodman, S. Brewster, and P. Gray. Using field experiments to evaluate mobile guides. In *6th International Symposium on Human Computer Interaction with Mobile Devices and Services (Mobile HCI), International Workshop on HCI with Mobile Guides*, Glasgow, UK, 2004.

[7] J. Indulska and D. D. Roure. Workshop on advanced context modelling, reasoning and management. In *6th International Conference on Ubiquitous Computing (UbiComp)*, Nottingham, UK, 2004.

[8] P. Ingwersen, K. Jaervelin, and N. Belkin. Workshop on information retrieval in context. In *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005. Royal School of Library and Information Science, Copenhagen, Denmark.

[9] P. Ingwersen, K. van Rijsbergen, and N. Belkin. Workshop on information retrieval in context (irix). In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, 2004.

[10] J. Kjeldskov, C. Graham, S. Pedell, F. Vetere, S. Howard, S. Balbo, and J. Davies. Evaluating the useability of a mobile guide: The influence of location, participants and resources. *Behaviour and Information Technology*, 24(1):51–65, 2005.

[11] D. M. Mountain and A. MacFarlane. Geographic information retrieval in a mobile environment: Evaluating the needs of mobile individuals. *Journal of Information Science*, 33(5):515–530, 2007.

[12] I. Ruthven, P. Borlund, P. Ingwersen, N. Belkin, A. Tombros, and P. Vakkari, editors. *Information Interaction in Context. 1st International Symposium on Information Interaction in Context, IIiX 2006.* ACM Press, Copenhagen, Denmark, 2006.

[13] D. Salber, A. K. Dey, and G. D. Adowd. The context toolkit: Aiding the development of context-enabled applications. In *Conference on Human Factors in Computing Systems (CHI)*, pages 434–441, Pittsburgh, PA, USA, 1999. ACM Press.

[14] A. Schmidt. *Ubiquitous Computing - Computing in Context.* Phd thesis, Lancaster University, 2002.

[15] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2nd edition, 2005.

[16] A. Zipf. User-adaptive maps for location-based services (lbs) for tourism. In *9th Int. Conf. for Information and Communication Technologies in Tourism (ENTER 2002)*, pages 329–337, Innsbruck, Austria, 2002. Springer Verlag.

[4] `http://sourceforge.net/projects/piirexs/`

Contextual evaluation of mobile search

Ourdia Bouidghaghen
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
bouidgha@irit.fr

Lynda Tamine
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
lechani@irit.fr

Mariam Daoud
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
daoud@irit.fr

Cécile Laffaire
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
laffaire@irit.fr

ABSTRACT

We discuss the issue of evaluating our context-based personalized mobile search approach with a methodology based on a combination of two evaluation approaches: context simulation and user study. Our personalized approach aims at exploiting some context-aware user profiles through a personalized score to re-rank initial search results obtained from a standard search system. We use Yahoo!'s open search web services platform BOSS [1] as a baseline. The context simulation allows us to simulate user locations and their related user interests. The user study involves real users who give their relevance judgments to the top 20 documents returned by yahoo and by our approach through an assessment tool available on the web platform OSIRIM[2]. The experimental results show the effectiveness of our personalized approach according to the proposed evaluation protocol.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback

Keywords

mobile search, context, user profile, evaluation protocol

1. INTRODUCTION

The proliferation of mobile technologies such as (PDAs and mobile phones, ...) and, with them, of mobile users, have moved the static world of classical and Web IR towards an always changing context-based world. The notion of context, roughly described as the situation the user is in, is exploited in the development of new IR systems. Starting from considering only a low number of contextual features (location, time and interests), such systems are faced to a new challenge for IR, that is how those contextual data can enhance user satisfaction. Another important issue is how to evaluate the strategies and techniques involved in these new systems. It is commonly accepted that the traditional evaluation methodologies used in TREC, CLEF and INEX campaigns are not always suitable for considering the contextual dimensions in the information access process. Indeed, laboratory-based or system oriented evaluation is challenged by the presence of contextual dimensions such as user profile or environment which significantly impact on the relevance judgments or usefulness ratings made by the end user [17]. To alleviate such limitations, contextual evaluation methodologies have been proposed to support simulated user profile through contextual simulations [16] or real evaluation scenarios through user studies [5].

As an initial approach, yet allowing meaningful observations, we present here, the evaluation protocol aiming to evaluate empirically the performance of a novel context-based personalized mobile search system. For this purpose, we compare the performance of retrieval: without personalization and with personalization. We compare our approach to the results obtained from yahoo BOSS web search service, which did not implement itself any personalization capability. This paper discusses the methodology adopted and presents the results obtained. We first briefly survey IR evaluation methodologies in mobile contexts (Sec. 2). We then presents our approach for mobile search personalization, and introduce our contextual IR evaluation protocol (Sect. 3). Finally, we conclude and give perspectives for future works.

2. EVALUATION OF IR IN MOBILE CONTEXTS

Context-awareness in mobile IR focuses on context models including user profiles and environmental data (time, location, near persons, device and networks). The state-of-the-art highlights that significative theoretical and technological progress has been achieved in this area over the last few years, encouraged by the growing interest to co-located human-human communications and large scale location-based applications ([10, 15]). In the development of an IR system for mobile environments, evaluation plays an important role, as it allows to measure the effectiveness of the system and to better understand problems from both the system and the

[1] http://developer.yahoo.com/search/boss/
[2] https://osirim.irit.fr developed at IRIT lab

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

user interaction point of view. However, evaluation remains challenging because of the main following reasons ([4, 11]): 1) environmental data should be available and several usage scenarios should be evaluated across them, 2) evaluation, if present, concerns a specific application (eg.tourist guide), generalization to a wide range of information access applications is difficult. Both user-centered and benchmark evaluation approaches are adopted. However, as mobile IR systems are strictly related to users and their environment, the user-centered evaluation live (user studies [3, 14, 8]) or in laboratory (context-simulation framework [4, 9]) seem to be the most natural one. In [8] for example, a user-centered, iterative, and progressive evaluation has been adopted combining IR evaluation methods with human-computer interaction development techniques. The authors consider mainly the following guidelines: involve the right participants that are either current users or likely future; choose the right situations considering the different aspects of the environment; set relevant tasks that make participants seek information and are in accordance with situations that have been identified; use relevant evaluation approach and measures according to the different sub-goals (effectiveness, usability) within the overall objective evaluation. The main limitations introduced by user studies is that experiments are not repeatable and that they induce an extra costs. Within the mobile IR field, a benchmark evaluation has been used in [13, 12], they demonstrated the efficacy of the benchmark approach to evaluate an early stage of their system.

3. EVALUATION OF OUR CONTEXT-BASED PERSONALIZED SEARCH

In this section, we first introduce our context-based personalized approach for mobile search, we then present our evaluation protocol devoted for our proposed approach.

3.1 Situation-aware user profile

Our context-aware approach to personalize search results for mobile users [2] aims to adapt search results according to user's interests in a certain situation. A user U is represented by a set of situations with their corresponding user profiles (interests), denoted : $U = \{(S^i, G^i)\}$, where S^i is a situation and G^i its corresponding user profile. A situation S^i refers to the geographical and/or temporal context of the user when submitting a query to the search engine. User profiles are built over each identified situation by combining graph-based query profiles. A query profile G_q^s is built by exploiting clicked documents D_r^s by the user and returned with respect to the query q^s submitted at time s. First a keyword query context K^s is calculated as the centroid of documents in D_r^s:

$$K^s\left(t\right) = \frac{1}{|D_r^s|}\sum_{d \in D_r^s} w_{td} \ . \qquad (1)$$

K^s is matched with each concept c_j of the ODP[3] ontology represented by single term vector $\vec{c_j}$ using the cosine similarity measure. The scores of the obtained concepts are propagated over the semantic links as explained in [6]. We select the most weighted graph of concepts to represent the query profile G_q^s at time s. The user profile G_i^0, within each identified situation S^i, is initialized by the profile of the first query submitted by the user at the situation S^i. It is updated by combining it with the query profile G_q^{s+1} of a new query for the same situation, submitted at time $s+1$. A case-based reasoning approach [1] is adopted for selecting a profile G^{opt} to use for personalization according to a new situation by exploiting a similarity measure between situations as explained in [2]. Personalization is achieved by re-ranking the search results of queries related to the same search situation. The search results are re-ranked by combining for each retrieved document d_k, the original score returned by the system $score_o(q^*, d_k)$ and a personalized score $score_c(d_k, G^{opt})$ obtaining a final $score_f\left(d_k\right)$ as follows:

$$score_f\left(d_k\right) = \gamma * score_o\left(q^*, d_k\right) + \left(1-\gamma\right) * score_c\left(d_k, G^{opt}\right) \qquad (2)$$

Where γ ranges from 0 to 1. Both personalized and original scores could be bounded by varying the values of γ. The personalized score $score_c(d_k, G^{opt})$ is computed using the cosine similarity measure between the result d_k and the top ranked concepts of the user profile C^{opt} as follows:

$$score_c\left(d_k, G^{opt}\right) = \sum_{c_j \in C^{opt}} sw\left(c_j\right) * \cos\left(\vec{d_k}, \vec{c_j}\right) \qquad (3)$$

Where $sw\left(c_j\right)$ is the similarity weight of the concept c_j in the user profile G^{opt}.

3.2 Evaluation of contextual personalization

In the absence of a standard evaluation framework, a formal evaluation of contextualization techniques may require a significant amount of extra feedback from users in order to measure how much better a retrieval system can perform with the proposed techniques than without them. In this case, the standard evaluation measures from the IR field require the availability of manual content ratings with respect to query relevance and specific user preference (i.e., constrained to the context of his search). For this aim we build a testbed consisting of a search space corpus, a set of queries, and a set of hypothetic context situations. A user study was conducted, participants were asked to provide ratings, in a blind test, for two retrieval scenarios: 1) top 20 documents returned by Yahoo BOSS, 2) top 20 documents returned by our personalized approach. In the following, we describe our experimental data sets and our evaluation protocol.

3.2.1 Contexts and Queries

Since the contextualization techniques are applied as the time goes, we have defined a set of *six* short use cases as part of the evaluation setup. Each use case is composed of a set of queries within a given geographical context, and a narrative describing the relevance of a document regarding a query and a geographical context. We have simulated a set of six geographical contexts defined by a location type (*zoo, music store, cinema, library, garden* and *museum*). We have created a set of totally *25* different queries, *5* queries belonging to each geographical context. Since mobile search queries are known to be short (and thus ambiguous), our queries are generally short (query length $\leq$ 3) and some of them are consequently ambiguous (eg. *jaguar*) and are tested within different geographical contexts (eg. the query *"water lilies"* is tested within the two contexts *"garden"* and *"museum"*), totalizing a number of *30* queries within the six contexts. Our goal was to verify whether the consideration

[3] The Open Directory Project (ODP): http://www.dmoz.org

of geographical contexts and user profiles can enhance the performance of the search engine to respond to such ambiguous queries. Table 1 gives an example of the use case of the context *museum*.

3.2.2 Document collection

The document collection consists of a set of about 3750 web pages retrieved from the web by yahoo BOSS as response to our set of queries. It is built by collecting the 150 first retrieved documents per query.

3.2.3 User profile

The user profiles are integrated in the evaluation strategy according to a simulation algorithm that generates them using hypothetic user interactions for each query. They are constructed based on a manual judgments of the <query, narrative, document> tuples for all the document in the collection. These, so built profiles, simulate user click-through data.

3.2.4 Evaluation protocol

Our experimental design consists of evaluating the effectiveness of our personalized approach when using the user profile in the IR model over a sequence of user contexts. In the absence of an initial score of the document results list of yahoo BOSS, the re-ranking procedure is done based only in the personalized score (ie. $\gamma = 0$ in equation 2). The evaluation scenario is based on the k-fold cross validation like in [7] explained as follows:

- for each use case, divide the query set into k equally-sized subsets, and using $k-1$ training subsets for learning the user interests and the remaining subset as a test set,
- for each query in the training set, an automatic process generates the associated profile based on its top n relevant documents listed in the manually constructed relevance judgments file.
- update the user profile concept weights across the queries in the training set and use it for re-ranking the search results of the queries in the test set.

In order to evaluate the performance of our proposed approach, a user study is conducted to compare the 20 top ranking output of our approach and of Yahoo BOSS. Using an assessment tool available on the web platform OSIRIM, *six* users who participated to the experiment were asked to judge each tuple <query, document, narrative> within the 20 top ranking output of both our approach and of Yahoo BOSS. Participants were unaware of the system they judge. Relevance judgments have been made using a three level relevance scale: relevant, partially relevant, or not relevant.

3.3 Results and Discussion

We evaluate the effectiveness of the personalized search over the six use cases and we compare the obtained results to the initial ones from Yahoo BOSS. To better estimate the quality of the search results at the top of the ranked list (since mobile users are unlikely to scroll long lists of retrieved items), we estimate the DCG@10 for all the queries.

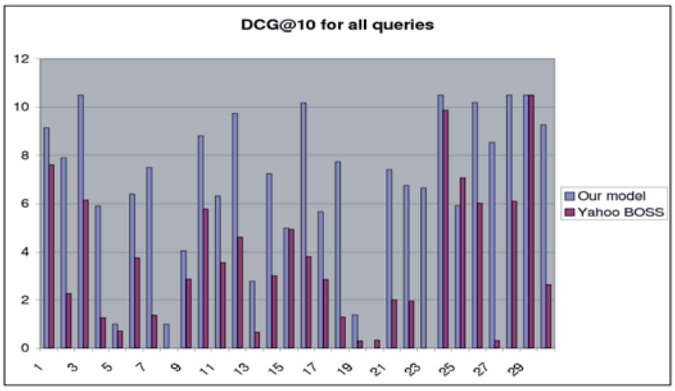

Figure 1: DCG@10 comparison between our personalized search and Yahoo BOSS over all queries

Table 2: Average Top-n precision comparison between our personalized search and Yahoo BOSS over all queries

	Average precision over all queries at:			
	P@5	P@10	P@15	P@20
Yahoo BOSS	0,37	0,39	0,38	0,36
Our model	0,70	0,64	0,59	0,55
Improvement	**87,50%**	**63,56%**	**53,49%**	**50,92%**

Figure 1 compares the effectiveness obtained by the initial yahoo search lists and the re-ranked ones obtained by our approach over all the queries. We observe that in general, our approach enhances the initial DCG@10 obtained by the standard search and improve the quality of the top search results lists. We have also computed the percentage of improvement of personalized search comparatively to the standard search computed at different cut-off points P@5, P@10, P@15 and P@20 averaged over all the queries. Results are presented in Table 2. Results prove that personalized search achieves higher retrieval precision of almost the queries in the six simulated contexts. Best performance are achieved by the personalized search in terms of average precision at different cut-off points achieving an improvement of 87,50% at P@5, 63,56% at P@10, 53,49% at P@15 and 50,92% at P@20 comparatively to Yahoo BOSS. However, precision improvement varies between queries, Figure 2 gives an example of this improvement variation between the queries of the context *museum*. This is probably due to the difference between the degree of ambiguity of the queries, which can not be explained only by the difference in query length. In fact, it depends also on the contents of the documents present in the collection.

4. CONCLUSION

In this paper we have presented our evaluation protocol of a context-aware personalization approach for mobile search. It is based on a combination of context simulation and user study. More precisely, we exploit context simulation to create user contexts and profiles in one hand. On the other hand, we exploit Yahoo's BOSS web search service and real user judgments, through a user study, to evaluate the search effectiveness of our approach comparatively to a standard search. We evaluated our approach according to the pro-

Table 1: an example of the use case *"museum"*

Context	QueryID	Query terms	Narrative
museum	M17	da Vinci	A document is relevant if it speaks about *da Vinci* painter and or his paintings
	M23	sunflowers	A document is relevant if it speaks about the painting *sunflowers* and or its painter *Van Gogh* and or his paintings
	M24	woman with a parasol	A document is relevant if it speaks about the painting *woman with a parasol* and or its painter *Claude Monet* and or his paintings
	M25	Edgar Degas	A document is relevant if it speaks about painter *Edgar Degas* and or his paintings
	M21	water lilies	A document is relevant if it speaks about the painting *water lilies* and or its painter *Claude Monet* and or his paintings

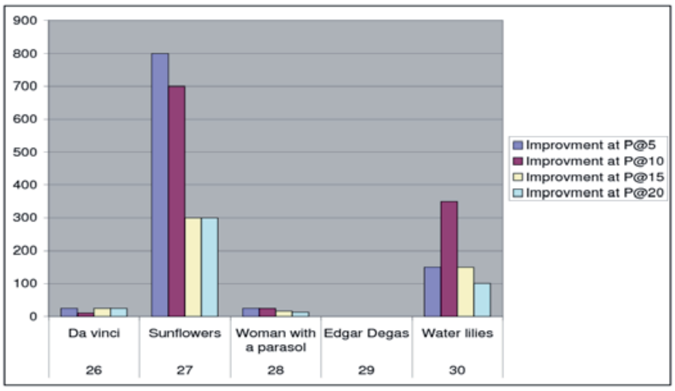

Figure 2: Improvement at P@5, P@10, P@15 and P@20 for the queries of the context "museum"

posed evaluation protocol and show that it is effective. In future work, we plan to extend this protocol by using real user data provided from a search engine log file. Extending the protocol aims at testing the effectiveness of the personalized search based on real mobile search contexts and click-through data available in the log file.

5. ACKNOWLEDGMENTS

The authors acknowledge the support of the project QUAERO, directed by OSEO agency, France, and thank PhD students at IRIT for their participation in the experiment.

6. REFERENCES

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1), 1994.

[2] O. Bouidghaghen, L. Tamine-Lechani, and M. Boughanem. Dynamically personalizing search results for mobile users. In *Proc. of Flexible Query Answering Systems*, pages 293–298, 2009.

[3] N. O. Bouvin, B. G. Christensen, K. Grønbæk, and F. A. Hansen. Hycon: a framework for context-aware mobile hypermedia. *Hypermedia*, 9(1):59–88, 2003.

[4] M. Bylund and F. Espinoza. Testing and demonstrating context-aware services with quake iii arena. *Communications of the ACM*, 45(1), 2002.

[5] V. Challam, S. Gauch, and A. Chandramouli. Contextual search using ontology-based user profiles. In *Proceedings of RIAO 2007*, 2007.

[6] M. Daoud, L. Tamine, M. Boughanem, and B. Chebaro. A session based personalized search using an ontological user profile. In *ACM Symposium on Applied Computing (SAC)*, pages 1031–1035, 2009.

[7] M. Daoud, L. Tamine-Lechani, and M. Boughanem. Using a concept-based user context for search personalization. In *Proc. of the 2008 Internat. Conf. of Data Mining and Knowledge Engineering*, 2008.

[8] A. Göker and H. I. Myrhaug. Evaluation of a mobile information system in context. *Information Processing and Management*, 44(1):39–65, 2008.

[9] F. Gui, M. Adjouadi, and N. Rishe. A contextualized and personalized approach for mobile search. In *2009 Internat. Conf. on Advanced Information Networking and Applications Workshops*, pages 966–971.

[10] R. Iqbal, J. Sturm, O. Kulyk, J. Wang, and J. Terken. User-centred design and evaluation of ubiquitous services. In *Proc. of the 23rd annual internat. conf. on Design of communication*, pages 138–145, 2005.

[11] J. Kjeldskov and C. Graham. A review of mobile hci research method. In *Human-Computer Interaction with Mobile Devices and Services-5th Internat. Symposium, Mobile HCI 2003 proceedings*, 2003.

[12] D. Menegon, S. Mizzaro, E. Nazzi, and L. Vassena. Benchmark evaluation of context-aware web search. In *Proc. of ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation.*

[13] S. Mizzaro, E. Nazzi, and L. Vassena. Retrieval of context-aware applications on mobile devices: how to evaluate? In *Proc. of IIiX'08*, pages 65–71, 2008.

[14] C. Panayiotou, M. Andreou, G. Samaras, and A. Pitsillides. Time based personalization for the moving user. In *Proc. of the International Conference on Mobile Business (ICMB'05)*, pages 128–136, 2005.

[15] W. Schwinger, C. Grün, B. Pröll, W. Retschitzegger, and A. Schauerhuber. *Context-awarness in mobile tourism guides- a comprehensive survey.* Technical Report,Johannes Kepler University Linz, IFS/TK, 2005.

[16] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proc. of the 16th ACM conference on information and knowledge management*, pages 525–534, 2007.

[17] L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowledge and Information Systems, Springer*, 2009.

A Proposal for the Evaluation of Adaptive Personalized Information Retrieval

Séamus Lawless, Alexander O'Connor, Catherine Mulwa

Centre for Next Generation Localization
School of Computer Science and Statistics
Trinity College Dublin,
Dublin, Ireland

seamus.lawless@sccs.tcd.ie, alex.oconnor@sccs.tcd.ie, mulwac@sccs.tcd.ie

ABSTRACT

Personalisation in Information Retrieval is achieved using a range of contextual information such as information about the user, the task being conducted and the device being used. This information is used to devise the most suitable response for the individual's need. As such personalised Information Retrieval and response composition approaches become more widely used, traditional evaluation measures become less effective and applicable. This paper proposes that contextual, and specifically personalised, approaches to Information Retrieval could benefit from the experience of the Adaptive Hypermedia community. The paper details the approaches to evaluation commonly used by the IR and AH communities and proposes a means of combining and enhancing these disparate approaches in a unified framework for the evaluation of personalised IR systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia;

General Terms

Experimentation, Measurement, Performance

Keywords

Information Retrieval, Personalisation, Adaptive Hypermedia, Evaluation

1. INTRODUCTION

Contextual information is increasingly being used to facilitate personalisation in Information Retrieval (IR). The personalised identification, retrieval and presentation of resources can provide the user with a tailored information seeking experience. Such tailored experiences can produce a more informative response than a traditional ranked list approach.

As personalised information retrieval and response composition become more widely used, traditional approaches to IR evaluation become less effective and applicable. It is the contention of this paper that contextual, and specifically personalised, approaches to IR could benefit from the experience of the Adaptive Hypermedia (AH) community.

Recently, research has been undertaken exploring how to enhance and combine key aspects of AH research with IR research to provide advanced annotation, slicing, retrieval and composition of multilingual digital content drawn from corporate documents repositories as well as open corpus sources [1][2]. We call such systems, which combine AH and IR approaches to deliver personalised information seeking and access, Adaptive Information Retrieval Systems (AIRS).

AH systems have traditionally functioned using low volumes of content with associated rich metadata descriptions, in contrast to web-based IR systems which tend to function using statistical methods on very large scale collections of unannotated content. In order to facilitate effective personalised information retrieval at the scale required by web-based IR systems, a hybrid approach to both the implementation and evaluation of these systems is necessary. This paper proposes an approach to the evaluation of AIRS. Having introduced the disciplines of IR and AH, the paper details the approaches to evaluation commonly used by these communities. The paper then continues by proposing a means of combining and enhancing these disparate approaches to produce a framework for the evaluation of adaptive, personalized information retrieval systems.

2. EVALUATION APPROACHES

2.1 Information Retrieval

The Cranfield model [3] remains the dominant approach to the evaluation of IR systems. This evaluation model uses test collections to assess and compare the performance of IR systems. Typical test collections include a document corpus, a set list of search queries and manually assigned relevance assessments for each query. Such collections are utilised in IR communities such as the Text REtrieval Conference (TREC)[1], the Cross Language

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
`http://www.irit.fr/CIRSE/`

[1] The Text Retrieval Conference – http://trec.nist.gov

Evaluation Forum (CLEF)[2] and the National Institute of Informatics Test Collection for IR Systems (NTCIR)[3].

Precision and recall [4] are the metrics upon which the majority of traditional IR evaluation approaches are based, including the Cranfield model. Precision is the fraction of documents retrieved in response to a query that are relevant to the users information need when making that query. Recall is the fraction of the total set of relevant documents in a collection which are returned for a given query. Another common measure of performance is van Rijsbergen's f_1-measure [5], which calculates the weighted harmonic mean of precision and recall.

However, differences between the structure of the WWW and typical document collections and key differences between the users of web-based IR systems and traditional IR systems challenge many of the assumptions of these metrics. The primary concern is that both metrics are highly subjective, as relevancy can only be assigned based upon the users intent when submitting a specific query. In relation to the WWW, recall is an ineffective means of evaluation, as recall requires the entire collection of relevant documents to be known in advance of each query performed [6].

Whilst the precision measure for a set of results is valuable, it does not take into account the importance of ranking in modern, web-based, information retrieval. The average precision is a measure used to estimate the performance of an IR system by placing emphasis on relevant results appearing high up the ranked list. Average precision is calculated by finding the average of the individual precision calculations for each relevant result in the list, if the list were truncated directly after that result. The Mean Average Precision (MAP) for an IR system is then calculated by summing the average precision value for each query conducted and dividing by the total number of queries. This gives an indication of the overall performance of the system with relation to precision and ranking.

The Cranfield model of test collections aims to ensure controlled experimentation across systems, using metrics such as those introduced above, so that conclusions can be formed about the comparative performance of a variety of IR systems. However, this approach is limited as assumptions must be made about user intent and behaviour. Due to the mainstream accessibility of web-based IR systems, assumptions cannot be made regarding a user's level of knowledge with regard to either the system or the subject domain. The IR system must cope with users across the spectrum of knowledge and ability [7].

2.2 Adaptive Hypermedia

Adaptive Hypermedia (AH) is an alternative to the traditional "one-size-fits-all" approach to the development of Hypermedia Systems. Adaptive Hypermedia Systems (AHS) can potentially adapt on any attributes of a user, for example in Adaptive Educational Hypermedia Systems (AEHS) models can be constructed which describe the goals, preferences, knowledge or other attributes of each individual user. This model can then be used during any interactions with the user, in order to adapt aspects of the systems functionality or presentation strategy to the needs of that user [8]. AH approaches are applied in educational hypermedia, on-line information systems, on-line help systems, institutional hypermedia and systems for managing personalized views.

2.2.1 Current Evaluation of Adaptive Systems

In order to produce effective results, evaluation should occur throughout the entire design cycle and provide feedback for design modification [9].

User-centered approach: User-centered evaluation (UCE) can serve three goals: verifying the quality of an AHS, detecting problems in the system functionality or interface, and supporting adaptivity decisions [10]. These functions make UCE a valuable tool for developers of all kinds of systems, because they can justify their efforts, improve upon a system or help developers to decide which version of a system to release. The benefits of the user-centered approach are savings in terms of time and cost, ensuring the completeness of system functionality, minimizing required repair efforts, and improving user satisfaction [11].

Empirical approach: Empirical evaluations, also known as controlled experiments, refer to the appraisal of a theory by observation in experiments. These evaluations help to estimate the effectiveness, efficiency and usability of a system and may uncover certain types of errors in the system that would remain otherwise undiscovered. The key to good empirical evaluation is the proper design and execution of the experiments so that the particular factors to be tested can be easily separated from other confounding factors. This method of evaluation is derived from empirical science and cognitive and experimental psychology [12].

Layered approach: The layered evaluation approach [13][14] separates the 'interaction assessment' and the 'adaptation decision'. Both layers should be evaluated separately in order to effectively interpret the evaluation results. Evaluating AHS on a layer by layer basis has been recommended as a more comprehensive approach [14][15]. In contrast to approaches that focus on the overall user's performance and satisfaction [16], layered evaluation in particular assesses the success of adaptation by decomposing it into different layers and evaluating each layer individually. This has a number of advantages over other approaches, such as useful insight into the success or failure of each separate adaptation stage, facilitation of improvements, generalization of evaluation results and re-use of successful practices.

Utility-based approach: Current evaluation practices attempt to evaluate adaptation as a whole, with user satisfaction or performance as the overall metric for success, based on identified measurable criteria. In the utility-based approach the evaluation can be seen as a utility function X that maps a system, given some user context, to a quantitative representation of user satisfaction or performance. For example if one compares an adaptive system with its non-adaptive counterpart, the value of adaptation is the difference in utility between the two systems.

As described above, the main advantage of layered evaluation methods are that they break the utility function into several distinct functions. For example suppose there is a utility X_1 that maps the interaction assessment and the resulting user model to a real number that represents its correctness. Suppose there is also a

[2] The Cross Language Evaluation Forum - http://www.clef-campaign.org/

[3] The National Institute of Informatics Test Collection for IR Systems - http://research.nii.ac.jp/ntcir/

utility function U_2 that maps a system, given some user model, to a real number that represents user satisfaction or performance. In this case the whole utility function can be expressed as $X = X_1 X_2$. It is clear that the latter utility function better indicates the usability of an adaptive hypermedia system. Utility-based evaluation of adaptive systems [17] offers a perspective of how to reintegrate the different layers.

Heuristic approach: A heuristic is a general principle or rule of thumb that can be used to critique existing decisions or guide a design decision. An approach which integrates layered evaluation and heuristic evaluation has been proposed [18]. The use of heuristics ensures that the entire system can be evaluated in depth and specific problems can be discovered at an early design stage before releasing a running prototype of a system [19]. This approach can help evaluators by improving the detection and diagnosis of potential usability problems.

2.2.2 Challenges in Evaluating of Adaptive Systems

The evaluation of AHS is a difficult task due to the complexity of such systems, as shown by many studies [20][21]. It is of crucial importance that the adaptive features of the system can be easily distinguished from the general usability of the designed tool. Issues arise in the selection of applicable criteria for the evaluation of adaptivity. Many metrics can be used to measure performance, for example: knowledge gain (AEHS), amount of requested materials, duration of interaction, number of navigation steps, task success, usability (e.g., effectiveness, efficiency and user satisfaction). The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls when evaluating adaptive systems. Examples of pitfalls [22] include:

- Difficulty in attributing cause: is the adaptation causing the measured effect or another aspect of system functionality or design (e.g. system usability).
- Statistically insignificant results: Adaptivity is typically used when individual users differ. However, differences in approach and preferences are likely to lead to a large variance in performance results, which makes it more difficult to produce statistically comparable results. In order to produce significant results, large volumes of queries and users are required. There are few general guidelines for the selection of these measurements.
- Difficulty in defining the effectiveness of adaptation: It can be difficult to define what constitutes a useful or helpful adaptation.
- Insufficient resources: To fully evaluate an adaptive system it is often necessary to have a large number of individuals interacting with the system. This is in part due to the expected variance between participants mentioned above.
- Too much emphasis on summative rather than formative evaluation: Evaluations often measure only how good or bad a system is rather than providing information on where the problems are and how a system can be improved.

The selection of the metrics to be used in the evaluation of AHS is crucial. There are currently no agreed evaluation methodology standards, thus making AH evaluation a difficult, complex and time consuming task**.**

3. EVALUATION METHODOLOGY

There are many methods of combining the techniques used in AH and IR. However, in order to define a common evaluation mechanism, we classify a content composition as the output from an AIRS. A content composition is an aggregated set of resources ordered according to the user's needs and preferences.

In order to sufficiently evaluate both the adaptive functionality and the retrieval performance of an AIRS, a hybrid approach is necessary. This involves user-centric assessment, layered evaluation of the personalisation which has been applied and quantitative performance metrics relating to the content delivered.

The layered approach to evaluation would allow the assessment of each individual aspect of personalisation applied within the AIRS to tailor the experience delivered to the user. By assessing each piece of functionality in isolation, it can be determined which provide the most value in relation to the experience delivered to the user. The layers which must be evaluated will differ for each system as the adaptation and personalization techniques used will vary depending on the range of models and adaptive presentation modalities available in the system.

There are a set of necessary elements of a hybrid AIRS evaluation model which can be defined irrespective of the system being evaluated. The key challenge is to be able to adequately combine the data-driven approach to assessing retrieval from IR evaluation with the more user-focused approach to evaluation from AH. The hybrid framework for AIRS evaluation proposed by this paper is as follows:

1. **Evaluating Query Formation.** Traditional IR methods are driven by keyword-driven queries. However, with the introduction of query term personalization and expansion, the effectiveness of these components must be evaluated. This can be achieved by assessing the correctness of the inference of a users' intention and information need.
2. **Evaluating Retrieval Effectiveness.** This is an assessment of the AIRS system in terms of traditional retrieval tasks. There are some additional constraints, however, as systems which use adaptive presentation methods are required to retrieve ordered sequences of content, which creates a more complex dependency than a list of independent candidate documents.
3. **User-Centric Evaluation.** A user-centric evaluation of the system is necessary to assess the assumptions made and adaptivity performed. Key metrics in user-centric approaches to evaluation include satisfaction, effectiveness and efficiency.
   a. **Evaluating Adaptivity Effectiveness and Efficiency**. The effectiveness of adaptivity performed by the system can be measured using domain appropriate performance indicators, for example, knowledge gain and knowledge retention in educational scenarios and information need fulfillment in IR. Different metrics can be assessed depending on the nature of the AIRS application and its use. This is necessarily a user-driven evaluation, because the behavior of the system depends on the properties of the user. Efficiency can be measured by examining user performance when conducting a set of defined tasks. Metrics measured can include time taken. Time taken can include two types of measurements; classical IR evaluations such as time to get the first relevant document and number of documents retrieved in exactly 10mn [23], number of queries needed etc.
   b. **User Satisfaction.** User satisfaction can be examined by eliciting direct user feedback. While there are many

means of eliciting such information, among the most widely used are qualitative questionnaires such as the System Usability Score (SUS) [24].

4. **System efficiency.** The creation of complex AIRS requires a wide variety of content and metadata. As such, the overall cost of provisioning a particular AIRS, and the performance of that system in terms of resource requirements and responsiveness must be assessed. This is particularly important as systems approach mass, web-scale use.

4. SUMMARY

This paper has presented an overview of the evaluation approaches commonly used in the IR and AH communities, with the objective of finding a means of combining these approaches to assess the next generation of personalization in IR. These approaches have traditionally varied in focus: the majority of IR evaluation is focused on the numerical performance and effectiveness of the retrieval of documents, while AH systems are evaluated with the users as an integral part.

The layered evaluation model proposed in this paper is an attempt to draw from the AH community's experience with the complex dependencies of different intelligent components in AHS. This is combined with traditional IR evaluation methods to gain a broad spectrum assessment of the entire system.

User-driven evaluation of IR is not new in itself, but the effects of personalization on creating reproducible, large scale experiments can, in the opinion of the authors, best be addressed by incorporating elements of AH evaluation techniques.

5. ACKNOWLEDGMENTS

This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE/I1142) as part of the Centre for Next Generation Localization (www.cngl.ie). The authors are grateful for the suggestions of the reviewers for this paper.

6. REFERENCES

[1] G. J. F. Jones & V. Wade. Integrated content presentation for multilingual and multimedia information access. In F.C. Gey, N. Kando, C.Y. Lin and C. Peters (eds.), New Directions in Multilingual Information Access, vol. 40, pp. 31–39, 2006.

[2] B., Steichen, S., Lawless, A., O'Connor & V., Wade. Dynamic Hypertext Generation for Reusing Open Corpus Content. In the Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, Hypertext 2009, Torino, Italy. 29th June – 1st July, 2009.

[3] C. W. Cleverdon, J. Mills, & M. Keen. Factors determining the performance of indexing systems. Vol. 1 - Design. ASLIB Cranfield Project. Technical Report, 1966.

[4] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[5] K., van Rijsbergen. Information Retrieval. London, England, Butterworths & Co. Ltd. 1979.

[6] S., Chakrabarti, M., van den Berg & B., Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In The International Journal of Computer and Telecommunications Networking, Vol. 31(11-16), Elsevier, pp. 1623-1640, May 1999.

[7] C., Hölscher & G., Strube. Web Search Behavior of Internet Experts and Newbies. In Proceedings of the 9th International Conference on the World Wide Web, WWW9, Amsterdam, The Netherlands. pp 337-346, 2000.

[8] P. Brusilovsky, E. Schwarz & G. Weber. A Tool for Developing Adaptive Electronic Textbooks on WWW, In the Proceedings of WebNet '96 World conference of the web society, pp. 64-69, 1996.

[9] C. Gena & S. Weibelzahl. Usability Engineering for the Adaptive Web, vol. 4321, LNCS, pp. 720-762, 2007.

[10] M. De Jong & P. Schellens. Reader-Focused Text Evaluation. An overview of goals and methods, vol. 11(4), pp. 402-432, 1997.

[11] J. Nielsen. Usability Engineering, Boston: MA: Academic Press, 1993.

[12] C. Gena. Methods and techniques for the evaluation of user-adaptive systems, The knowledge engineering review, vol 20:1, pp. 1-37, United Kingdom: Cambridge University Press, 2005.

[13] C. Karagiannidis & D. Sampson. Layered evaluation of adaptive applications and services, International Conference on adaptive hypermedia and adaptive applications and services, 2000.

[14] P. Brusilovsky, C. Karagiannidis & D. Sampson. The benefits of layered evaluation of adaptive applications and services, In the proceedings of the first workshop on empirical evaluation of adaptive systems, UM2001, Sonthofen, Germany, pp. 1-8, 2001.

[15] S. Weibelzahl & G. Weber. Advantages, opportunities and limits of empirical evaluations, Evaluating adaptive systems, vol. 3(2), 2002.

[16] D. Chin. Empirical evaluation of user models and user-adapted systems, pp. 181-194, 2001.

[17] E. Herder. Utility-based evaluation of adaptive systems, In the proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International Conference on User Modeling, UM2003, Pittsburg, USA, pp. 25-30, 2003.

[18] G. Magoulas, S. Chen & K. Papanikolaou. Integrating layered and heuristic evaluation for adaptive learning environments, In the proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International Conference on User Modeling, UM2003, Pittsburg, USA, pp. 5-14, 2003.

[19] L. Fu, G. Salvendy & L. Turley. Effectiveness of user testing and heuristic evaluation as a function of performance classification. Behaviour and information technology, vol. 21, pp. 137-143, 2002.

[20] F. Del Missier & F. Ricci. Understanding recommender systems: Experimental evaluation challenges, pp. 31-40, 2003.

[21] T. Lavie, J. Meyer, K. Beugler & J. Coughlin, The evaluation of in-vehicle adaptive systems, User Modeling: Work on the EAS, pp. 9-18, 2005.

[22] N. Tintarev and J. Masthoff, "Evaluating Recommender Explanations: Problems Experienced and Lessons Learned for the," *UMAP 2009,* p. 54, 2009.

[23] E. Crestan, C. de Loupy, "Browsing Help for a Faster Retrieval; COLING 2004," 2004, pp. 576-582; Geneve, Suisse.

[24] J. Brooke. SUS: a "quick and dirty" usability scale. In Usability Evaluation in Industry, P.W. Jordan, B. Thomas, B.A. Weerdmeester & A.L. McClelland (eds.), London: Taylor and Francis, 1996

On Search Topic Variability in Interactive Information Retrieval

Ying-Hsang Liu
School of Information Studies
Charles Sturt University
Wagga Wagga NSW 2678, Australia
+61 2 6933 2171
yingliu@csu.edu.au

Nina Wacholder
School of Communication and Information
Rutgers University
New Brunswick NJ 089091, USA
+1 732 932 7500 ext. 8214
ninwac@rutgers.edu

ABSTRACT

This paper describes the research design and methodologies we used to assess the usefulness of MeSH (Medical Subject Headings) terms for different types of users in an interactive search environment. We observed four different kinds of information seekers using an experimental IR system: (1) search novices; (2) domain experts; (3) search experts and (4) medical librarians. We employed a user-oriented evaluation methodology to assess search effectiveness of automatic and manual indexing methods using TREC Genomics Track 2004 data set. Our approach demonstrated (1) the reusability of a large test collection originally created for TREC, (2) an experimental design that specifically considers types of searchers, system versions and search topic pairs by Graeco-Latin square design and (3) search topic variability can be alleviated by using different sets of equally difficult topics and well-controlled experimental design for contextual information retrieval evaluation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*query formulation*, *search process*

General Terms

Measurement, Human Factors

Keywords

Information retrieval evaluation, Search topic variability, interactive information retrieval

1. INTRODUCTION

The creation and refinement of test design and methodologies for IR system evaluation have been one of the greatest achievements in IR research and development. In the second Cranfield project [6], the main purpose is to evaluate the effectiveness of indexing techniques at a level of abstraction where users are not specifically considered in a batch mode experiment.

The test design and methodology following the Cranfield paradigm culminated in the TREC (Text REtrieval Conference) activities since the 1990s. TREC has provided a research forum for comparing the search effectiveness of different retrieval techniques across IR systems in a laboratory and controlled environment [30]. The very large test collection used in TREC provided a test bed for researchers to experiment the scalability of retrieval techniques, which had not been possible in previous years. However, how we specifically take into account different aspects of user contexts within a more realistic test environment has been challenging in part because it is difficult to isolate the effects of user, search topic and system in IR experiments (see e.g., [7, 17] for recent efforts).

In batch experiments the search effectiveness of different retrieval techniques is achieved by comparing the search performance of queries. IR researchers have widely used the micro-averaging method of performing statistics on the queries in summarizing precision and recall values for comparing the search effectiveness of different retrieval techniques in order to meet the statistical requirements (see e.g., [25, 27]). The method of micro-averaging is intended to obtain reliable results in comparing search performance of different retrieval techniques by giving equal weights to each query.

However, within an interactive IR search environment that involves human searchers, it is difficult to use a large set of search topics. Empirical evidence has demonstrated that the search topic set size of 50 is necessary to determine the relative performance of different retrieval techniques in batch evaluations [3], because the variability of search topics has an overriding effect on search results. Another possible solution is to use different sets of topics in a non-matched-pair design [5, 21, 22], but theoretically it requires a very large sample of independent searches.

This problem has been exacerbated by the fact that we have little theoretical understanding about the nature and properties of search topics for evaluation purposes [20]. From a systems perspective, recent in-depth failure analyses of variability in search topics for reliable and robust retrieval performance (e.g., [11, 28]) have contributed to our preliminary understanding of how and why IR systems fail to do well across all search topics. It is still elusive what kinds of search topics can be used to directly control the topic effect for IR evaluation purposes.

This study was designed to assess the search effectiveness of MeSH terms by different types of searchers in an interactive search environment. By an experimental design that controls searchers, system versions and search topic pairs and the use of a relatively large number of search topics, we were able to demonstrate an IR user experiment that specifically controls the

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

search topic variability and assesses the user effect on search effectiveness within the laboratory IR framework (see e.g., [14, 15] for recent discussions).

2. METHOD

Thirty-two searchers from a major public university and nearby medical libraries in the northeast area of the US participated in the study. Each searcher belonged to one of four groups: (1) Search Novice (SN), (2) Domain Experts (DE), (3) Search Experts (SE) and (4) Medical Librarians (ML).

The experimental task was to conduct a total of eight searches to help biologists conduct their research. Participants searched either using a version of the system in which abstracts and MeSH terms were displayed (MeSH+) or another version in which they had to formulate their own terms based only on the display of abstracts (MeSH–). Participants conducted four searches each with two different systems: in one, they browsed a displayed list of MeSH terms (MeSH+) and in the other (MeSH–). Half the participants used MeSH+ system first; half used MeSH– first. Each participant was allowed to conduct searches on eight different topics.

The experimental setting for most searchers was a university office; for some searchers, it was a medical library. Before they began searching participants were briefly trained in how to use the MeSH terms. We kept search logs that recorded search terms, a ranked list of retrieved documents, and time-stamps.

2.1 Subjects

We used the purposive sampling method for recruiting our subjects since we were concerned with the impact of specific searcher characteristics on search effectiveness. The key searcher characteristics were different levels of domain knowledge in the biomedical domain and whether they had substantial search training. The four types of searchers were distinguished by their levels of domain knowledge and search training.

2.2 Experimental design

The experiment was a 4×2×2 factorial design with four types of searchers, two versions of an experimental system and controlled search topic pairs. The versions of a system, types of searchers (distinguished by levels of domain knowledge and search training) and search topic pairs were controlled by a Graeco-Latin square balanced design [8]. The possible ordering effects have been taken into account by the design. The requirement for this experimental design is that the examined variables do not interact and each variable has the same number of levels [16]. The treatment layout of a 4×4 Graeco-Latin square design is illustrated in Figure 1.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SN	DE	SE	ML	DE	SN	ML	SE	SE	ML	SN	DE	ML	SE	DE	SN
38	12	29	50	38	12	27	45	29	50	27	45	42	46	9	36
12	38	50	29	12	45	38	27	50	29	29	27	46	36	42	9
29	50	12	38	27	38	45	12	27	45	45	50	9	42	36	46
50	29	38	12	45	27	12	38	45	27	50	29	36	9	46	42
42	46	32	15	9	36	30	20	2	43	1	49	2	43	33	23
46	42	15	42	36	9	20	30	43	1	49	2	43	2	23	33
32	15	42	46	30	20	9	36	1	49	2	43	33	23	2	43
15	32	46	32	20	30	36	9	49	2	43	1	23	33	43	2

Note. Numbers 1-16 refers to participant ID; SN, DE, DE and ML refer to types of searchers, SN=Search Novices, DE=Domain Experts; SE=Search Experts; ML=Medical Librarians; Shaded and non-shaded blocks refer to MeSH**+** and MeSH**–** versions of an experimental system; Numbers in blocks refer to search topic ID number from TREC Genomics Track 2004 data set; 10 search topic pairs, randomly selected from a pool of 20 selected topics, include (38, 12), (29, 50), (42, 46), (32, 15), (27, 45), (9, 36), (30, 20), (2, 43), (1, 49) and (33, 23).

Figure 1. 4×4 Graeco-Latin square design

Because of the potential interfering effect of search topic variability on search performance in IR evaluation, we used a design that included relatively large number of search topics. In theory, the effect of topic variability and topic-system interaction on system performance could be eliminated by averaging the performance scores of the topics (micro-averaging method), together with the use of very large number of search topics. The TREC standard ad hoc task evaluation studies ([1, 3]) and other proposals of test collections (e.g., [20-22, 24, 29]) have been concerned with the large search topic variability in batch experiments. However, in a user-centered IR experiment it is not feasible to use as many as 50 search topics because of human fatigue.

We controlled search topic pairs by a balanced design in order to alleviate the overriding effect of search topic variability. We assumed that all the search topics are equally difficult, since we do not have a good theory about what makes some search topics more difficult than others. By design we ensured that each search topic pair was assigned to all types of searchers and was searched at least two times by the same type of searchers. This design required a total of 10 search topic pairs and a minimum of 16 participants.

2.3 Search tasks and incentive system

The search task was designed to simulate online searching situations in which professional searchers look for information on behalf of users. We decided to use this relatively challenging task for untrained searchers because choosing realistic tasks such as this one would enhance the external validity of the experiment. Considering the relatively difficult tasks, we were concerned that searchers may have problems completing all searches. Because research literature has suggested that the motivational characteristics of participants are possible sources of sample bias [23], we designed an incentive system to motivate the searchers.

We promised monetary incentives according to the participant's search effectiveness. Each subject was paid $20 for participating and was also paid up to $10.00 dollars more based on the average number of relevant documents in the top ten search results across all search topics; on average each participant received an additional $4.40, with a range of $2.00 - $8.00.

2.4 Experimental procedures

After signing the consent form, the participant filled out a searcher background questionnaire before the search assignment. After a brief training session, they were assigned to one of the arranged experimental conditions and conducted search tasks. They completed a search perception questionnaire and were asked to indicate the relevance of two pre-judged documents when they were done with each search topic. A brief interview was conducted when they finished all search topics. Search logs with search terms and ranked retrieved documents were recorded.

The MeSH Browser [19], an online vocabulary look-up aid, prepared by U.S. National Library of Medicine, was designed to help searchers find appropriate MeSH terms and display hierarchy of terms for retrieval purposes. The MeSH Browser was only available when participants were assigned to the MeSH+ version of an experimental system; in the MeSH– version, participants had to formulate their own terms without the assistance of MeSH Browser and displayed MeSH terms in bibliographic records.

Because we were concerned that the topics were so hard that even the medical librarians would not understand them, we used a questionnaire regarding search topic understanding after each topic. The testing items of two randomly selected pre-judged documents, one definitely relevant and the other definitely not relevant, were prepared from the data set [26].

Each search topic was allocated up to ten minutes. The last search within the time limit was used for calculating search performance. To keep the participants motivated and reward their effort, they were asked to orally indicate which previous search result would be the best answer when the search task was not finished within ten minutes.

2.5 Experimental system

For this study, it was important for participants to conduct their searches in a carefully controlled environment; our goal was to offer as much help as possible while still making sure that the help and search functions did not interfere with our ability to measure the impact of the MeSH terms. We built an information retrieval system based on the Greenstone Digital Library Software version 2.70 [9] because it provides reliable search functionality, customizable search interface and good documentation [31].

We prepared two different search interfaces using a single system using Greenstone: MeSH+ and MeSH– versions. One interface allowed users to use MeSH terms; the other required them to devise their own terms. One interface displayed MeSH terms in retrieved bibliographic records and the other did not. Because we were concerned that the participant responds to the cue that may signal the experimenter's intent, the search interfaces were termed 'System Version A' and 'System Version B' for 'MeSH+ Version' and 'MeSH– Version' respectively (see http://comminfo.rutgers.edu/irgs/gsdl/cgi-bin/library/). The MeSH– version was used as baseline system for an automatic indexing system, whereas the MeSH+ version served as performance of a manual indexing system. That is, MeSH terms added another layer of document representation to the MeSH+ version.

The experimental system was constructed as Boolean-based system with ranked functions by the TF×IDF weighting rule [32]. More specifically, MGPP (MG++), a re-implementation of the mg (Managing Gigabytes) searching and compression algorithms, was used as indexing and querying indexer. Basic system features, including fielded searching, phrase searching, Boolean operators, case sensitivity, stemming and display of search history, were sufficient to fulfill the search tasks. The display of search history was necessary because it provided useful feedback regarding the magnitude of retrieved documents for difficult search tasks that usually required query reformulations.

Since our goal was specifically to investigate the usefulness of displayed MeSH terms, we deliberately refrained from implementing certain system features that allow users to take advantage of the hierarchical structures of MeSH terms, such as the hyperlinked MeSH terms, explode function that automatically includes all narrower terms and automatic query expansion (see e.g. [13, 18]) available on other online search systems. The use of those features would have invalidated the results by introducing other variables at the levels of search interface and query processing, although a full integration of those system features would have increased the usefulness of MeSH terms.

2.6 Documents

The experimental system was set up on a server, using bibliographic records from the 2004 TREC Genomics document set [26]. TREC Genomics Track 2004 Data Set document test collection was a 10-year (from 1994 to 2003) subset of MEDLINE with a total of 4,591,108 records. The test collection subset fed into the system used 75.0% of the whole collection, a total of 3,442,321 records, excluding the records without MeSH terms or abstracts.

We prepared two sets of documents for setting up the experimental system: MeSH+ and MeSH– versions. One interface allowed users to use MeSH terms; the other did not provide this search option. The difference was also reflected in retrieved bibliographic records.

2.7 Search topics

The search topics used in this study were originally created for TREC Genomics Track 2004 for the purpose of evaluating the search effectiveness of different retrieval techniques (see Figure 3-9 for an example). They covered a range of genomics topics typically asked by biomedical researchers. Besides a unique ID number for each topic, the topic was constructed in a format that included the title, need and context fields. The title field was a short query. The need field was a short description of the kind of material the biologists are interested in, whereas the context field provides background information for judging the relevance of documents. The need and context fields were designed to provide more possible search terms for system experimentation purposes.

> ID: 39
> Title: Hypertension
> Need: Identify genes as potential genetic risk factors candidates for causing hypertension.
> Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.

Figure 2. Sample search topic

Because of the technical nature of genomics topics, we wondered whether the search topics could be understood by human searchers, particularly for those without advanced training in the biomedical field. TREC search topics were designed for machine runs with little or no consideration for searches by real users. We selected 20 of the 50 topics using the following procedure:

1. Consulting an experienced professional searcher with biology background and a graduate student in neuroscience, to help make a judgment as to whether the topics would be comprehensible to the participants who were not domain experts. Topics that used advanced technical vocabulary, such as specific genes, pathways and mechanisms, were excluded;
2. Ensuring that major concepts in search topics could be mapped to MeSH by searching the MeSH Browser. For instance, topic 39 could be mapped to MeSH preferred terms hypertension and risk factors;
3. Eliminating topics with very low MAP (mean average precision) and P10 (precision at top 10 documents) score in the relevance judgment set because these topics would be too difficult;

The selected topics were then randomly ordered to create ten search topic pairs for the experimental conditions (see Figure 1 for search topic pairs).

2.8 Reliability of relevance judgment sets

We measured search outcome using standard precision and recall measures for accuracy and time spent for user effort [6] because we were concerned with the usefulness of MeSH terms on search effectiveness by using TREC assessments [12].

Theoretically speaking, the calculation of recall measure requires relevance judgments from the whole test collection. However, it is almost impossible to obtain these judgments from a test collection with more than 3 million documents. For practical reasons the recall measure used a pooling method that created a set of unique documents from the top 75 documents submitted by 27 groups participated in the TREC 2004 Genomics Track ad hoc tasks [26]. Empirical evidence has shown that recall calculated with a pooling method provides a reasonable approximation, although the recall is likely to be overestimated [33]. But as a result of this approach, there was an average pool size of 976 documents, with a range of 476-1450, which had relevance judgments for each topic [12].

It was quite likely that some of the participants in our experiment would retrieve documents that had not been judged. The existence of un-judged relevant documents, called sampling bias in pooling method, is concerned with the pool depth and the diversity of retrieval methods that may affect the reliability of relevance judgment set [2]. The assumption that the pooled judgment set is a reasonable approximation of complete relevance judgment set may become invalid when the test collection is very large.

To ensure that the TREC pooled relevance judgment set was sufficiently complete and valid for the current study, we analyzed top 10 retrieved documents from each human runs (32 searchers × 8 topics = 256 runs). Cross-tabulation results showed that about one-third of all documents retrieved in our study had not been judged in the TREC data set. More specifically, for a total of 2277 analyzed documents, 762 (33.5 %) had not been assigned relevant judgments. There existed large variations in percentage of un-judged documents for each search topic, with a range of 0–59.3%.

To assess the impact of incomplete relevance judgments, we compared the top 10 ranked search results between the judged document set and the pooled document set for each topic. The judged document set was composed of the documents that matched TREC data, i.e., combination of judged not relevant and judged relevant. The un-judged documents, added to the pooled document set, were considered 'not relevant' in our calculations of search outcome. We used precision oriented measures, MAP (mean average precision), P10 (precision at top 10 documents) and P100 (precision at top 100 documents) to estimate the impact of incomplete judgments.

The paired t-test results by search topic revealed significant differences between the two sets in terms of MAP ($t(19) = -3.69, p < .01$), P10 ($t(19) = -3.89, p < .001$) and P100 ($t(19) = -3.95, p < .001$) measures. The mean of the differences for MAP, P10 and P100 was approximately 2.7%, 9.9% and 4.9% respectively. We concluded that the TREC relevance judgments are applicable to this study.

2.9 Limitations of the design

This study was designed to assess the impact of MeSH terms on search effectiveness in an interactive search environment. One limitation of the design was that participants were a self-selected group of searchers that may not be representative of the population. The interaction effects of selection biases and the experimental variable, i.e., the displayed MeSH terms, were another possible factor that limits the generalizability of this study [4]. The use of relatively technical and difficult search topics in the interactive search environment posed threat to external validity, since those topics might not represent typical topics received by medical librarians in practice.

The internal validity of this design was enhanced by specifically considering several aspects: We devised an incentive system to consider the possible sampling bias of searchers' motivational characteristics in experimental settings. Besides levels of education, participants' domain knowledge was evaluated by a topic understanding test. The variability of search topics was alleviated by using a relatively large number of search topics by experimental design. Selected search topics were intelligible in consultation with domain expert and medical librarian. A concept analysis form was used to help searchers recognize potentially useful terms. The reliability of relevance judgment sets was ensured by additional analysis of top 10 search results from our human searchers.

3. DISCUSSION AND CONCLUSION

The Cranfield paradigm has been very useful for comparing search effectiveness of different retrieval techniques at the level of abstraction that simulates user search performance. Putting users in the loop of IR experiments is particularly challenging because it is difficult to separate the effects of systems, searchers and topics and the search topics have had dominating effects [17]. To alleviate search topic variability in interactive IR experiments, we consider how to increase the topic set size by experimental design within the laboratory IR framework.

This study has demonstrated that a total of 20 search topics can be used in an interactive experiment by Graeco-Latin square balanced design and using different sets of carefully selected topics. We assume that the selected topics are equally difficult since we do not have a good theory of search topics that can directly control the topic difficulty for evaluation purposes. Recent attempts to use reduced topic sets and use non-matched topics (see e.g., [5, 10]) indirectly support our experimental design considerations of search topic variability and topic difficulty. However, an important theoretical question remains. How can we better control the topic effects in batch and user IR experiments?

4. ACKNOWLEDGMENTS

This study was funded by NSF grant #0414557, PIs. Michael Lesk and Nina Wacholder. We thank anonymous reviewers for their constructive comments.

5. REFERENCES

[1] Banks, D., Over, P. and Zhang, N.-F. 1999. Blind men and elephants: Six approaches to TREC data. Inform Retrieval, 1, 1/2 (April 1999), 7-34. DOI=http://dx.doi.org/10.1023/A:1009984519381

[2] Buckley, C., Dimmick, D., Soboroff, I. and Voorhees, E. 2007. Bias and the limits of pooling for large collections. Inform Retrieval, 10, 6 (December 2007), 491-508. DOI=http://dx.doi.org/10.1007/s10791-007-9032-x

[3] Buckley, C. and Voorhees, E. M. 2005. Retrieval system evaluation. In Voorhees, E. M. and Harman, D. K. (Eds.), TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, Cambridge, MA, 53-75.

[4] Campbell, D. T., Stanley, J. C. and Gage, N. L. 1966. Experimental and Quasi-Experimental Designs for Research. R. McNally, Chicago.

[5] Cattelan, M. and Mizzaro, S. 2009. IR evaluation without a common set of topics. In Proceedings of the 2nd International Conference on the Theory of Information Retrieval (Cambridge, UK, September 10-12, 2009). ICTIR 2009. Springer, Berlin, 342-345. DOI=http://dx.doi.org/10.1007/978-3-642-04417-5_35

[6] Cleverdon, C. W. 1967. The Cranfield tests on index language devices. Aslib Proc, 19, 6 (1967), 173-193. DOI=http://dx.doi.org/10.1108/eb050097

[7] Dumais, S. T. and Belkin, N. J. 2005. The TREC Interactive Track: Putting the user into search. In Voorhees, E. M. and Harman, D. K. (Eds.), TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, Cambridge, MA, 123-152.

[8] Fisher, R. A. 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.

[9] Greenstone Digital Library Software (Version 2.70). 2006. Department of Computer Science, The University of Waikato, New Zealand. Available at: http://prdownloads.sourceforge.net/greenstone/gsdl-2.70-export.zip

[10] Guiver, J., Mizzaro, S. and Robertson, S. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. ACM Trans. Inf. Syst., 27, 4 (November 2009), 1-26. DOI=http://doi.acm.org/10.1145/1629096.1629099

[11] Harman, D. and Buckley, C. 2009. Overview of the Reliable Information Access Workshop. Inform Retrieval, 12, 6 (December 2009), 615-641. DOI=http://dx.doi.org/10.1007/s10791-009-9101-4

[12] Hersh, W., Bhupatiraju, R., Ross, L., Roberts, P., Cohen, A. and Kraemer, D. 2006. Enhancing access to the Bibliome: The TREC 2004 Genomics Track, Journal of Biomedical Discovery and Collaboration, 1, 3 (March 2006). DOI=http://dx.doi.org/10.1186/1747-5333-1-3

[13] Hersh, W. R. 2008. Information Retrieval: A Health and Biomedical Perspective. Springer, New York.

[14] Ingwersen, P. and Järvelin, K. 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Dordrecht.

[15] Ingwersen, P. and Järvelin, K. 2007. On the holistic cognitive theory for information retrieval. In Proceedings of the First International Conference on the Theory of Information Retrieval (ICTIR) (Budapest, Hungary, 2007). Foundation for Information Society.

[16] Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. 1995. Brooks/Cole, Pacific Grove, CA.

[17] Lagergren, E. and Over, P. 1998. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, 1998). SIGIR '98. ACM Press, New York, NY, 164-172. DOI=http://doi.acm.org/10.1145/290941.290986

[18] Lu, Z., Kim, W. and Wilbur, W. Evaluation of query expansion using MeSH in PubMed. Inform Retrieval, 12, 1 (February 2009), 69-80. DOI=http://dx.doi.org/10.1007/s10791-008-9074-8

[19] MeSH Browser (2003 MeSH). 2004. U.S. National Library of Medicine. Available at: http://www.nlm.nih.gov/mesh/2003/MBrowser.html

[20] Robertson, S. E. 1981. The methodology of information retrieval experiment. In Sparck Jones, K. (Ed.), Information Retrieval Experiment, Butterworth, London, 9-31.

[21] Robertson, S. E. 1990. On sample sizes for non-matched-pair IR experiments. Inform Process Manag, 26, 6 (1990), 739-753. DOI=http://dx.doi.org/10.1016/0306-4573(90)90049-8

[22] Robertson, S. E., Thompson, C. L. and Macaskill, M. J. 1986. Weighting, ranking and relevance feedback in a front-end system. Journal of Information and Image Management, 12, 1/2, (January 1986), 71-75. DOI=http://dx.doi.org/10.1177/016555158601200112

[23] Sharp, E. C., Pelletier, L. G. and Levesque, C. 2006. The double-edged sword of rewards for participation in psychology experiments. Can J Beh Sci, 38, 3 (Jul 2006), 269-277. DOI=http://dx.doi.org/10.1037/cjbs2006014

[24] Sparck Jones, K. and van Rijsbergen, C. J. 1976. Information retrieval test collections. J Doc, 32, 1 (March 1976), 59-75. DOI=http://dx.doi.org/10.1108/eb026616

[25] Tague-Sutcliffe, J. 1992. The pragmatics of information retrieval experimentation, revisited. Inform Process Manag, 28, 4 1992), 467-490. DOI=http://dx.doi.org/10.1016/0306-4573(92)90005-K

[26] TREC 2004 Genomics Track document set data file. 2005. Available at http://ir.ohsu.edu/genomics/data/2004/

[27] van Rijsbergen, C. J. 1979. Information Retrieval. Butterworths, London.

[28] Voorhees, E. M. 2005. The TREC robust retrieval track. SIGIR Forum, 39, 1 (June 2005), 11-20. DOI=http://doi.acm.org/10.1145/1067268.1067272

[29] Voorhees, E. M. On test collections for adaptive information retrieval. Inform Process Manag, 44, 6 (November 2008), 1879-1885. DOI=http://dx.doi.org/10.1016/j.ipm.2007.12.011

[30] Voorhees, E. M. and Harman, D. K. 2005. TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge, MA.

[31] Witten, I. H. and Bainbridge, D. 2007. A retrospective look at Greenstone: Lessons from the first decade. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (Vancouver, Canada, June 18-23, 2007). JCDL '07. ACM Press, New York, NY, 147-156. DOI=http://doi.acm.org/10.1145/1255175.1255204

[32] Witten, I. H., Moffat, A. and Bell, T. C. 1999. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco.

[33] Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, 1998). SIGIR '98. ACM Press, New York, NY, 307-314. DOI=http://doi.acm.org/10.1145/290941.291014