SIGIR 2010

Geneva, Switzerland
July 19-23, 2010

Simulation of Interaction

Automated Evaluation of Interactive IR

Workshop of the 33rd Annual International
ACM SIGIR Conference
on Research and Development
in Information Retrieval

Organised by
Leif Azzopardi
Kalervo Järvelin
Jaap Kamps
Mark D. Smucker

Association for
Computing Machinery

Proceedings of the SIGIR 2010 Workshop on the

Simulation of Interaction

Automated Evaluation of Interactive IR

Leif Azzopardi, Kalervo Järvelin, Jaap Kamps,
Mark D. Smucker (editors)

July 23, 2010
Geneva, Switzerland

Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction. Held in Geneva, Switzerland. July 23, 2010.
Published by: IR Publications, Amsterdam. ISBN 978-90-814485-3-6.

Preface

These proceedings contain the posters of the SIGIR 2010 Workshop on the Simulation of Interaction, Geneva, Switzerland, on 23 July, 2010. The workshop will consist of three main parts:

- First, a set of keynotes by Donna Harman and Ryen White that help frame the problems, and outline potential solutions.
- Second, a poster and discussion session with seventeen papers selected by the program committee from 22 submissions (a 77% acceptance rate). Each paper was reviewed by at least two members of the program committee.
- Third, break out sessions on different aspect of the simulation of interaction with reports being discussed in the final slot.

When reading this volume it is necessary to keep in mind that these papers represent the ideas and opinions of the authors (who are trying to stimulate debate). It is the combination of these papers and the debate that will make the workshop a success.

We would like to thank ACM and SIGIR for hosting the workshop, and Gianni Amati for his outstanding support in the organization. Thanks also go to the program committee, the authors of the papers, and all the participants, for without these people there would be no workshop.

July 2010

Leif Azzopardi
Kalervo Järvelin
Jaap Kamps
Mark D. Smucker

Organization

Program Chairs

Leif Azzopardi
Kalervo Järvelin
Jaap Kamps
Mark D. Smucker

Program Committee

Nicholas Belkin
Pia Borlund
Ben Carterette
Paul Clough
Georges Dupret
Donna Harman
Claudia Hauff
Evangelos Kanoulas
Jaana Kekäläinen
Diane Kelly
Heikki Keskustalo
Birger Larsen
Jeremy Pickens
Benjamin Piwowarski
Ian Ruthven
Mark Sanderson
Falk Scholer
Andrew Turpin
Ryen White
Max Wilson

Table of Contents

Preface III

Organization V

Table of Contents VII

Making Simulations Work

Simulation of the IIR User: Beyond the Automagic 1
Michael J. Cole

Bridging Evaluations: Inspiration from Dialog System Research 3
Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando

A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction 5
Catherine Mulwa, Wei Li, Séamus Lawless, Gareth Jones

Using QPP to Simulate Query Refinement 7
Daniel Tunkelang

Generating and Modeling Queries and Interaction

Focused Relevance Feedback Evaluation 9
Shlomo Geva, Timothy Chappell

Modeling the Time to Judge Document Relevance 11
Chandra Prakash Jethani, Mark Smucker

Session Track at TREC 2010 13
Evangelos Kanoulas, Paul Clough, Ben Carterette, Mark Sanderson

Graph-Based Query Session Exploration Based on Facet Analysis 15
Heikki Keskustalo, Kalervo Järvelin, Ari Pirkola

A Probabilistic Automaton for the Dynamic Relevance Judgement of Users 17
Peng Zhang, Ulises Cerviño Beresi, Dawei Song, Yuexian Hou

Creating Simulations using Search Logs

Creating Re-Useable Log Files for Interactive CLIR 19
Paul Clough, Julio Gonzalo, Jussi Karlgren

Simulating Searches from Transaction Logs 21
Bouke Huurnink, Katja Hofmann, Maarten de Rijke

A Methodology for Simulated Experiments in Interactive Search 23
Nikolaos Nanas, Udo Kruschwitz, M-Dyaa Albakour, Maria Fasli, Dawei Song, Yunhyong Kim, Ulises Cerviño Beresi, Anne De Roeck

Browsing and User Interfaces

Recovering Temporal Context for Relevance Assessments 25
Omar Alonso, Jan Pedersen

Simulating User Interaction in Result Document Browsing 27
Paavo Arvola, Jaana Kekäläinen

Query and Browsing-Based Interaction Simulation in Test Collections ... 29
Heikki Keskustalo, Kalervo Järvelin

Evaluating a Visualization Approach by User Simulation 31
Michael Preminger

Automatic Evaluation of User Adaptive Interfaces for Information Organization and Exploration 33
Sebastian Stober, Andreas Nuernberger

Author Index 35

Simulation of the IIR User: Beyond the Automagic

Michael J. Cole
School of Communication & Information
Rutgers, The State University of New Jersey
4 Huntington St.
New Bunswick, New Jersey 08901
m.cole@rutgers.edu

ABSTRACT

Simulation of human users engaged in interactive information retrieval (IIR) may be a key resource to enable evaluation of IR systems in interactive settings in ways that are scalable, objective, and cheap to implement. This paper considers generation of simulated users in an operationalist framework. It identifies several challenges due to the cognitive nature of IIR and its highly conditionalized interactions with information systems and suggests a program for user simulation must rely on results from user studies and will need to overcome several difficult modeling problems.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation

General Terms

Measurement, Experimentation, Human Factors

Keywords

User studies, simulation, interactive information retrieval

1. INTRODUCTION

The problems of addressing IIR within the TREC framework are a significant motivator of the interest in employing simulated users in a new IR system evaluation paradigm. The difficulties that make it hard for TREC to handle interaction leads one to ask about the nature of interaction with IR systems and how interaction makes system evaluation difficult. Why is it hard to model interaction? Does that also mean simulating users is fundamentally difficult?

This paper identifies challenges for simulation of real users and argues simulation must rest on studies of humans engaged in realistic tasks. The cognitive nature of IIR grounds issues that make difficult both modeling and identifying rules to simulate user decision-making. Five specific issues tied to the nature of interaction and the cognitive and interactive aspects of IIR are raised.

2. A COGNITIVE VIEW OF SIMULATION

Edwin Boring [3] outlined a five-step operationalist approach to providing an objective model of the mind's functions, free of the taint of mentalism:

(B1) Make an inventory of cognitive functions by analyzing mental capacities, for example *learning.*

(B2) Describe those functional capacities as properties of the human-as-system by describing the input-output pairs. That is, treat the mind as a black box and operationalize its functional properties.

(B3) Reformulate the functions as properties of an abstract system, e.g. a collection of Turing machines.

(B4) Design and realize physical artifacts by implementing the abstract machines. These things can then simulate the mind.

(B5) Show that mentalistic references have been removed by explaining the artifact's activity as a mechanical system.

The essence of this program must be achieved to simulate human IIR users, although the mental functions to be analyzed can be restricted to an appropriate information behavior realm, say exploratory search. To simulate real users, steps (B1), (B2), and (B5) are critical.

Translation of cognitive functions (B1) into input-output pairs (B2) is non-trivial because we must go from the user's intentions and plans to learning how mental functions are exercised in specific situations of knowledge, task, and cognitive constraints. Further, it is not enough to observe individual actions in a user situation, instead the observable outputs come in the form of coherent *action sequences* of variable length. This raises many practical issues for modeling in computationally-useful forms.

For Boring (B5) was crucial because he wanted to place human simulation on wholly objective grounds and the details of this explanatory process go to the heart of the shortcomings of operationalism in cognitive psychology. For IIR simulations (B5) is essential for evaluation and explanation. Ultimately, system performance depends on the capacity of the system to recognize a user's situation and provide optimal information access and support for the user actions to satisfy their task. To do this requires development of user models with explanatory value. That explanatory value must reside in the process of understanding how the information behavior functionality expressed in (B1) is operationalized in (B2) and reduced to a computable form (B3).

3. FIVE CHALLENGES FOR SIMULATION

Challenges for simulation arise from the cognitive nature of IIR and the concept of interaction and affect both modeling and the abstraction of rules for simulated user actions in a given situation. Five specific issues can be identified:

(1) IIR is obviously highly contextual and conditioned on multiple levels of cognitive processes ranging from immediate processing (200ms) to bounded rational processes over minutes to years [1]. Perfect user simulation rests on solving the basic

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.
.

problem of modeling the architecture of cognitive systems, a leading challenge for machine learning [4]. Of course, in any rational program of simulation, perfect user simulation is not necessary. The degree to which a simulation must embed aspects of the cognitive model to reach acceptable fidelity with decision making by real users can only be determined in user studies.

(2) From a system perspective, the issue is not simply to simulate users. Rather, one must simulate users and *uses of the system,* e.g. tasks, because an optimal system must be able to recognize the results of that joint construct[2, 5]. This raises the possibility of a combinatorial explosion for real simulations and questions about the computable characteristics of the problem. The degree to which user and task representations can be handled independently is an active area of investigation for information science, cf [6]. It is hard to see how these issues can be addressed without designed user studies.

(3) A formal problem with practical difficulties comes from the nature of interaction as two spaces with different properties – one for users and one for systems [5]. The user space from action to action is a variable dimension space: the selection options and choices can change at each step. In contrast, the system is a Turing machine and has a fixed dimension space. This modeling space mismatch vastly complicates the problem of assigning the input-output pairs (B2) to a simulated user in the machine's system space, especially when variable length behavior sequences are considered.

(4) Another formal problem for interaction is use of similarity to compare mathematical representations. Interaction means each observation can be conditional on previous actions and the situation. The assumption observations are iid (independent and identically distributed) is essential for defining similarity measures, so use of similarity measures is only approximately valid at best. The relative validity of deriving simulations using similarity calculations during modeling can only rest on empirical evidence, i.e. user studies.

(5) We are still discovering the essential functions and dependencies of human information behavior as illustrated, for example, by point (2) above. Further, it is not clear what observational units are appropriate: user actions? action sequences? goal-segmented sequences? These basic questions require study of real users.

Presenting these issues is not to claim effective user simulation is impossible – indeed one of the practical goals of Human-Computer Information Retrieval (HCIR) is to model users in situations of task, etc. in order to design better information systems. Rather, it is to say that in an operationalist approach the simulation of users cannot avoid studying real users and learning how to build useful models from that data. The absence of a free lunch, i.e. simulating synthetic users, is explained by the very complexity of human information behavior in interactions with systems. From the perspective of the process of creating simulated users, these challenges spring from several foundational sources. Gaining ground truth is necessary to fix both essential properties of the simulations and model parameters for approximations that can achieve acceptable fidelity for the purposes of IIR system evaluation.

4. IIR SYSTEM EVALUATION

What do we want to achieve in simulating a user? The goal for a simulated user project is to build objective system evaluation frameworks that allow for evaluation of individual IIR systems and comparison across systems. Can an evaluation framework capable of saying $Performance(SystemA) > Performance(SystemB)$ be justified without being able to explain qua the evaluation framework *why* System A is better than System B for users?

Summative evaluation is inadequate and this is one shortcoming of the current TREC paradigm. Useful system evaluation must be analytic if it is to connect with real users engaged in tasks that differ from those used to evaluate the systems. This explanatory challenge is the same as Boring's (B5) and is answered in the operationalist approach because the simulation achieved in (B3) depends on the grounding of input-output pairs in (B2) based on observations expressing real user functionalities in IIR (B1).

5. CONCLUSIONS

In view of the scope and nature of these challenges to simulation, it seems best to proceed incrementally in two steps to make IIR user simulation more realistic and less "automagic". That is, one should work from modeling real users performing in real task settings and then abstract learned user and task features into computable models. The first step is demanding and expensive, but necessary. To start with a focus on user system interaction arbitrarily restricts the discovery of rules and relationships between the user and the system to the system components and their interface. The result is to essentially restrict the simulation of users to a class of machine and so deny a reasonable sense of 'real' in simulating real users.

Whatever route exists to simulation of real users must be based on models of real users. Shortcuts that leave users out of the picture run the substantial risk that simulated users are mere reflections of system interface interactions with the essence of interaction abstracted away. Investing in the production of such simulacra is of dubious value.

6. ACKNOWLEDGMENTS

This work was supported by IMLS grant LG-06-07-0105-07.

7. REFERENCES

[1] ANDERSON, J. R. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science 26*, 1 (2002), 85–112.

[2] BELKIN, N. J. Some(what) grand challenges for information retrieval. *SIGIR Forum 42*, 1 (2008), 47–54.

[3] BORING, E. G. Mind and mechanism. *The American Journal of Psychology 59*, 2 (1946), 173–192.

[4] DIETTERICH, T., DOMINGOS, P., GETOOR, L., MUGGLETON, S., AND TADEPALLI, P. Structured machine learning: the next ten years. *Machine Learning 73*, 1 (2008), 3–23.

[5] FUHR, N. A probability ranking principle for interactive information retrieval. *Information Retrieval 12* (2008). http://dx.doi.org/10.1007/s10791-008-9045-0.

[6] LIU, J., COLE, M. J., LIU, C., BELKIN, N. J., ZHANG, J., GWIZDKA, J., BIERIG, R., AND ZHANG, X. Search behaviors in different task types. In *Proceedings of JCDL 2010* (Gold Coast, Australia, June 2010), ACM.

Bridging Evaluations: Inspiration from Dialogue System Research

Tsuneaki Kato
The University of Tokyo
3-8-1 Komaba Meguro-ku
Tokyo, Japan
kato@boz.c.u-tokyo.ac.jp

Mitsunori Matsushita
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka, Japan
mat@res.kutc.kansai-u.ac.jp

Noriko Kando
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
kando@nii.ac.jp

ABSTRACT

This paper discusses methodologies to quantitatively evaluate exploratory information access environments exploiting interactive information visualization. A novel evaluation framework for such a highly interactive environment is proposed, which is inspired by two concepts adopted in research into spoken language dialogue systems: the PARADISE framework and the use of a simulated user. It aims to bridge two types of evaluation and to use the results of the more cost-effective one to predict the other through a version of simulation. A course of studies is also discussed for making this framework feasible, which includes empirical user studies in sophisticated settings.

Categories and Subject Descriptors:
H.3.m [**Information Systems**]: Miscellaneous
General Terms: Measurement
Keywords: Exploratory Search, Visualization, Evaluation

1. INTRODUCTION

The authors are studying methodologies to quantitatively evaluate exploratory information access environments exploiting interactive information visualization. Using interactive information visualization undoubtedly achieves richer environments for exploratory information access through providing the users with representations of informative information and an intuitive means for handling it. It, however, simultaneously makes its evaluation much harder because both information provided by the system and interactive moves made by a user become more diverse and complex. This paper discusses a novel framework for evaluating such a highly interactive environment. It aims to bridge two types of evaluation and to use the results of the more cost-effective one to predict the other through a version of simulation. A course of studies is also discussed for making this framework feasible, which includes empirical user studies in sophisticated settings.

2. BASIC IDEA

There are two directions in evaluating interactive systems. One is empirical user studies in which subjects are requested to accomplish a given task in a controlled situation, and through observing the process, systems are evaluated and the degree of the achievement quantified. When the task is adequately designed, it is very helpful to obtain data in a real world situation, but takes significant amounts of time and resources, especially to compare different systems. The TREC interactive tracks [2] are a representative of this direction. The other is benchmark tests, in which components of systems and their specific functions are evaluated. It is relatively cost effective, but to be convincing, needs to show the results properly reflect the system's utility or quality in a real setting.

A framework is proposed to bridge these two evaluation methods and to predict the results of empirical user studies using those of benchmark tests. It is inspired by the PARADISE framework, which is motivated by similar concerns in research into spoken language dialogue systems [7]. By developing this idea accompanied with using a simulated user, another idea that also came from the same research field, the framework is expected to be applicable to an interactive information access environment.

In the PARADISE framework, the system's primary objective is set to maximize user satisfaction, which is users' subjective rating for controlled experiments. Both task success and costs associated with the interaction are considered to contribute to user satisfaction. Dialogue efficiency and quality are, in turn, considered to contribute to dialogue costs. The PARADISE framework posits that a performance function can then be derived by applying multivariate liner regression with user satisfaction as the dependent variable and task success, dialogue quality and dialogue efficiency measures as independent variables. Dialogue efficiency and quality are represented in several metrics such as elapsed time and mean recognition rate, which can be obtained via controlled experiments by analyzing logged data automatically or by hand. Modeling user satisfaction in such a way leads to predictive performance models of dialogue systems, through which user satisfaction could be predicted on the basis of a number of simpler metrics that can be measured from the logged data. It saves the need for extensive experiments with users to assess their satisfaction.

The PARADISE framework is insufficient to our objective because some metrics still need to be obtained through empirical user studies. It, however, could be developed

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

so as to enable it to construct predictive models of genuine quality for interactive systems, which are usually obtained through empirical user studies, by using data obtained through benchmark tests. In those models, if successfully obtained, the relationship between the dependent and independent variables is too complex to be expressed as a weighted liner combination, however.

To identify this relationship, another notion from the same research field may help. This is using a simulated user, which is used in reinforcement learning of dialogue systems for optimizing dialogue strategy [4]. The simulated user is a stochastic generative model that produces user utterances as a response to a dialogue action taken by the system. The parameters of the simulated user should be estimated from an annotated dialogue corpus. Once a simulated user is available, it can be used in a generative mode to interact with the dialogue system. A two-stage approach is expected to work well. At the first stage, simulation using a simulated user predicts various metrics of task success and dialogue costs, which are measured from the system logs in the PARADISE framework, and then, at the second stage, those metrics are combined and predict the system's primary objective, such as user satisfaction.

3. A COURSE OF STUDIES

Several problems must be overcome to implement this idea to evaluate exploratory information access environments. For example, we have still not agreed on the metric of task success. As pointed out by [2], in interactive environments precision and recall are not competitive enough. In that paper, aspect/instance recall achieved in a given time is thought to be representative of a realistic interactive retrieval task, but this claim has not, and so should be, verified. The most serious problem is that we are not sharing a taxonomy or ontology of either interaction moves of systems and users or evaluation metrics of several aspects of systems, which are related to those moves. These are indispensable for designing metrics of dialogue costs and a simulated user. Various kinds of moves users can make in an interactive information access environment were examined by [1]. That range drastically expands when exploiting interactive information visualization, and their taxonomy/ontology becomes complex, intertwining several factors. Although some candidates have been proposed by [6] and [8], for example, the efforts have to be continued to elaborate them in order to construct those that can be shared in the community.

Thus, empirical user studies, which are a preparatory to make feasible simulation-based interactive exploratory information access evaluation, should be continued for a while longer. They are needed, first, to design and investigate a taxonomy/ontology of moves in exploratory information access environments and their measures, and then to establish a methodology for simulated user configuration and construct predictive models of genuine quality of interactive environments using data obtained through benchmark tests.

Two possible task settings are being designed and elaborated for those studies. The first one is an interactive version of complex question answering (CQA). In CQA, users gather information on events and their interests by asking questions in natural language. Interactive CQA is to add a process of elaboration and refinement to such a one-shot question answering though interaction with visual or textual information representation. The objective is to gather as many relevant nuggets as possible, which is similar to achieving high aspect/instance recall. In addition, in the case of event-list questions, for example, each nugget consists of an event description, which has several facets suitable to visualize, such as the time and place it occurs. Test sets constructed in ACLIA task in NTCIR-7 and 8 [5], the organization of which the authors were engaged in, must be used for that purpose.

The second is summarizing trend information, in which users are requested to summarize the trend of given time series information in a given period, such as the changes in cabinet approval ratings over the past year. This is an example of creative, intelligent works interacting with information access. Summarization has a method for automatic evaluation, and it allows the measurement of its task success to be tractable. Moreover, the materials of summarization range from textual information to numerical information, which can be visualized as charts, and recognizing interrelations among them is important for summarization. These characteristics promote rich interaction across media. The authors have organized the MuST workshop, which deals with several aspects of summarizing trend information, and constructed research resources for this topic [3]. These are significant advantages in pursuing this task.

4. CONCLUSION

A novel evaluation framework for interactive information environments was proposed, which is inspired by two concepts adopted in research into spoken language dialogue systems: the PARADISE framework and the use of a simulated user. The framework will enable genuine quality of systems in the real setting to be predicted using data of benchmark tests. This framework is worth investigating, though it still has a lot of problems and its feasibility is not clear. As a preparatory to make this framework feasible, empirical user studies should be continued for a while longer. Two tasks were proposed for those empirical user studies, which are expected to derive fruitful outcomes for this purpose.

5. REFERENCES

[1] M.J. Bates: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. In *Online Information Review*, Vol. 13, No. 5, pp. 407-424, 1989.

[2] S.T. Dumais and N.J. Belkin: The TREC Interactive Tracks: Putting the User into Search. In E.M. Voorhees and D.K. Harman ed. *TREC Experiment and Evaluation in Information Retrieval*. pp. 123-152. The MIT Press, 2005.

[3] T. Kato and M. Matsushita: Overview of MuST at the NTCIR-7 Workshop – Challenges to Multi-modal Summarization for Trend Information. In *Procs. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 475-488, 2008.

[4] E. Levin, R. Pieraccini, and W. Eckert: A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. In *IEEE Trans. of Speech and Audio Processing*, Vol. 8, No. 1, pp. 11-23, 2000.

[5] T. Mitamura, E. Nyberg, et al.: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Procs. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 21-25, 2008.

[6] B. Shneiderman: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. In *Procs. of the 1996 IEEE Symposium on Visual Languages*, pp.336-343, 1996.

[7] M.A. Walker, D.J. Litman, et al.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Procs. of ACL/EACL 97*, pp. 271-280, 1997.

[8] J. Zhang: *Visualization for Information Retrieval*. Springer, 2008.

A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction

Catherine Mulwa[1], Wei Li[2], Séamus Lawless[1], Gareth J. F. Jones[2]

Centre for Next Generation Localisation
School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland[1]
School of Computing, Dublin City University, Dublin, Ireland[2]

mulwac@scss.tcd.ie, wli@computing.dcu.ie, seamus.lawless@scss.tcd.ie, gjones@computing.dcu.ie

ABSTRACT

The Centre for Next Generation Localisation (CNGL) is involved in building interactive adaptive systems which combine Information Retrieval (IR), Adaptive Hypermedia (AH) and adaptive web techniques and technologies. The complex functionality of these systems coupled with the variety of potential users means that the experiments necessary to evaluate such systems are difficult to plan, implement and execute. This evaluation requires both component-level scientific evaluation and user-based evaluation. Automated replication of experiments and simulation of user interaction would be hugely beneficial in the evaluation of adaptive information retrieval systems (AIRS). This paper proposes a methodology for the evaluation of AIRS which leverages simulated interaction. The hybrid approach detailed combines: (i) user-centred methods for simulating interaction and personalisation; (ii) evaluation metrics that combine Human Computer Interaction (HCI), AH and IR techniques; and (iii) the use of qualitative and quantitative evaluations. The benefits and limitations of evaluations based on user simulations are also discussed.

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

Relevance Feedback, Simulation

1. INTRODUCTION

The Centre for Next Generation Localisation (CNGL) is developing novel technologies which address the key challenges in localisation. Localisation refers to the process of adapting digital content to culture, locale and linguistic environments at high quality and speed. The technologies being developed combine techniques from natural language processing, information retrieval and Adaptive Hypermedia. The complex functionality offered by these systems and the variety of users who interact with them, mean that evaluation can be extremely difficult to plan, implement and execute. Both component-level scientific evaluation and extensive user-based evaluation are required to comprehensively assess the performance of an application. It is critically important that such experiments are thoroughly planned and conducted to ensure the quality of application produced. The potential number of experiments needed to gain a full understanding of the systems being developed means that carrying out these repeated investigations using real interactive user studies is impractical. As a result, simulated interaction is vital to enable these experiments to be replicated and recursively executed in a controlled manner.

SimInt'10, at SIGIR 2010, July 19–23, 2010, Geneva, Switzerland.

2. EVALUATION USING SIMULATED INTERACTION

IR evaluation experiments can be divided into four classes: i) observing users in real situations, ii) observing users performing simulated tasks, iii) performing simulations in the laboratory without users and iv) traditional laboratory research (no users and no interaction simulation) [1]. When simulating user interaction and replicating experiments it is essential that performance is measured using the most suitable evaluation metrics. The following sections detail metrics which can be used to evaluate AIRS, particularly experiments which use simulated interaction.

2.1 IR Evaluation Metrics

IR is classically evaluated in terms of precision and recall, which tell us about the accuracy and scope of the retrieval of relevant documents. These metrics are, of course, very valuable in measuring the effectiveness of real world search tasks. They are also used to evaluate retrieval effectiveness with test collections in laboratory IR experimental settings. However, the standard assumption, in laboratory IR experiments, that the relevance of individual documents is constant for multiple search interactions limits the suitability of such test collections for the evaluation of simulated interactive search.

An experimental framework is needed to capture simulated explicit or implicit feedback from a user and exploit this for relevance feedback and subsequent experiments. This framework could also potentially modify the identified set of relevant documents to reflect: (i) relevant information found in previous iterations of the experiment; and (ii) the development of the user's information need. For example, documents may become relevant as the search progresses and the user's knowledge of a subject grows having seen previous relevant documents. This concept of a user interacting with an IR system and providing feedback which modifies the systems response has similarities with AH systems, from which we next consider relevant evaluation principles.

2.2 AH Evaluation Metrics

Numerous measures of the performance of adaptivity in adaptive systems have been proposed [2]. These metrics aim to address

both component-level scientific evaluation and user-based evaluation of the adaptivity offered by the system.

Personalised Metrics: Personalisation in IR can be achieved using a range of contextual information such as information about the user, the task being conducted and the device being used. Contextual information is increasingly being used to facilitate personalisation in IR. The personalised identification, retrieval and presentation of resources can provide the user with a tailored information seeking experience [2]. Personalisation metrics aim to express the effort necessary to exploit a system [3] e.g. **MpAC:** Minimum personalisation Adaptive Cost which indicates the percentage of entities which are personalised in an AIRS system. This metric considers only the minimum number of entities necessary to make a system adaptive.

Interaction Metrics: These metrics aim to provide information on the quality of the AIRS system's functionality. This is achieved by evaluating the variation in the interaction between administrators or users and the adaptive and non-adaptive versions of a system [4]. Examples include: i) **AiAI:** Administrator Interaction Adaptivity Index. This metric compares the actions performed by administrator to manage the system before and after the addition of adaptivity; ii) **UiAI**: User interaction Adaptivity Index. This metric compares the actions performed by a user to access the functionality of a system both before and after the addition of adaptivity. Whenever an action differs, an additional action is needed or an action is missing, this index increases by one. Interaction metrics assist in the comparative evaluation of AIRS systems from an adaptive perspective.

Performance metrics: Many metrics can be used to measure performance e.g., knowledge gain (AEHS), amount of requested materials, duration of interaction, number of navigation steps, task success, usability (effectiveness, efficiency and user satisfaction). Such metrics concern aspects of the system related to response time, improvement of response quality through adaptivity and the influence of performance factors on the adaptive strategies.

2.3 Simulation of Interaction Techniques

Simulation techniques enable multiple changes of system configuration, running of extensive experiments and analysing results. The simulation assumes the role of a searcher, browsing the results of an initial retrieval [5]. The content of the top-ranked documents in the first retrieved document set constitutes the information space that the searcher explores. All the interaction in this simulation is with this set and it is assumed that searchers will only mark relevant information via the interaction. The authors are interested in the use of this technique to determine how to evaluate the change in retrieval effectiveness when an AIRS system adapts to a query in a standard way, and to incorporate user and domain models and investigate how to exploit these.

2.4 Simulation-Based Evaluation Challenges

The main challenges in the use of simulation methods include: i) determining what data must be gathered in order to replicate experiments; ii) deciding how to gather this data; iii) identifying how to replicate the variety of user behaviours and personalisation offered by the system; iv) the simulation of relevance, for instance simulating the characteristics of relevant documents successfully over a search session; v) validating the simulation's query evaluation times against the actual implementation; vi) selecting what method to use to collect implicit feedback; and vii) deciding how to filter the collected implicit feedback.

3. PROPOSED METHODOLOGY

It is essential that the correct methods are used when evaluating AIRS systems [6]. In order to sufficiently evaluate both the adaptive functionality and the retrieval performance of these systems a hybrid approach is proposed which combines IR, AH and Simulation-based evaluation methods. The techniques and metrics required are: i) **simulation-based techniques** where simulation assumes the role of a searcher, browsing the results of an initial retrieval; ii) **user-centred methods** for simulating interaction and personalisation; and iii) evaluation metrics borrowed from **AH** and **IR**. During a search the information state and need of the user changes and this must be modelled in each simulation so that the information viewed so far by the user can be used to influence the generation of a subsequent query. An objective of AIRS is to minimise the amount of information that must be viewed in order to gain a certain amount of knowledge. Thus the user must be shown relevant information in correct order. This is related to both IR and AH, where personalised responses are created for a domain-specific information need. Thus, for an information need, it is necessary to assess not only the relevance of documents to a topic, but also the order in which these should be presented. The number of documents which must be viewed over a search session to satisfy the information need can be further measured. At each point, search effectiveness can be measured with respect to the current information state of the simulated user. One of the main objectives of this work is to explore the potential of using user and domain models to reduce the user search effort. The potential benefits of the proposed methodology include: retrieval accuracy, completeness of system functionality, cost saving, user satisfaction, adaptivity, time, satisfied customer goal, user ratings, quality, appropriateness, accessibility, assistance, richness, availability, completeness, self-evidence, usability, user-retention, consistency, functionality, performance, predictability, portability, reliability and reuse.

4. CONCLUSION AND FUTURE WORK

Simulation-driven evaluation is not new, but the effects of personalisation on creating reproducible, large scale experiments can be addressed by incorporating AH and IR techniques and evaluation metrics. Further work is required in order to test the proposed methodology using systems being developed by CNGL.

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 Localisation (www.cngl.ie).

6. REFERENCES

[1] H. Keskustalo et al. "Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value". Information Retrieval, vol.11, pp.209-228, 2008.

[2] Séamus Lawless, et al., "A Proposal for the Evaluation of Adaptive Personalised Information Retrieval," presented at the CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation, Milton Keynes, UK, 2010.

[3] L. Masciadri and C. Raibulet, "Frameworks for the Development of Adaptive Systems: Evaluation of Their Adaptability Feature Through Software Metrics,"4th International conference SEA 2009, pp. 309-312.

[4] C. Raibulet and L. Masciadri, "Evaluation of Dynamic Adaptivity Through Metrics: an Achievable Target?," 4th International conference SEA 2009.

[5] R. White, et al., "A simulated study of implicit feedback models," Advances in Information Retrieval, pp. 311-326, 2004.

[6] P. Brusilovsky, et al., "Layered evaluation of adaptive learning systems," International Journal of Continuing Engineering Education and Life Long Learning, vol. 14, pp. 402-421, 2004.

Using QPP to Simulate Query Refinement

[Position Paper]

Daniel Tunkelang
Google
New York, United States
dt@google.com

ABSTRACT

Standard test collections evaluate information retrieval systems through batch analysis of query-response pairs. Such analysis neglects the richness of user interaction in interfaces. Many interactive information retrieval (IIR) researchers favor user studies over test collections, but these studies suffer from high costs and raise concerns about generalizability. In this position paper, we propose using query performance prediction (QPP) to model the fidelity of communication between user and system, thus helping IIR researchers to simulate query refinement with standard test collections.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human information processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*

General Terms

Algorithms, Performance, Experimentation, Human Factors

Keywords

interactive information retrieval, models, evaluation

1. THE CRANFIELD PARADIGM AND IIR

The Cranfield paradigm for information retrieval system evaluation fixes the corpus and information needs, and assumes that relevance judgments are user-independent [12]. Its advocates argue that this paradigm offers broad utility and strong experimental power at low cost [13].

Critics (e.g., [1]) object that the Cranfield paradigm is unable to accommodate the study of people in interaction with information systems. These critics generally favor user studies as an evaluation methodology.

The Cranfield model is most suitable for evaluating query-response systems. Such systems are designed to retrieve relevant results in a single query without access to further user context. For this interaction model, the user-agnostic assumptions of the Cranfield model at least seem plausible.

Interactive information retrieval (IIR) research emphasizes user responsibility and control over the information seeking process [9]. IIR techniques include query and term suggestions [7], result clustering [6], and faceted search [11].

Unfortunately, the Cranfield paradigm has not adapted well to IIR. The effectiveness of systems that go beyond ranked result retrieval cannot be neatly summarized in terms of relevance judgments. As a result, IIR researchers depend predominantly on user studies for evaluation. As noted above, these studies suffer from high costs and raise concerns about generalizability.

2. FIDELITY OF COMMUNICATION

In the query-response model, the primary goal is that the system return relevant results. In contrast, IIR views information seeking as a sequence of interactions between the user and the system. From an IIR perspective, the effectiveness of an information-seeking support system depends on the fidelity of the communication channel through which the user and the system interact.

Because the query-response model has only one round of user-system interaction, its fidelity of communication can be summarized by measuring the relevance of results. In an IIR system, however, communication is iterative and bidirectional. While each round of interaction offers an opportunity to advance the user and system's shared state of knowledge, it also creates a risk that the user and system will diverge because of a breakdown in communication.

How do we measure fidelity of communication in an IIR system? In work on evaluating relevance feedback based on simulated users [8, 14], researchers assume that users are perfect judges of relevance, and thus that we can simulate relevance feedback behavior based on the relevance judgments in a test collection. This assumption is already a rough approximation to reality, but it breaks down entirely when we expect users to apply such perfect judgment to suggested richer response elements, such as query suggestions and result clusters. Instead, we need a way to model and measure the likelihood that a user will correctly assess the utility of a query refinement to his or her information need.

3. QUERY PERFORMANCE PREDICTION

In order to evaluate a system that offers users query refinement options, we need to address these two questions:

1) Do the query refinement options enable the user to make progress in the information seeking task?

2) Can the user interpret the query refinement options and predict the relevance of the results to which they lead?

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

Given a standard test collection with relevance judgments, we can model an answer to the first question in terms of those judgments. For example, we can simulate an opportunistic user who, whenever presented with refinement options, selects the option that immediately leads to to the most relevant result set (e.g., one that optimizes for average precision). We can also simulate a more far-sighted user who explores a set of paths using iterative query refinements and ultimately selects the path that lead to the best destination. Such an approach is similar to the expected benefit framework proposed by Fuhr [4].

For the second question, we turn to query performance prediction (QPP), also known as query difficulty estimation [2]. QPP estimates the likelihood that a system will succeed in addressing a user's information need for a particular query–without access to a set of relevance judgments. Pre-retrieval measures, such as query coherency [5], estimate performance based solely on analzying the query; while post-retrieval measures, such as query clarity[3] and robustness[15], also analyze the results.

While QPP has been used primarily to identify human-generated queries that pose difficulty to information retrieval systems, we believe it can be applied more generally to evaluate or guide query refinements generated by IIR systems. Even if a query refinement leads to results that advance the user's progress towards fulfilling his or her information need, QPP estimates the likelihood that the user will pick up the requisite information scent [10] to follow it.

Now we have all of the ingredients we need: a test collection with relevance judgements, a standard relevance measure such as average precision, and a QPP measure such as query clarity. Combining these, we can generalize Fuhr's expected-benefit model to simulate how a user interacts with a system that offers query refinements.

For each interaction, we combine (e.g., multiply) the relevance and QPP measures to obtain a score for each refinement. How we apply these scores depends on the information-seeking strategy we want to simulate. For example, we can simulate an opportunistic user by always following the highest-scoring refinement. We can then judge the system's effectiveness of the system based on the relevance of the final result set and the system's efficiency based on the length of the simulated session. This approach generalizes to simulating other information-seeking strategies.

This approach relies on a variety of assumptions–in particular, that QPP works for query refinements generated by IIR systems. Given that query refinements are often user-generated queries mined from logs, this assumption seems reasonable. Nonetheless, it requires experimental validation. Also, not all query refinement approaches are amenable to all QPP measures. In particular, pre-retrieval methods require that the refinements by expressible as explicit queries. Nonetheless, we feel this approach has broad applicability.

4. SUMMARY

We have proposed an approach that applies QPP to IIR in order to use standard test collections for evaluating IIR systems. While this proposal is barely a sketch, we hope it suggests a productive research direction for the IIR community. We believe that bridging the gap between the Cranfield model and IIR hinges on modeling the fidelity of communication between the user and the system, and we hope that future research will validate the potential of this approach.

5. REFERENCES

[1] N. Belkin. Some (what) grand challenges for information retrieval. *ACM SIGIR Forum*, 42:47–54, 2008.

[2] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool, 2010.

[3] S. Cronen-Townsend, Y. Zhou, and W. Croft. Predicting query performance. In *Proc of 25th ACM SIGIR*, 299–306, 2002.

[4] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[5] J. He, M. Larson, and M. De Rijke Using coherence-based measures to predict query difficulty. In *Proc of 30th ECIR*, 689–694, 2008.

[6] M. Hearst. Clustering versus faceted categories for information exploration. *CACM*, 49(4):61, 2006.

[7] D. Kelly, K. Gyllstrom, and E. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proc of 32nd ACM SIGIR*, 371–378, 2009.

[8] H. Keskustalo, K. Järvelin, and A. Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228, 2008.

[9] G. Marchionini. Toward human-computer information retrieval. *ASIST Bulletin*, 32(5):20, 2006.

[10] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[11] D. Tunkelang. *Faceted search*. Morgan & Claypool, 2009.

[12] E. Voorhees. The philosophy of information retrieval evaluation. In *2nd CLEF Workshop on Evaluation of CLIR Systems*, 143–170, 2002.

[13] E. Voorhees. On test collections for adaptive information retrieval. *Information Processing & Management*, 44(6):1879–1885, 2008.

[14] R. White, J. Jose, C.J. Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *Proc of 26th ECIR*, 311–326, 2004.

[15] Y. Zhou and W. Croft. Ranking robustness: a novel framework to predict query performance. In *Proc of 15th ACM CIKM*, 574, 2006.

Focused Relevance Feedback Evaluation

Shlomo Geva
Computer Science, QUT
2 George St
Brisbane Q4001 Australia
+617 3138 1920
s.geva@qut.edu.au

Timothy Chappell
Computer Science, QUT
2 George St
Brisbane Q4001 Australia
+617 3138 1920
timothy.chappell@qut.edu.au

ABSTRACT
This paper describes a refined approach to the evaluation of Focused Relevance Feedback algorithms through simulated exhaustive user feedback. As in traditional approaches we simulate a user-in-the loop by re-using the assessments of ad-hoc retrieval obtained from real users who assess *focused* ad-hoc retrieval submissions. The evaluation is extended in several ways: the use of *exhaustive* relevance feedback over entire runs; the evaluation of *focused retrieval* where both the retrieval results and the feedback are *focused*; the evaluation is performed over a closed set of documents and complete focused assessments; the evaluation is performed over executable implementations of relevance feedback algorithms; and finally, the entire evaluation platform is reusable. We present the evaluation methodology, its implementation, and experimental results that demonstrate its utility.

Categories and Subject Descriptors
Focused Relevance Feedback, Relevance Feedback, Information Retrieval, IR, RF, User Simulation, Search Engine, Evaluation, INEX http://www.inex.otago.ac.nz/

General Terms
Algorithms, Measurement, Performance, Benchmark.

Keywords
Relevance feedback evaluation.

1. INTRODUCTION
Information retrieval systems are most effective when used by skilled operators who are capable of forming queries appropriate for retrieving relevant documents. The vast majority of users of information retrieval systems are unlikely to be skilled users. It is a trivial observation that user will sooner reformulate a query than they would scan the initial result list to any depth beyond the first page of results. As query reformulation may be a difficult and time-consuming task, machine-assisted query reformulation based on the requirements of the user is an important part of information retrieval. An early and rather effective mechanism for improving the effectiveness of search interfaces is known as *relevance feedback* where by query reformulation is automated. We are concerned with the *evaluation* of this approach. This paper describes an extension of the *Incremental Relevance Feedback* approach, described by IJsbrand Jan Aalbersberg[1], to Focused Relevance Feedback and to the evaluation of executable implementations under uniform setting.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

1.1 Relevance Feedback Evaluation
This wealth of research and reported results on relevance feedback leads to an obvious problem – most of the earlier work is difficult to reproduce reliably, and certainly not without great difficulty in implementation of systems described by others. Of great importance to the task of comparing different methods of ranking and retrieval is having a standard, systematic way of evaluating the results so that it can be empirically validated, in a methodologically sound manner, that for a given test collection one particular method is better than another.

Ruthven and Lalmas[2] review alternate evaluation methods suited to relevance feedback: *Freezing, Residual ranking*, and *Test and control,* all intended to counter the effect where the documents marked as 'relevant' by the user are pushed to the top of the document ranking, artificially raising the mean precision of the results. *Freezing* is where the initially top-ranked documents are frozen in place and the relevance feedback system used to re-rank the remaining documents, and the precision/recall evaluation conducted on the entire document set. *Residual ranking* is where the top-ranked documents, used to train the relevance feedback system, are removed from the document set before evaluation. *Test and control groups*; where the document set is partitioned into two equal groups, the first used to train the relevance feedback system and the second used to evaluate the system. Aalbersberg[1] describes an incremental approach which we extend in this paper, where one document at a time is evaluated by the user until a given depth of results list is inspected.

1.2 Focused Relevance Feedback evaluation
In this paper we describe a refined approach to the evaluation of Relevance Feedback algorithms through simulated *exhaustive* incremental user feedback. The approach extends evaluation in several ways, relative to traditional evaluation. First, it facilitates the evaluation of retrieval where both the retrieval results and the feedback are ***focused***. This means that both the search results and the feedback are specified as passages, or as XML elements, in documents - rather than as whole documents. Second, the evaluation is performed over a closed set of documents and assessments, and hence the evaluation is exhaustive, reliable and less dependent on the specific search engine in use. By reusing the relatively small topic assessment pools, having only several hundred documents per topic, the search engine quality can largely be taken out of the equation. Third, the evaluation is performed over ***executable*** implementations of relevance feedback algorithms rather than being performed over result submissions. Finally, the entire evaluation platform is reusable and over time can be used to measure progress in focused

relevance feedback in an independent, reproducible, verifiable, uniform, and methodologically sound manner.

2. EVALUATION APPROACH

This approach is concerned with the simulation of a user in loop, in the evaluation of relevance feedback systems. This approach can be used to compare systems in an evaluation forum setting, or simply to evaluate improvements of variations to existing relevance feedback algorithms in the development process.

2.1 Use Case

The use-case of this track is similar to Aalbersberg[1] - a single user searching with a particular query in an information retrieval system that supports relevance feedback. Our user views and ***highlights relevant passages*** of text in a returned document (if exist) and provides this feedback to the information retrieval system. The IR system re-ranks the remainder of the unseen results list to provide the next ***assumed*** most relevant result to the user. The exact manner in which this is implemented is not of concern in this evaluation; here we test the ability of the system to use focused relevance feedback to improve the ranking of previously unseen results. Importantly, we extend Aalbersberg's approach to compare the improvement, if exists, which focused relevance feedback (FRF) offers over whole document feedback. This includes structured IR (e.g. XML documents).

2.2 Test Collection

The relevance feedback track will use the INEX Wikipedia XML collection. Evaluation will be based on the focused relevance assessments, which are gathered by the INEX Ad-Hoc track through the GPXrai assessment tool, where assessors highlight relevant passages in documents. The INEX Wikipedia test collection is semantically marked up. This facilitates the evaluation of FRF algorithms implementations, which take advantage not only of the (often) passage-sized feedback, but also the semantic mark-up of the relevant text.

2.3 Task

Participants will create one or more Relevance Feedback Modules (RFMs) intended to rank a collection of documents with a query while incrementally responding to explicit user feedback on the relevance of the results presented to the user. These RFMs will be implemented as dynamically linkable modules that will implement a standard defined interface. The Evaluation Platform (EP) will interact with the RFMs directly, simulating a user search session. The EP will instantiate an RFM object and provide it with a set of XML documents and a query. The RFM will respond by ranking the documents (without feedback) and returning the ranking to the EP. This is so that the difference in quality between the rankings before and after feedback can be compared to determine the extent of the effect the relevance feedback has on the results. The EP will then request the next most relevant document in the collection (that has not yet been presented to the user). On subsequent calls the EP will pass relevance feedback (in the form of passage offsets and lengths) about the last document presented by the RFM. This feedback is taken from the qrels of the respective topic, as provided by the Ad-Hoc track assessors. The simulated user feedback may then be used by the RFM to re-rank the remaining unseen documents and return the next most relevant document. The EP makes repeated calls to the RFM until all relevant documents in the collection have been returned.

The EP will retain the presentation order of documents as generated by the RFM. This order will then be evaluated as a submission to the ad-hoc track in the usual manner and with the standard retrieval evaluation metrics. It is expected that an effective dynamic relevance feedback method will produce a higher score than a static ranking method (i.e. the initial baseline rank ordering). Evaluation will be performed over all topics and systems will be ranked by the averaged performance over the entire set of topics, using standard INEX and TREC metrics.

Each topic consists of a set of documents (the topic pool) and a complete and exhaustive set of manual focused assessments against a query. Hence, we effectively have a "classical" Cranfield experiment over each topic pool as a small collection with complete assessments for a single query. The small collection size allows participants without an efficient implementation of a search engine to handle the task without the complexities of scale that the full collection presents.

As an example, Figure 1 depicts the performance improvement evaluation as obtained by using Rocchio with the Lucene search engine, when evaluated by trec_eval. Rocchio-based relevance feedback engine results in an improved mean average precision. The third line shown (the middle) is the best performing submission at INEX 2008, modified to conform to the trec_eval input format. It performs best out of the three in early precision, but precision suffers later and it has a lower average precision than the Lucene engine when using relevance feedback.

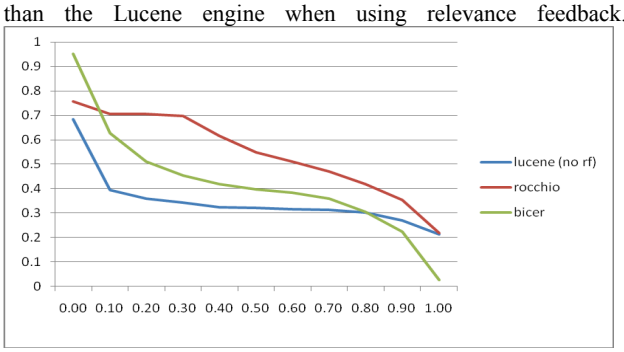


Figure 1. Evaluation with trec_eval, document retrieval.

The approach provides an interactive user session simulation in a focused relevance feedback setting. The evaluation provides a level playing field for the ***independent and completely reproducible*** evaluation of RF implementations in a standard setting and with a standard pool of documents for each topic.
The approach supports the accurate evaluation of any benefits that may (or may not) arise from the use of Focused IR, as opposed to document IR, be it passage based or XML Element based.

3. References

1. IJsbrand Jan Aalbersberg, Incremental Relevance Feedback, Proceedings of SIGIR 1992, pp 11-22.
2. I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.,* 18(2):95-145, 2003

Modeling the Time to Judge Document Relevance

Chandra Prakash Jethani
David R. Cheriton School of Computer Science
University of Waterloo
cpjethan@cs.uwaterloo.ca

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

ABSTRACT

We want to make timed predictions of human performance by modeling user interaction with a hypothetical user interface. Inherent in making a timed prediction via simulation is knowing how long various actions take. One of the most costly actions a user can take is judging a document for relevance. We present a model of the average time to judge the relevance of a document given the document's length. We produce two parameterized versions of the model based on two sets of user data. The models explain 26-45% of the variance in the average time to judge document relevance. Our models should allow for more accurate timed predictions of human performance with interactive retrieval systems.

1. INTRODUCTION

We believe that automated information retrieval (IR) evaluation should provide predictions of human performance. To make these predictions our approach has been to create user models that describe how a user interacts with a given interactive IR system. Each action that the simulated user takes has with it an associated time to complete. While the HCI community has established average times for low level actions such as pointing a computer mouse, the IR community has yet to establish times for high level actions such as the judging of document relevance. We next describe our model to predict the average time to judge a document for relevance.

2. METHODS AND MATERIALS

Our model for judging the relevance of a document is a simple one based solely on a document's length. To judge the relevance of the document, our model's hypothetical user scans the document looking for relevant material. The user examines areas of the page in more detail to make a relevance decision. On making a decision, the user then uses the computer mouse to enter the decision and go on to the next page. As such, the time to judge a document given a document's length in words, is:

$$T(w) = sw + ra + c \quad (1)$$

where s is the scan rate in seconds per word, w is the document length in words, r is the reading rate in seconds per word, a is the total length of the areas of interest, and c represents a constant overhead for judging in seconds. A simplifying assumption we make is to fix the size of a. Thus, we can rewrite the model as:

$$T(w) = sw + k \quad (2)$$

where k is a constant amount of time required for judging any document.

To parameterize and evaluate our model, we draw our data from a larger user study [2]. In this two phase study, 48 users used a different interface for each phase of the study. In the first phase, users judged for relevance a series of document summaries and full documents. In this paper, we only look at the time to judge full documents. The phase 2 interface presented 10 query-biased document summaries (snippets) per page. Clicking on the summary allowed the user to either save the document as relevant or do nothing and use the web browser's back button to go back and view the search result summaries. Each interface displayed full documents in a similar manner with query terms highlighted and instructions and buttons for judging displayed at the top of the page.

For each document, we compute its length by stripping it of any markup and splitting the text into whitespace separated words. Some documents are duplicates of each other. We mapped each set of duplicates to a unique identifier and set the document length equal to the average of the documents in the set.

Users worked at different rates. For this paper, we only use documents judged by 10 or more users. We compute for a document the average amount of time users took to judge its relevance. For documents in phase 2, we counted as a judgment both the saving of a document as relevant and also the choice to not save a viewed document. We used only the first time a user viewed a document. For example, if a user takes t_1 seconds to view a document and does nothing, but later takes t_2 seconds to revisit the document and save it as relevant, we will only use t_1 in our calculations. In particular, for duplicate documents, we only use the time to judge the first copy of the document.

Our user study [2] used 8 topics from the 2005 TREC Robust track, which used the AQUAINT collection of 1,033,461 newswire documents.

3. RESULTS AND DISCUSSION

Figure 1 shows the experimental data for phase 1 and 2 and a linear fit to each set of data. The fit model with standard errors for phase 1 is:

$$T(w) = 0.024 \pm 0.004w + 21 \pm 4 \quad (3)$$

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

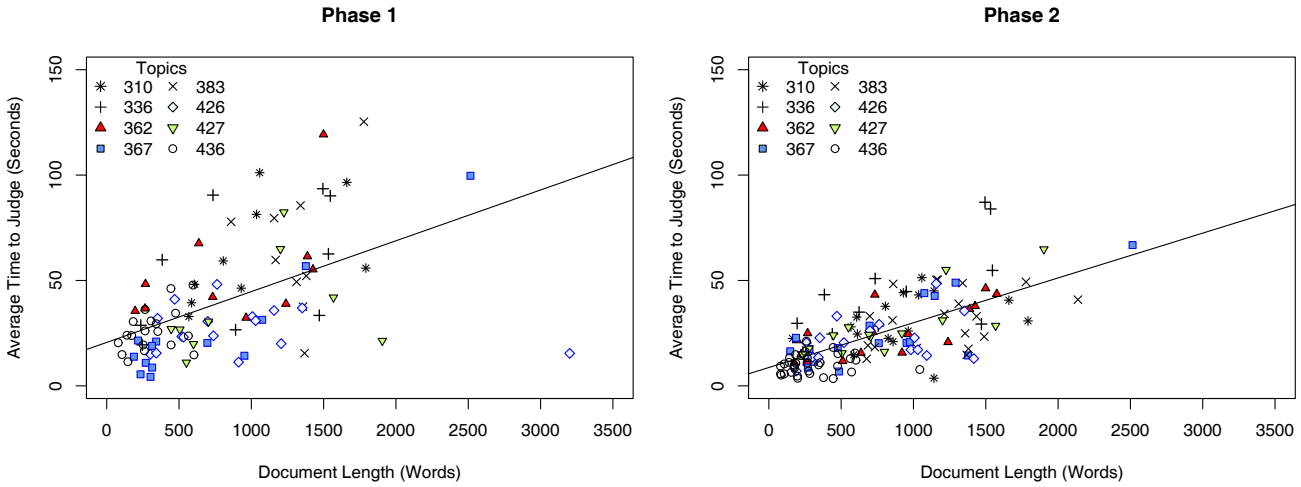


Figure 1: The average time to judge relevance vs. document length. Each point is a document and is the average of 10 or more users' time to judge the document. Each plot represents a different phase of the user study as explained in Section 2.

The adjusted R-squared for the phase 1 model is 0.26, i.e. 26% of the variance in the average time to judge a document in phase 1 is explained by the model. For phase 2, the fitted model is:

$$T(w) = 0.021 \pm 0.002w + 9 \pm 2 \quad (4)$$

and has an adjusted R-squared of 0.45. Interestingly, phase 2 is the more realistic interface for search, and here we are better able model the time to judge relevance.

For each phase, the number of seconds per word is approximately the same. A scan rate of 0.024 seconds per word is equivalent to 2500 words per minute (wpm). The average rate at which people can scan a text looking for a word is 600 wpm [1]. There are several possible reasons for our measured scan rate being more than 4 times the rate of conventional scanning.

Firstly, users are scanning documents with query term highlighting and are not performing a true lexical scan of 600 wpm. We expect users to be able to find areas of interest in documents at a rate faster than possible without highlighting. Secondly, users do not always need to view the full length of the document to make their decision. For example, Figure 1 shows for phase 1 a long document of greater than 3000 words that is judged in an average of 15 seconds. On examination of this document, it is clear that the document is non-relevant given its title and first sentences. Finally, some users appear to read more than others. In other words, users employ different strategies to comprehending documents, and our model does not capture this difference between users.

The constant amount of time required for judging a document varies greatly between the two phases. For phase 1, the constant is 21 seconds and for phase 2 it drops to 9 seconds. These times are similar to the times we measured for users to judge summaries in phase 1 and phase 2. In phase 1, users on average judged a summary in 15.5 seconds. In phase 2, users spent 9.1 seconds viewing the summaries before taking some action. Thus, it seems reasonable to hypothesize that users read in detail about as much material as a query-biased document summary. Our summaries were 2 sentences with a combined maximum length of 50 words.

Why should users take longer to judge documents in phase 1? In phase 1, users are forced to make relevance decisions for a given document. In phase 2, users only explicitly save relevant documents after making a decision to view the full document based on the document summaries.

4. CONCLUSION

We modeled the average time to judge document relevance as a function of document length. Using data from a previously conducted user study, we fit the model to the data produced by the study's two user interfaces. When fit to the interface that required users to make a judgment, the model explains 26% of the variance in the average time to judge. The second interface was similar to today's web search interfaces that display 10 query-biased document summaries and allow the user to click on a summary to view the full document. When fit to the data from this second interface, the model explains 45 percent of the variance. It appears that document length does have a significant influence on the time to judge document relevance.

5. ACKNOWLEDGMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by an Amazon Web Services in Education Research Grant, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] R. P. Carver. Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2):84–95, 1992.

[2] M. D. Smucker and C. Jethani. Human performance and retrieval precision revisited. In *SIGIR'10*. 2010.

Session Track at TREC 2010

Evangelos Kanoulas* Paul Clough* Ben Carterette† Mark Sanderson*

*Department of Information Studies
University of Sheffield
Sheffield, UK

†Department of Computer & Information Sciences
University of Delaware
Newark, DE, USA

ABSTRACT

Research in Information Retrieval has traditionally focused on serving the best results for a single query. In practice however users often enter ill-specified queries which then they reformulate. In this work we propose an initial experiment to evaluate the effectiveness of retrieval systems over single query reformulations. This experiment is the basis of the TREC 2010 Session track.

1. INTRODUCTION

Research in Information Retrieval has traditionally focused on serving the best results for a single query, e.g. the most relevant results, a single most relevant result, or a facet-spanning set of results. In practice, no matter the task, users often enter a sufficiently ill-specified query that one or more reformulations are needed in order to locate a sufficient number of what they seek. Early studies on web search query logs showed that half of all Web users reformulated their initial query: 52% of the users in 1997 Excite data set, 45% of the users in the 2001 Excite dataset [9]. A search engine may be able to better serve a user not by ranking the most relevant results to each query in the sequence, but by ranking results that help "point the way" to what the user is really looking for, or by complementing results from previous queries in the sequence with new results, or in other currently-unanticipated ways.

The standard evaluation paradigm of controlled laboratory experiments is unable to assess the effectiveness of retrieval systems to an actual user experience of querying with reformulations. On the other hand, interactive evaluation is both noisy due to the high degrees of freedom of user interactions, and expensive due to its low reusability and need for many test subjects. In this work we propose an initial experiment that can be used to evaluate the simplest form of user contribution to the retrieval process, a single query reformulation. This experiment is the basis of the TREC 2010 Session track.

2. EVALUATION TASKS

We call a sequence of reformulations in service of satisfying an information need a session, and the goals of our evaluation are: (**G1**) to test whether systems can improve their performance for a given query by using information about a previous query, and (**G2**) to evaluate system performance over an entire query session instead of a single query. We limit the focus of the track to sessions of two queries.This is partly for pragmatic reasons regarding the difficulty of obtaining session data, and partly for reasons of experimental design and analysis: allowing longer sessions introduces many more degrees of freedom, requiring more data from which to base conclusions.

A set of 150 query pairs (original query, query reformulation) is provided to TREC participants. For each such pair the participants are asked to submit three ranked lists of documents for three experimental conditions, (a) one over the original query (**RL1**), (b) one over the query reformulation, ignoring the original query (**RL2**), and (c) one over the query reformulation taking into consideration the original query and its search results (**RL3**). By using the ranked lists (RL2) and (RL3) we evaluate the ability of systems to utilize prior history (G1). By using the returned ranked lists (RL1) and (RL3) we evaluate the quality of ranking function over the entire session (G2).

3. QUERY REFORMULATIONS

There is a large volume of research regarding query reformulations which follows two lines of work: a descriptive line that analyzes query logs and identifies a taxonomy of query reformulations based on certain user actions over the original query (e.g. [6, 1]) and a predictive line that trains different models over query logs to predict good query reformulations (e.g. [4, 3, 8, 5]). Analyses of query logs showed a number of different types of query reformulations with three of them being consistent across different studies (e.g. [4, 6]):
Specifications: the user enters a query, realizes the results are too broad or that they wanted a more detailed level of information, and reformulates a more specific query.
Drifting/Parallel Reformulation: the user entered a query, then reformulated to another query with the same level of specification but moved to a different aspect or facet of their information need.
Generalizations: the user enters a query, realizes that the results are too narrow or that they wanted a wider range of information, and reformulated a more general query.

In the absence of query logs, Dang and Croft [2] simulated query reformulations by using anchor text, which is readily available. In this work we use a different approach. To construct the query pairs (original query, query reformulation) we start with the TREC 2009 Web Track diversity topics. This collection consists of topics that have a "main theme" and a series of "aspects" or "sub-topics". The Web

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

Track queries were sampled from the query log of a commercial search engine and the sub-topics were constructed by a clustering algorithm [7] run over these queries aggregating query reformulations occuring in the same session. We used the aspect and main theme of these collection topics in a variety of combinations to provide a simulation of an initial and second query. An example of part of a 2009 Web track query is shown below.

```
<topic number="4" type="faceted">
  <query>toilet</query>
  <description> Find information on buying, installing,
  and repairing toilets.
  </description>
  <subtopic number="1" type="inf">
    What different kinds of toilets exist, and how do
    they differ?
  </subtopic>
  ...
  <subtopic number="3" type="inf">
    Where can I buy parts for American Standard toilets?
  </subtopic>
  ...
  <subtopic number="6" type="inf">
    I'm looking for a Kohler wall-hung toilet. Where can
    I buy one?
  </subtopic>
</topic>
```

To construct **specification** reformulations we used the Web Track <query> element as the original query, selected a subtopic and considered it as the actual information need. We then manually extracted keywords from the sub-topic and used them as the reformulation. For instance, in the example above we used the query "toilet" as the first query, selected the information need ("I'm looking for a Kohler wall-hung toilet. Where can I buy one?"), extracted the keyword "Kohler wall-hung" and considered that as a reformulation. This query pair simulates a user that is actually looking for a Kohler wall-hung toilet, but poses a more general query first, possibly because they don't "know" what they need.

```
<topic number="1" reformtype="specification">
  <query>toilet</query>
  <reformulation>Kohler wall-hung toilet</reformulation>
  <description>I'm looking for a Kohler wall-hung toilet.
  Where can I buy one?
  </description>
</topic>
```

To construct **drifting** reformulations we selected two subtopics, used the corresponding <subtopic> elements as the description of two separate information needs, extracted keywords out of the subtopic, and used these keywords respectively as the query and query reformulation. For instance, in the example above we selected subtopics 3 and 6 as the two information needs. Then we extracted the keywords "parts American Standard" and "Kohler wall-hung toilet" and used them as the original query and the query reformulation. This pair simulates a user that first wants to buy toilet parts from American Standard and then decides that they also want to purchase Kohler wall-hungs while browsing the results.

```
<topic number="2" reformtype="drifting">
  <query>parts American Standard</query>
  <description>Where can I buy parts for
  American Standard toilets?</description>
  <reformulation>Kohler wall-hung toilet</reformulation>
  <rdescription>I'm looking for a Kohler
  wall-hung toilet. Where can I buy one?
  </rdescription>
</topic>
```

Finally, to construct **generalization** reformulations we followed one of two methods. In the first method we selected one of the subtopics and we extracted as many keywords as possible to construct an over-specified query, e.g. from subtopic 1 of the example topic we may extract the keywords "different kinds of toilets", which seems to be a lexical over-specification. We then used a subset of these keywords to generalize the original query (e.g. "toilet"). This is meant to simulate a user that first wants to find what types of toilets exist, but lexically over-specifies the need; the retrieved results are expected to be poor and therefore the user needs to reformulate.

```
<topic number="3" reformtype="generalization">
  <query>different kinds of toilets</query>
  <reformulation>toilets</reformulation>
  <description> What different kinds of toilets
  exist, and how do they differ?</description>
</topic>
```

For the second method we selected one of the subtopics or the query description from the Web Track topics as the information need, extracted keywords from a different subtopic that seemed related but essentially it was a mis-specification of something very narrow, and extracted keywords from the subtopic used as information need.

```
<topic number="4" reformtype="generalization">
  <query>American Standard toilet</query>
  <reformulation>toilet</reformulation>
  <description>Find information on buying,
  installing, and repairing toilets.</description>
</topic>
```

4. CONCLUSIONS

Simulating a user is a difficult task. A test collection and accompanying evaluation measures already provide a rudimentary simulation of such users. We have chosen to extend this by considering one more aspect of typical searchers, their reformulation of a query.

5. REFERENCES

[1] P. Bruza and S. Dennis. Query reformulation on the internet: Empirical data and the hyperindex search engine. In *Proceedings of RIAO*, pages 488–500, 1997.

[2] V. Dang and B. W. Croft. Query reformulation using anchor text. In *Proceedings of WSDM*, pages 41–50, 2010.

[3] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of CIKM*, pages 77–86, 2009.

[4] B. J. Jansen, D. L. Booth, and A. Spink. Patterns of query reformulation during web searching. *JASIST*, 60(7):1358–1371, 2009.

[5] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of WWW*, 2006.

[6] T. Lau and E. Horvitz. Patterns of search: analyzing and modeling web query refinement. In *Proceedings of UM*, pages 119–128, 1999.

[7] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of WWW*, pages 1171–1172, New York, NY, USA, 2010. ACM.

[8] X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of CIKM*, pages 479–488, 2008.

[9] D. Wolfram, A. Spink, B. J. Jansen, and T. Saracevic. Vox populi: The public searching of the web. *JASIST*, 52(12):1073–1074, 2001.

Graph-Based Query Session Exploration Based on Facet Analysis

Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola
Department of Information Studies and Interactive Media
FI-33014 University of Tampere, FINLAND
heikki.keskustalo@uta.fi, kalervo.jarvelin@uta.fi, ari.pirkola@uta.fi

ABSTRACT
We explain a simulation approach based on topical facet analysis and a graph-based exploration of query sequences in test collections. First, major facets and their logical relationships are identified in the topics and in the corresponding relevant documents. Secondly, expressions (query terms) in relevant documents are collected, classified and annotated by test persons. Third, term combinations (queries) are formed systematically so that one query corresponds to one vertex of a graph G representing a topic. Query formulation strategies manifest as edges in G. Session strategies manifest as paths in G. We close by discussing the significance of this approach for IR evaluation.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Selection process

General Terms
Measurement, Performance, Theory

Keywords
Iterative search process, session-based evaluation, simulation

1. INTRODUCTION
In real life searchers often utilize sequences of short queries (1-3 query terms) based on limited query modifications. The traditional Cranfield-style setting assuming one query per topic does not model such behavior. We suggest modeling search sessions as an iteration of querying and browsing, and studying the limits of short query sessions through a query graph based on query terms having "searcher warrant" and "literary warrant".

We assume the following real life search behavior. The user will issue an initial query and inspect some top-N documents retrieved; if an insufficient number of relevant documents are found, the user will repeatedly launch the next query until the information need is satisfied or (s)he gives up. Due to the costs involved in query formulation, the user may attempt to maximize the total benefits in relation to the costs during session by rapidly trying out short queries.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

Our research question is: How effective are short query sessions if we allow only limited query modifications, limited browsing of the retrieved result, and success is defined as being able to locate one (highly) relevant document. We propose a graph-based approach to study the effectiveness of short query sessions.

2. MULTI-LAYERED GRAPHS
To attain a "searcher warrant" *topics* are first analyzed by test searchers who suggest the major facets and their logical relationships which are justified for being used in searching. Secondly, to guarantee "literary warrant" the facets and their expressions are recognized in the relevant *documents*. This creates a search thesaurus suggesting search facets for each topic and reasonable expressions to use during a topical search session. Finally, the expressions need to be represented as character strings suited to the properties of the particular retrieval system and index type [1]. For each topic we suggest forming two separate layers: a *concept graph C = (F,M)* and *an expression level graph G = (V,E)* (cf. [3]). In *C* the facets and facet combinations constitute the set of vertexes *F* and the *abstract* moves (e.g., "add a facet") reflect as edges *M*. The nodes between the two layers are linked in such a way that corresponding to each node in C there is a flock of expression level nodes (queries) in G which conform to the same logical form as the particular facet node. For example, let us assume that a Boolean structure is used. In that case a facet node "[A] and [B]" is related to expressions such "a1 and b1" and "a1 and b2". This structure can be traversed to systematically experiment the limits of various query modification strategies.

3. EXPLORING THE EXPRESSION LEVEL GRAPH
We next report an simplified experiment related to the idea described above. We asked test persons to suggest search terms for short queries based on test topics of a test collection (41 topics from TREC 7 and TREC 8 ad hoc tracks) [2]. We did not have facet analysis performed for the relevant documents or for the topics. Therefore, having only a limited set of terms available suggested by test persons we simply formed *all 5-word combinations* as query candidates using an implicit *#sum* operator of search engine *Lemur*.

We next performed a search for each query individually. Each query corresponds to one vertex *V* in a topical graph *G = (V, E)* and an effectiveness result can be associated with each vertex.

This graph supports exploring the effectiveness of various paths (topical query sessions) defined by the allowed transitions (i.e., edges, *E*) from vertex to vertex. The effectiveness of short query sessions can be considered in retrospect using such a graph.

The set of 5 query terms *A, B, C, D, E* produced 2^5 query term combinations (32 vertexes) which are arranged into a diamond-shaped figure (Figure 1).

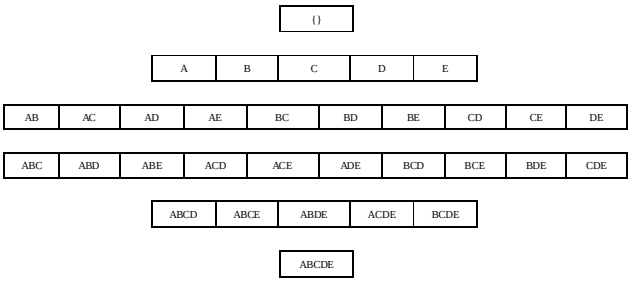


Figure 1. *Query combinations (graph vertexes) arranged by the number of search terms.*

For example, the search terms A-E for topic #351 consist of an ordered set *{petroleum, exploration, south, atlantic, falkland}*. The vertex BC corresponds to the query *#sum(exploration south)*.

4. THE EXPERIMENT

The test collection was organized under the retrieval system *Lemur*, and the existing relevance judgments of the collection were done using a four-point scale. As explained above, the search terms were suggested by test persons (here seven undergraduate information science students) who had no interaction with the search system.

Any desired effectiveness values can be computed for the graph vertexes. Figure 2 reports P@5 results for one topic (#351). Next the numbers in the cells in Figure 2 are interpreted from the point of view of binary success. We define *success* as finding at least one highly relevant document within the top-5, i.e., P@5>0, and failure otherwise. We label the successful vertexes by a plus ('+') sign and the failed vertexes by a minus ('-') sign, thereby the information in Figure 2 can be expressed as follows:

```
#351 ----+ ---+--+--- --+-++-++- -++++ +
```

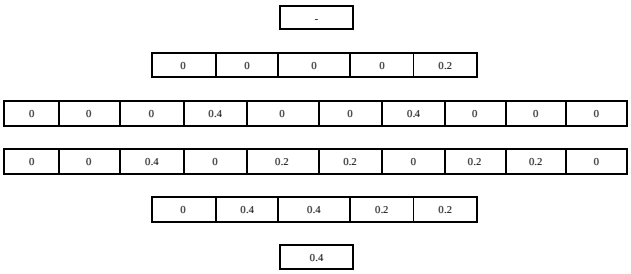


Figure 2. *Effectiveness (P@5) (%) for topic #351 ("petroleum exploration south atlantic falkland") measured by stringent relevance threshold, for each query combination. 14 highly relevant documents exist for the topic. Legend: cells having value above zero indicate success (+) and zeros indicate failure (-) for any particular query combination.*

To make the diagram readable we have arranged it into groups of 5, 10, 10, 5, and 1 symbol, corresponding to query combinations having one, two, three, four, and five query terms. By expressing the topical data this way for *every* topic a visual map is created. It gives information regarding the query combinations available for topical sessions based on a specific success criterion (Table 1).

```
#351 ----+ ---+--+--- --+-++-++- -++++ +
#353 ----- --------+- ----+----+ ---+- -
#355 +++-+ ++++++++++ ++++++++++ +++++ +
#358 ----- -+++---+-- ---+++---+ ---+- +
#360 -+--- +----++--- -++-----+- --+-- -
```

Table 1. *Binary session map for five sample topics and all query combinations. Legend: plus ('+') or minus ('-') symbols correspond to the 31 non-empty vertexes in topical graph, traversed left to right, and rows traversed from top to bottom. Plus indicates success, i.e., P@5 > 0 (here at stringent relevance threshold) and minus indicates a failure (P@5 = 0).*

The binary session map for *all* topics (not shown) can be used to analyze the success of various queries/query types/session types. We observed that the first single-word query ('A') succeeded for 9 topics (the first symbol of the first group) out of 38 (highly relevant documents exist for 38 topics). Assuming that the user started the session this way and in case of failure continued by trying out the second single-word query ('B')(substitution of the term), it succeeded for 6 additional topics (the second symbol of the first group). Assuming, that the user continued instead by *adding* one word ('AB'), session succeeded for 10 additional topics. If the session was started by trying out a two-word query (the first two words given by the simulated users: 'AB') it succeeds for 17 topics out of 38. Three-word query session ('ABC') immediately succeeds for 21 topics.

5. CONCLUSION

In real life sequences of short queries are popular, and the terms are selected from among several alternatives. Such search sessions can be studied in retrospect by considering the query graphs. Considering the systems effectiveness from this point of view may help us towards understanding better the success achievable by various adaptive querying and browsing tactics.

6. REFERENCES

[1] K. Järvelin, J. Kristensen, T. Niemi, E. Sormunen, H. Keskustalo (1996) A deductive data model for query expansion. In SIGIR'96, 235-243.

[2] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, M. Lykke (2009) Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries. In AIRS'09, 63-74.

[3] M. Whittle, B. Eaglestone, N. Ford, V. J. Gillet, A. Madden (2007) Data Mining of Search Engine Logs. JASIST, 58(14), 2382-2400.

A Probabilistic Automaton for the Dynamic Relevance Judgements of Users

Peng Zhang, Ulises Cerviño Beresi
School of Computing
The Robert Gordon University
United Kingdom
p.zhang1, prs.cervino-beresi@rgu.ac.uk

Dawei Song
School of Computing
The Robert Gordon University
United Kingdom
d.song@rgu.ac.uk

Yuexian Hou
School of Computer Sci & Tec
Tianjin University
China
yxhou@tju.edu.cn

ABSTRACT

Conventional information retrieval (IR) evaluation relies on static relevance judgements in test collections. These, however, are insufficient for the evaluation of interactive IR (IIR) systems. When users browse search results, their decisions on whether to keep a document may be influenced by several factors including previously seen documents. This makes user-centred relevance judgements not only dynamic but also dependent on previous judgements. In this paper, we propose to use a probabilistic automaton (PA) to model the dynamics of users' relevance judgements. Based on the initial judgement data that can be collected in a proposed user study, the estimated PA can further simulate more dynamic relevance judgements, which are of potential usefulness for the evaluation of IIR systems.

Category and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Formal Model

Keywords: Interactive IR, Dynamic Relevance Judgement, Probabilistic Automaton, Simulation

1. INTRODUCTION

Relevance judgements in TREC test collections are static. However, in an interactive IR (IIR) environment, when users inspect the results of a search, relevance judgements become dynamic and dependent on each other. Consider the following two scenarios, the first one involving two complementary documents. Each document provides a portion of the solution but in combination they both provide a full solution to a user's problem. Either document in isolation is likely to be judged partially relevant (if not irrelevant), however, when combined, both are likely to be judged relevant. The second scenario is related to the comparison effects between two relevant documents, one being slightly more relevant than the other. Should the less relevant one be encountered first, the user is likely to want to keep it. When faced with the more relevant one, however, this decision may change.

Despite the recognition for the dynamic nature of relevance judgements [2], to the best of our knowledge, little attention has been paid to their formal modelling in terms of the judgement interference, i.e., the interference among relevance judgements for different documents. In this paper, we use the probabilistic automaton (PA) to model the changing process of judgement scores, as the result of the judgement interference. Specifically, the states of the PA can represent users' judgement states, which correspond to judgement scores. Thus the dynamic changes among judgement scores are modelled by the transitions among the PA states. Additionally, each symbol α in the alphabet Σ of the PA denotes (or defines) one type of interference, and the corresponding interference effects are reflected by the PA states transitions, for which the transition probabilities are represented in the transition matrix $M(\alpha)$. Suppose that the judgement for a document d_i is interfered by the judgement for another document d_j. Our method is that after the judgement for d_j, a symbol α will be generated and input to the PA, which triggers d_i's judgement changed, and the change obeys the transition probabilities in the $M(\alpha)$.

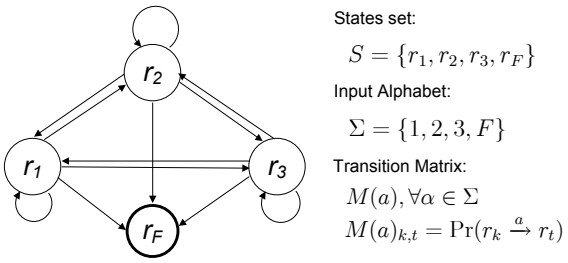


Figure 1: A probabilistic automaton (PA) with three judgement states r_1, r_2 and r_3, and one final state r_F.

In this paper, we focus on the inter-document judgement interference. Nonetheless, the proposed method can be generalised to incorporate other factors that may affect relevance judgements, such as evolving information needs or contexts, by encoding these factors into the PA alphabet.

2. DYNAMIC RELEVANCE JUDGEMENT

2.1 Modelling Judgement Interference

The probabilistic automaton (PA) is a generalisation of Markov Chains [4]. In general, the PA can have any finite number of states, which means that it can model the changing process of any finite number of judgement scores. For simplicity, suppose there are three judgement scores, i.e. 1, 2 and 3, where higher score means higher relevance.

As shown in Figure 1, the PA is a 5-tuple (S, Σ, M, $\mathbf{w}$, r_F). The states (also called judgement states) r_1, r_2 and r_3 in S correspond to the graded scores 1, 2 and 3, respectively. The states distribution $\mathbf{w}$ is a row vector that represents the probabilities of all judgement states, where the k^{th} element of $\mathbf{w}$ denotes the probability of the state r_k . We neglect the final state r_F in $\mathbf{w}$ to make the description simpler. For each symbol $\alpha \in \Sigma$, the $M(\alpha)$, which is a stochastic matrix (i.e.,

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

its each row sums to 1), represents transition probabilities among judgement states.

Let states distribution $\mathbf{w}_i$ be the initial states distribution of a document d_i to represent its static judgement (e.g., TREC judgement). The judgement of d_i, as interfered by the judgement of another document d_j, can be computed as:

$$\mathbf{w}_i^{\alpha} = \mathbf{w}_i \times M(\alpha). \quad (1)$$

where $\mathbf{w}_i^{\alpha}$ denotes the interfered judgements of d_i, and the α is the symbol generated after the judgement of d_j.

For example, if d_i's static judgement score is 2, then the $\mathbf{w}_i = [0, 1, 0]$, meaning that 100% probability of the state r_2. After judging d_j with the score 1, suppose the input symbol corresponds to d_j's judgement score, i.e., $\alpha = 1$ is generated, and accordingly the $M(\alpha) = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$, where the $M(\alpha)_{k,t}$=Pr($r_k \xrightarrow{\alpha} r_t$) denotes the transition probability from the state r_k to r_t. Based on Formula 1, the interfered judgement result $\mathbf{w}_i^{\alpha}$ is [0, 0.8, 0.2]. This means that after judging a less relevant d_j with score 1, the user may heighten the judgement score of d_i, i.e., the user has 0.8 probability of keeping his original judgement (score 2) for d_i, and 0.2 probability of changing d_i's score from 2 to 3.

This example may be over simplistic to assume that the symbols correspond exactly to judgment scores. One should consider the inter-document dependency, or other factors, e.g. the task and context, in the alphabet encoding.

2.2 Simulating Dynamic judgements

Based on a TREC collection $\mathcal{C}$, our aim is to simulate the dynamic judgements of every $d_i \in \mathcal{C}$, assuming the users are judging a list of documents $d_1 \cdots d_n$. It could be extremely complex to consider all the possible judgement interferences. Therefore, we design and run a user study on the inter-document judgement interference of the representative document pairs (d_i, d_j), in order to obtain the judgement transition matrices $(\forall\alpha)M(\alpha)$. The user study will be discussed in the next subsection. Here we assume that transition matrices are available and present how to use them to simulate the dynamic judgements of d_i.

Suppose the judgement of each d_j in the list $d_1 \cdots d_n$ generates an input symbol $\alpha(d_j)$ for the PA. After the user judged the documents $d_1 \cdots d_p$, the dynamic judgements of d_i at the position p ($p < n$), can be computed by:

$$\mathbf{w}_i^{x_p} = \mathbf{w}_i \times M(x_p). \quad (2)$$

where the string $x_p = \alpha(d_1) \cdots \alpha(d_p)$, and the $M(x_p) = \prod_{j=1}^{p} M(\alpha(d_j))$ is also a transition matrix. Note that the dynamic judgement modeling is also connected to the design of novel evaluation metrics considering user behavior [5, 3].

2.3 User Study Methodology

This proposed user study is to collect the initial judgement interference data for the simulation task described in Section 2.2. We need to study the inter-document judgement interference of each document pair (d_i, d_j), selected from TREC collections. The selection process should consider two factors between d_i and d_j, i.e., the document dependency (e.g., similarity) and judgement score difference, which are possibly related to the judgement interference. We plan to adopt the statistical document dependency, rather than the document dependency measured by ourselves or users, since we aim at using the collected interference data on selected documents to simulate dynamic judgements for a large number of other documents, for which it is too expensive to obtain the human-measured document dependency. We stick to use binary judgement in the initial stage, since it can reduce the efforts and randomness of user evaluation. The document pairs will be selected and assigned to several categories, and under the same category, all document pairs have the same document dependency degree and the same judgement score difference. We let each category correspond to each symbol α in the PA. We then study the pairwise-document judgement interference in each category, to learn the transition matrix $M(\alpha)$ for the corresponding symbol α.

For each document pair (d_i, d_j), we need to consider two situations, i.e., $rank(d_i) > rank(d_j)$ and $rank(d_i) < rank(d_j)$. In the first situation, users would be presented with d_i first and be asked to judge its relevance with respect to an information need (represented for instance as a request or a simulated work task [1]). Once d_i was judged, users would be asked to judge d_j's relevance. To close the circle, users would be asked to judge d_i again, i.e. to reconsider their previous judgement. In the second situation, two separate user groups would be involved. Users in the first group would be asked to judge d_j first and then judge the d_i, while in the second group, users would be only asked to judge the relevance of d_i. Collected data in both situations could help us to study how the judgement of d_i can be interfered by the judgement of d_j. We will further investigate the user study design under two situations in the future.

3. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to use a probabilistic automaton (PA) to model the dynamics of user-centred relevance judgements. We then further presented how to use PA, which can be trained through a user study, to simulate more dynamic judgement data, potentially useful for the IIR evaluation. In the future, we will adopt more appropriate strategies for encoding the alphabet Σ to include other factors that affect relevance judgement. We also plan to refine the user study methodology and carry out an extensive user study, of the best possible quality, as we can. Our ultimate goal is to simulate an IIR evaluation including the derivation of the appropriate evaluation metrics.

4. ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their constructive comments. This work is supported in part by the UK's EPSRC grant (No.: EP/F014708/2).

5. REFERENCES

[1] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):8–3, 2003.

[2] S. Mizzaro. Relevance: The whole (hi)story. *Journal of the American Society for Information Science*, 48:810–832, 1996.

[3] C. Olivier, M. Donald, Z. Ya, and G. Pierre. Expected reciprocal rank for graded relevance. In *CIKM '09*, pages 621–630, 2009.

[4] W.-G. Tzeng. Learning probabilistic automata and markov chains via queries. *Machine Learning*, 8:151–166, 1992.

[5] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Incorporating user behavior information in ir evaluation. In *SIGIR 2009 Workshop on Understanding the user- Logging and interpreting user interactions in information retrieval*, page 3, 2009.

Creating Re-Useable Log Files for Interactive CLIR

Paul Clough
Department of Information Studies
University of Sheffield
Sheffield UK
p.d.clough@sheffield.ac.uk

Julio Gonzalo
UNED
C/ Juan del Rosal
16 28040 Madrid, Spain
julio@lsi.uned.es

Jussi Karlgren
SICS
Isafjordsgatan 22
120 64 Kista, Sweden
jussi@sics.se

ABSTRACT

This paper discusses the creation of re-useable log files for investigating interactive cross-language search behaviour. This was run as part of iCLEF 2008-09 where the goal was generating a record of user-system interactions based on interactive cross-language image searches. The level of entry to iCLEF was made purposely low with a default search interface and online game environment provided by the organisers. User-system interaction and input from users was recorded in log files for future investigation. This novel approach to running iCLEF resulted in logs containing more than 2 million lines of data.

1. INTRODUCTION

iCLEF is the interactive track of CLEF (Cross-Language Evaluation Forum), an annual evaluation exercise for Multilingual Information Access systems. In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. Over the last years, iCLEF participants have typically designed one or more cross-language search interfaces for tasks such as document retrieval, question answering or text-based image retrieval. Experiments were hypothesis-driven, and interfaces were studied and compared using controlled user populations under laboratory conditions. This experimental setting has provided valuable research insights into the problem, but there are problems: user populations are small in size, and the cost of training users, scheduling and monitoring search sessions is very high. In addition, the target notion of relevance does not cover all aspects that make an interactive search session successful; other factors include user satisfaction with the results and usability of the user interface.

In 2008 and 2009 iCLEF organisers decided to try something different: we provided a default multilingual search system (called `Flickling`) which accessed images from `Flickr`. Many images in `Flickr` contain metadata in multiple languages thereby providing a semi-realistic search scenario. For users of `Flickling` the task was kept very simple: given an image find it again. Users did not know in advance the languages in which the image was annotated; therefore searching in multiple languages was essential to get optimal results. The iCLEF interactive search task was publicised to attract as many users as possible from all around the world and the whole evaluation event was centered around an online: the more images found, the higher a user was ranked. This helped to encourage participation but also make the evaluation task engaging and addictive. The focus of iCLEF 2008-09 was on search log analysis rather than the more traditional focus of system design.

Interaction by users with the system was recorded in custom log files which were shared with iCLEF participants for further analyses. `Flicking` also gathered input from users such as their language skills, reasons for aborting a search task and comments on their search experience. These logs, a form of simulated interaction data[1], provide a resource for further study of interactive cross-language search behaviour. Our experiences in iCLEF have been positive and for the first time in this kind of evaluation setting we have been able to produce a re-useable output from interative experiments.

2. ICLEF METHODOLOGY

Our primary goal for iCLEF 2008-09 was to harvest a large search log of users performing multilingual searches on Flickr. Participants in iCLEF 2008-09 could perform two tasks: (1) analyse log files based on all participating users (default option) and, (2) execute their own interactive experiments with the interface provided by the organisers.

Generation of search logs. Participants could mine data from the search session logs, for example looking for differences in search behaviour according to language skills, or correlations between search success and search strategies. Overall 435 subjects contributed to the logs.

Interactive experiments. Participants could recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc. iCLEF organisers provided assistance with defining appropriate user groups and image lists, for example, within the common search interface. Overall 6 groups participated in iCLEF and conducted experiments using `Flickling`.

2.1 Flickling

[1]These logs can be downloaded from `http://nlp.uned.es/iCLEF/`

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

The Flickling system was created as a default search application for the evaluations [1]. The intention was to provide a standard baseline interface that was independent of any particular approach to cross-language search assistance. Functionality provided by the system included: user registration and recording of language skills, localisation of the interface, monolingual and multilingual search modes, automatic term-by-term query translation facility, query translation assistant allowing users to select/remove translations and add their own, query refinement assistant allowing users to refine or modify terms suggested by Flickr, control of game-like features and post-search questionnaires (launched after image found/failed, and a final questionnaire launched after the user had searched for 15 images). Customised logging recorded a rich amount of user-system interaction including explicit success/failure of searches, users' profiles, and post-search questionnaires for every search (over 7,500 questionnaires)

2.2 Search Task

The task was organised as an online game: the more images found, the higher a user was ranked. Depending on the image, the source and target languages, this could be a very challenging task. To have an adaptive level of difficulty, we implemented a hints mechanism. At any time whilst searching, the user was allowed to quit the search (skip to next image) or ask for a hint. The first hint was always the target language (and therefore the search became mono or bilingual as opposed to multilingual). The rest of the hints were keywords used to annotate the image. Each image found scores 25 points, but for every hint requested, there was a penalty of 5 points. Initially a five minute time limit per image was considered, but initial testing indicated that such a limitation was not natural and changed users' search behaviour. We therefore decided to remove time restrictions from the task definition.

2.3 Generated Logs

Overall, the logs collected and released during the iCLEF 2008 and 2009 campaigns contain more than 2 million lines. Table 1 summarises the most relevant statistics of both search logs. The log files record various user-system interactions such as qureries, results, items clicked, selected query translations, query modifications, feedback from users and navigational actions (e.g. next/previous page). In total 435 users contributed to the logs and generated 6,182 valid search sessions (a session is when a user logs in and carries out a number of searches). The logs provide a rich source of information for studying multilingual search from a user's perspective. iCLEF participants analysed the log files in various ways [2] [3]. For example, to discover actions leading to aborting a search task, investigating the effects of language skills on search behaviour, observing the switching behaviour of users between languages within a search session, investigating when users added translation terms to the Flickling dictionary and why, and the effects of ambigious search terms on search results and user behaviour.

Table 1: Statistics of iCLEF 2008/2009 search logs.

	2008	2009
subjects	305	130
log lines	1,483,806	617,947
target images	103	132
valid search sessions	5,101	2,410
successful sessions	4,033	2,149
unsuccessful sessions	1,068	261
hints asked	11,044	5,805
queries in monolingual mode	37,125	13,037
queries in multi-lingual mode	36,504	17,872
manually promoted translations	584	725
manually penalised translations	215	353
image descriptions inspected	418	100

3. DISCUSSION

What we have done in iCLEF is to observe the user-system interactions of users recruited to perform assigned tasks (real users/interactions, simulated tasks). The focus has moved from comparing different aspects of cross-language search assistance using more classical TREC (Interactive) style of experiment to using a default system and (simple) set of search tasks to provide a more realistic setting in which to conduct experiments and record and analyse user-system interactions. The data collection has not been constrained to subjects recruited by participating groups, but also involved recruiting subjects on an individual basis with the aim of contributing to the search log. This community (and game-like) approach to generating search logs is perhaps one way to generate resources that can be used by researchers and without the limitations and ethical concerns imposed when using logs from commercial web search engines.

However, although the user-system interaction logs provide a useful re-useable resource for studying user-system interaction and search behaviours, we also recognise the limitations of our approach. For example, the logs reflect only a specific known-item search task and participants experiment only with a single pre-defined search interface. In the future one could imagine using the log files to record behaviour for a specific version of the Flicking interface with systematic modifications carried out and user behaviours compared and used to evaluate various forms of search assistance.

4. REFERENCES

[1] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F. (2009) FlickLing: a multilingual search interface for Flickr, In Working Notes for the CLEF 2008 Workshop.

[2] Gonzalo, J., Clough, P. and Karlgren, J. (2009), Overview of iCLEF2008: Search Log Analysis for Multilingual Image Retrieval, In Proceedings of 9th Workshop of the Cross-Language Evaluation Forum (CLEF'08), September 17-19 2008, LNCS 5706, pp. 227-235.

[3] Gonzalo, J., Peinado, V., Clough, P. and Karlgren, J. (2009) Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy Environment, In Working Notes for the CLEF 2009 Workshop.

Simulating Searches from Transaction Logs

Bouke Huurnink bhuurnink@uva.nl
Katja Hofmann k.hofmann@uva.nl
Maarten de Rijke derijke@uva.nl

ISLA, University of Amsterdam
Science Park 107, Amsterdam, The Netherlands

ABSTRACT

Computer simulations have become key to modeling human behavior in many disciplines. They can be used to explore, and deepen our understanding of, new algorithms and interfaces, especially when real-world data is too costly to obtain or unavailable due to privacy or competitiveness reasons.

In information retrieval, simulators can be used to generate inputs from simulated users—including queries, clicks, reformulations, and judgments—which can then be used to develop a deeper understanding of user behavior and to evaluate (interactive) retrieval systems.

The trust that we put in simulators depends on their validity, which, in turn, depends on the data sources used to inform them. In this paper we present our views on future directions and challenges for simulation of queries and clicks from transaction logs.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms: Experimentation

Keywords: Simulation, Transaction Logs, Evaluation

1. INTRODUCTION

Simulation offers a source of experimental data when real-world information is either unobtainable or too expensive to acquire. This makes simulation a valuable solution for evaluating information retrieval (IR) theories and systems, especially for interactive settings.

One approach to creating simulators for IR evaluation is to build simulation models that incorporate manually created queries and relevance judgments, as in the Cranfield tradition. A problem here is that it is not clear in how far such explicit judgments reflect how users would interact with a real IR system.

In this paper we discuss search engine transaction logs as a source of data for building simulators for IR evaluation. Transaction logs typically record, among other things, sequences of queries and result clicks issued by a user while using a search engine [6]. The data is collected unobtrusively, thereby capturing the actions of users "in the wild." In addition, large quantities of searches and clicks can be gathered. This makes them a rich source of information.

For privacy and competitiveness reasons transaction logs are rarely made publicly available. However, simulators can be developed to generate artificial transaction log data not linked to any real users. Such a simulator can then be released to outside parties, for example in order to test our theories about searcher behavior or to evaluate the performance of (interactive) retrieval systems. In the process of creating such simulators, theories about search engine users could be tested by incorporating them in the simulation models.

So far, relatively little research has been conducted into developing simulators based on the searches in transaction logs. Below, we will discuss future directions for such research. We start by giving a brief overview of the state-of-the-art in simulation for retrieval evaluation in Section 2. Next, we outline possible application areas for transaction log-based simulators, and discuss some of the challenges that need to be addressed, in Section 3. We summarize our views in Section 4.

2. SIMULATION FOR IR

We now sketch some recent developments in simulation for IR purposes. In particular, we focus on the specific aspects of user interaction that have been simulated, and on how simulators have been validated.

A common form of data used to inform simulators in IR is manually created sets of queries and explicit relevance judgments. There are multiple scenarios where such data has been exploited. For example, an approach for evaluating novel interfaces for presenting search results is to create a simulated user who clicks on relevant documents that appear on the screen [4, 10]. Another example is in simulating queries for retrieval evaluation, for example by creating new queries or sequences of queries from an existing query and set of relevant documents [1, 8]. Here all simulated queries generated from the same truth data are associated with the same set of truth judgments.

When simulators are not informed by manually created queries and document judgments, key challenges are to (1) create queries, and (2) identify relevant documents for a given query. One solution is to use document labels to identify groups of related "relevant" documents: these documents are then considered relevant to the queries generated from their combined text [7] (the document labels themselves are not used in the query generation process). Another solution is to address retrieval tasks where only one document is considered relevant to a query, as in for example known-item search. Here Azzopardi et al. [2] used document collections in multiple European languages to simulate pairs of queries and relevant documents, and compared them to

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

sets of manually created known-item queries.

For the simulation approaches mentioned above, it is unclear to what extent they reflect real-world queries and result interactions. This can be addressed through the use of transaction logs, as we discuss in the next section.

3. LOG-BASED SIMULATORS

Following on from our brief description of simulation in IR in general, we turn to directions for research in the development of transaction log-based simulators for IR.

The first direction for research is the realistic simulation of queries and clicks. We ourselves investigated this task using transaction logs from an archive search engine [5]. We validated each simulator by ranking different retrieval systems on its output data, and comparing this to a "gold-standard" retrieval system ranking obtained by evaluating the systems on actual log data. We found that incorporating information about users in the simulation model improved the simulator output. Another approach to simulation of queries and clicks was taken by Dang and Croft [3], who worked in a web setting. Here, anchor texts were available, which they used as simulated queries. The purpose here was to evaluate the effect of different query reformulation techniques. The authors compared retrieval performance on the simulated queries to retrieval performance on queries and clicks taken from an actual search log, and found that the simulated queries showed retrieval performance similar to real queries.

A second direction for research is the simulation of *sessions*—sequences of queries and clicks. Retrieval evaluation with explicitly judged queries and documents generally considers each query in isolation. Transaction logs, however, offer us a wealth of information about query modification behavior, and there is broad interest in the IR community about using such information for retrieval system evaluation. What is needed here is to develop and incorporate into the simulator insights about possible "moves" in a session, based on different assumptions about user intent.

A key problem when designing a simulator is ensuring that it is valid for the purpose for which that simulator is developed. Sargent [9] identifies three types of validity in the simulation model: *conceptual model validity*, the validity of the underlying assumptions, theories, and representations of the model; *model verification*, the correctness of the programming and implementation of a conceptual model; and *operational validity*, the accuracy of the output data created by a simulation model. Transaction logs, due to the large number of interactions that they record, offer a wealth of data for quantitatively determining operational validity. The measure of validity will vary according to the purpose of the simulator. For example, when generating simulated queries and clicks for comparing retrieval systems, a valid simulator is one that produces output data that scores different retrieval systems in the same way as data derived from actual transaction logs. Here a rank correlation coefficient such as Kendall's τ can be used to compare the rankings of retrieval systems on real and simulated output [5].

An open question in developing transaction log-based simulators is whether those simulators are transferable to new domains. The data contained in a transaction log represents the actions of a specific set of users on a specific search engine. A simulator that captures general aspects of user behavior could successfully be applied to new collections.

4. SUMMARY

In this position paper we have discussed the potential of transaction log data for developing and validating simulators for IR experiments. In particular we have discussed the creation of simulators for two scenarios: generating evaluation testbeds consisting of artificial queries and clicks; and creating simulations of session behavior in terms of sequences of queries and clicks. We discussed some of the challenges in creating such simulators, including the validation of simulation output. It is our view that transaction logs pose a rich source of information for simulator development and validation. By producing simulators that accurately reproduce the queries and clicks contained in transaction logs, we will not only be able to generate data for different retrieval tasks, but we will also obtain a better understanding of the behavior of users "in the wild."

Acknowledgements This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Center for Creation, Content and Technology (CCCT), by the Dutch Ministry of Economic Affairs and Amsterdam Topstad under the Krant van Morgen project, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.-512, 612.061.814, 612.061.815, 640.004.802.

REFERENCES

[1] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *SIGIR '09*, pages 556–563. ACM, 2009.

[2] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six European languages. In *SIGIR '07*, pages 455–462. ACM, 2007.

[3] V. Dang and B. W. Croft. Query reformulation using anchor text. In *WSDM '10*, pages 41–50. ACM, 2010.

[4] O. de Rooij and M. Worring. Browsing video along multiple threads. *IEEE Trans. Multimedia*, 12(2):121–130, 2010.

[5] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Validating query simulators: An experiment using commercial searches and purchases. In *CLEF '10*, 2010.

[6] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *JASIS&T*, 52(3):235–246, 2001.

[7] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *DL '06*, pages 286–295. ACM, 2006.

[8] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based IR evaluation needs extension toward sessions—a case of extremely short queries. In *AIRS '09*, pages 63–74. Springer-Verlag, 2009.

[9] R. Sargent. Verification and validation of simulation models. In *WSC '05*, pages 130–143. Winter Simulation Conference, 2005.

[10] R. White, I. Ruthven, J. Jose, and C. Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM TOIS*, 23(3):361, 2005.

A Methodology for Simulated Experiments in Interactive Search

Nikolaos Nanas
Centre for Research and Technology - Thessaly
Greece
n.nanas@cereteth.gr

Udo Kruschwitz, M-Dyaa Albakour, Maria Fasli
University of Essex
Colchester, United Kingdom

Dawei Song, Yunhyong Kim, Ulises Cerviño
Robert Gordon University
Aberdeen, United Kingdom

Anne De Roeck
Open University
Milton Keynes, United Kingdom

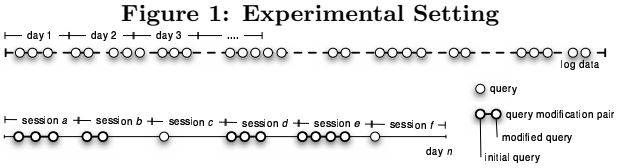
1. INTRODUCTION

Interactive information retrieval has received much attention in recent years, e.g. [7]. Furthermore, increased activity in developing interactive features in search systems used across existing popular Web search engines suggests that interactive systems are being recognised as a promising next step in assisting information search. One of the most challenging problems with interactive systems however remains evaluation.

We describe the general specifications of a methodology for conducting controlled and reproducible experiments in the context of interactive search. It was developed in the AutoAdapt project[1] focusing on search in intranets, but the methodology is more generic than that and can be applied to interactive Web search as well. The goal of this methodology is to evaluate the ability of different algorithms to produce domain models that provide accurate suggestions for query modifications. The AutoAdapt project investigates the application of automatically constructed adaptive domain models for providing suggestions for query modifications to the users of an intranet search engine. This goes beyond static models such as the one employed to guide users who search the Web site of the University of Essex[2] which is based on a domain model that has been built in advance using the documents' markup structure [6].

Over a period of more than two years we have collected a substantial query log corpus (more than 1 million queries) that records all queries and query modifications submitted to the University of Essex search engine. These logs include information about the searching session id, date, the queries and their modifications. Query modifications derive from the user selecting one of the modified queries suggested by the static domain model, from the suggestions proposed by the system which have been extracted from the top-matching snippets, or the user defining a new query in the provided text box. In any case, we are interested in the queries that a user has submitted to the system after the initial query within a session or an information-seeking dialogue (a "search mission").

Figure 1: Experimental Setting



2. SIMULATED QUERY RECOMMENDATION EXPERIMENTS

Here we propose a methodology for performing simulated query recommendation experiments based on log data of the type outlined above. The methodology can be used to perform both "static" and "dynamic" experiments. In particular, we treat the log data as a collection of query modification pairs (initial query – modified query) for building a domain model, but also for evaluating its ability to recommend accurate query modifications. Any log file that records the user queries along with a time stamp and a session id can be used. The log data are traversed in chronological order and in daily batches (see fig. 1). Within each day, subsequent queries submitted within the same searching session are treated as a query modification pair. For instance in the example of figure 1, there are eight query modification pairs within the six sessions of day n.

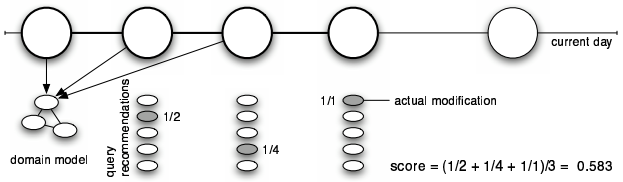
In the static experiments, we start with an existing domain model that remains unchanged during the evaluation process. The model's evaluation is performed on a daily basis as depicted in figure 2. It only takes place for days with at least one query modification pair. For example, let us assume that during the current day, three query modifications have been submitted (fig. 2). For each query modification pair, the domain model is provided with the initial query and returns a ranked list of recommended query modifications. We take the rank of the actual modified query (i.e., the one in the log data) in this list, as an indication of the domain model's accuracy. The assumption here is that an accurate domain model should be able to propose the most appropriate query modification at the top of the list of rec-

[1] http://autoadaptproject.org
[2] http://www.essex.ac.uk

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

ommended modifications. This is based on the observation that users are much more likely to click on the top results of a ranked list than to select something further down [4], and it seems reasonable to assume that such a preference is valid not just for ranked lists of search results but for lists of query modification suggestions as well. The underlying principle of a graded scoring is inherited from DCG [3].

Figure 2: Daily Model Evaluation



So for the total of three query modifications in the current day, we can calculate the model's accuracy score as $(1/r_1 + 1/r_2 + 1/r_3)/3$, where r_1 to r_3 are the ranks of the actual query modifications in the list of modifications recommended by the model in each of the three cases. In the figure's example the models score would be $1/2+1/4+1/1 = 0.583$. More generally, given a day d with Q query modification pairs, the model's accuracy score S_d for that day is given by equation 1 below.

$$S_d = (\sum_{i=1}^{Q} \frac{1}{r_i})/Q \quad (1)$$

Note that in the special case where the actual query modification is not included in the list of recommended modifications then $1/r$ is set to zero. The above evaluation process results in an accuracy score for each logged day for which at least a query modification pair exists. So overall, the process produces a series of scores for each domain model being evaluated. These scores allow the comparison between different domain models. A model M_1 can therefore be considered superior over a model M_2 if a statistically significant improvement can be measured over the given period.

In the case of dynamic experiments, the experimental process is similar. We start with an initially empty domain model, or an existing domain model. Like before, the model is evaluated at the end of each daily batch of query modifications, but unlike the static experiments it uses the daily data for updating its structure. This is essentially a continuous learning problem, where the domain model has to continuously learn from (adapt to) temporal query modification data. Again, we treat a model as superior over another (possibly static one) if an improvement can be observed that is significant.

3. DISCUSSION

The proposed methodology addresses one major weakness of interactive information retrieval, in that it does not involve users and is purely technical. However, the methodology cannot replace user experiments. One reason is that we cannot assume that the selection of a query suggestion will actually be successful in the sense that it leads to the right documents or narrows down the search as expected. A number of other issues remain. For example, we do not try to identify which query modifications within a session are actually related. We consider the entire session in this context. This implies that even subsequent queries that are not related are treated as a query modification pair, thus adding noise to the data. Automatically identifying the boundaries of sessions is a difficult task [2]. One of the reasons is that a session can easily consist of a number of *search goals* and *search missions* [5]. However, we assume that this noise does not affect the evaluation methodology because: a) it is common for all evaluated models and b) no model can predict an arbitrary, unrelated query modification from the initial query. In other words, all evaluated models will perform equally bad for such noisy query modification pairs. But note, that the fairly simplistic fashion of constructing query pairs can easily be replaced by a more sophisticated method without affecting the general methodology proposed in this paper.

Another issue is the question of how the presentation of query suggestions might influence the users' behaviour and how different ways of presenting such query modifications may affect their perceived usefulness.

4. NEXT STEPS

Our plan is to initially use the described methodology to evaluate a number of adaptive algorithms using the log data we have collected. We have already started conducting experiments, following this methodology, for static domain models as well as an adaptive model we have developed and which has been shown to be effective in learning term associations in a user study [1].

5. ACKNOWLEDGEMENTS

AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1.

6. REFERENCES

[1] S. Dignum, U. Kruschwitz, M. Fasli, Y. Kim, D. Song, U. Cervino, and A. De Roeck. Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs. In *Proceedings of WI'10*, Toronto, 2010. Forthcoming.

[2] A. Göker and D. He. Analysing web search logs to determine session boundaries for user-oriented learning. In *Proceedings of AH '00*, pages 319–322. Springer, 2000.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[4] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, pages 154–161, Salvador, Brazil, 2005.

[5] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceeding of CIKM*, pages 699–708, 2008.

[6] U. Kruschwitz. *Intelligent Document Retrieval: Exploiting Markup Structure*, volume 17 of *The Information Retrieval Series*. Springer, 2005.

[7] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 42:43–92, 2008.

Recovering Temporal Context for Relevance Assessments

Omar Alonso & Jan Pedersen
Microsoft Corp.
1056 La Avenida
Mountain View, CA 94043
{omalonso, jpederse}@microsoft.com

ABSTRACT

Relevance assessment is a key aspect of information retrieval evaluation. For a typical web search engine, an editor will look at the search engine results page and judge if the results are relevant or not.

For certain types of queries, those that are time sensitive, it is desirable to be looking at the search results as soon as the event is happening. Unfortunately, editors are not always available at the right time so it is difficult to assess relevance after the fact. This is, in particular, true in the case of real-time search: the need to be present observing the results otherwise the content maybe out of date. In this paper, we present the problem of gathering relevance assessments in the absence of temporal context and suggest a technique for addressing the problem.

Categories and Subject Descriptors

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

General Terms

Measurements, performance, experimentation

Keywords

Real-time search, user feedback, relevance assessment.

1. INTRODUCTION

Situational relevance is often defined as a dynamic concept that depends on users' judgments of quality of the relationship between information and information need at a certain point in time [2]. In the context of real-time search, time plays a crucial role.

Micro-blogging sites like Twitter are currently being used as a communication channel to break out events or comment on current episodes. If the event is considered to be important like an annual entry in a calendar (Mother's Day), a celebrity death (Gary Coleman), a major sports event (World Cup), or a finance crisis (Dow Jones), users would write at an incredible speed.

For the user who is looking at this information, there is a perception that if new content (status updates from Twitter or Facebook, for example) appear rapidly on the page, then the results from the search engine have "real-time" content. In other words, the effect is similar to observing a river: the flow is what makes it interesting.

Previous research has shown that recent information is one the factors that that affect relevance judgments [3]. For real-time content, we are interested in timely (up-to-date) information. We can argue that another criterion for relevance evaluation is the *temporal context* of the content that gets broadcasted by the users. A related feature is the *velocity* at which the updates appear in the web page creating the real-time stream effect.

Our goal is to use time data to reconstruct the situation or setting, for assessing relevance in the context of real-time search.

Research questions:

1. Does it make sense to assess relevance of real-time results?
2. How can we simulate that effect in real-time search when people are assessing content offline?
3. If so, how can we implement a solution?

2. THE SLIDE-SHOW TECHNIQUE

We simulate the interactivity of the stream flow by taking a number of screen captures every time interval *t* and replay them for the user. The idea is to use the slide show as an approximation of the stream. By adding a time-stamp, it is easy to see the content as it happened over time. Our approach is very similar in spirit to the stream-based/time-view for representing usage of a system [1].

The implementation works as follows. The prototype issues a query to the search engine and captures a number of screenshots along with the timestamp during a time interval as parameter. Say that we would like to simulate the stream for 1 hour and capture a screenshot every 5 minutes. After one hour, we would have 12 images that are replayed in a web page so the editor can assess the content.

We ask assessors to rate the slide-show content given a query, with the following scale:

- Very interesting, relevant, and timely.
- Relevant but somewhat old content.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva, Switzerland.

- Not Relevant

In the instructions we mention that relevance in our case means that the content is relevant and fresh. That is, the topic is discussed and has timely content.

Figure 1 shows two screenshots for the query {Mother's Day} captured on the day of the event (5/9/2010). The first one was captured at 7:23am and the second at 8:23am. Note that for the second screenshot, there is a tweet that says "good morning …" at time-stamp 8:23am.

In Figure 2, we show a couple of the screenshots for the France-Uruguay World Cup game (6/11/2010). In this example, we started capturing the stream before the game to analyze what users were saying. As expected, people make predictions about the final score before the start of the match. Then, all the content is mainly about what's going on during the game. In this example, our tool can be used to describe the match as it happened.

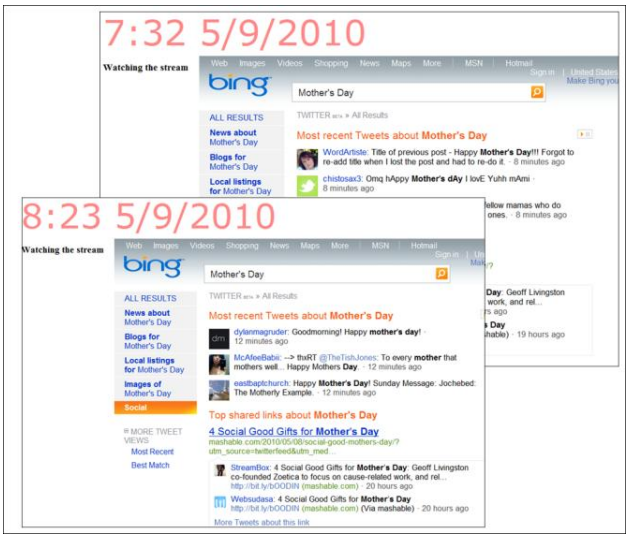


Figure 1. Slide show for the query {Mother's Day}.

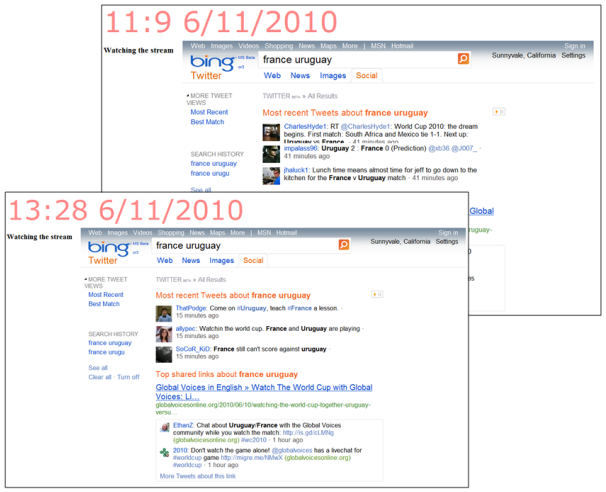


Figure 2. Slide show for the query {france uruguay}.

3. EXPERIMENTATION

To test our approach, we tried a number of examples that contained trending topics or highly volume queries. The goal was to collect feedback from the internal team and use it as proxy before using editors. Internal testing showed that people like the idea of watching the stream for two main reasons:

1. Observe (and record) if the search engine is serving time-sensitive events. This has more to do with presenting updates that are relevant to the query in a timely manner.
2. Watch how the link ranking changes over a particular period of time. For those queries that represent a breaking event, people would often share links. Exploring how those links appear and change over time in the results is also very useful.

The internal feedback was helpful for understanding more if velocity plays an important role in relevance evaluation criteria. People like the notion of activity or dynamic data change in real time. Of course, that is true for events that have a high load of content.

Other user feedback included a feature to replay the content at different speeds. Very much like fast forward or rewind, the internal team found it useful to debug certain behavior by pointing to a specific time-stamp

4. CONCLUSIONS

We presented a technique for simulating a stream of real-time results and make it available via a web page so human editors can assess the relevance giving the temporal context. Temporal context and velocity are important factors that can influence relevance.

The prototype was deployed internally and provided insights on how people perceived relevance for real-time updates.

People see the value of assessing real-time results. However, it is not clear what the best way is to test and/or evaluate this type of retrieval. We plan to continue refining our technique focusing on usage performance and metrics.

5. REFERENCES

[1] L. Azzopardi. "Usage Based Effectiveness Measures: Monitoring Application Performance in Information Retrieval". *Proc. of CIKM'09*, pp. 631-640, 2009.

[2] P. Borlund. "The Concept of Relevance in IR", *J. Am. Soc. Information Science and Technology*, 54(10):913–925, 2003.

[3] C. Barry and L. Schamber. "Users' Criteria for Relevance Evaluation: A Cross-Situational Comparison". *Information Processing and Management*, Vol. 34. No. 2/3, 1998.

Simulating User Interaction in Result Document Browsing

Paavo Arvola
Dept. of Information Studies and Interactive Media
University of Tampere, Finland
paavo.arvola@uta.fi

Jaana Kekäläinen
Dept. of Information Studies and Interactive Media
University of Tampere, Finland
jaana.kekalainen@uta.fi

ABSTRACT

A user can explore a result document in various ways. In this study we seek means to simulate a user browsing a result document. The simulation starts when the user accesses a result document and ends when the user leaves it. The outcome of a simulated strategy depends on two facts: Did the user find relevant information from the document, and how much effort the user wasted while exploring the document?

Categories and Subject Descriptors

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

General Terms

Design.

Keywords

Focussed retrieval, simulated interaction.

1. DOCUMENT BROWSING

In many cases the text document as a whole is not needed to satisfy the user's information need. Instead, the relevant parts embedded within the document need to be sought by the user. Focused retrieval systems aim to provide a better access to the relevant document parts. In any case, finding the relevant parts requires browsing within the document's content and clearly there are numerous ways to do so. In addition the user may stop exploring the document basically at any point.

The idea of this study is to form a general browsing model and define the alternatives the user has in finding the relevant information when exploring a single document. Based on the model, the user interaction with the document content can be simulated. Eventually, this study aims to contribute to the in context tasks of INEX [2], where results are grouped per article.

The flow diagram in Figure 1 represents the user's browsing model within a document. First, the user accesses a document and starts to browse. The browsing may simply lead either to reading the document from the beginning or some other point. For example, the user may utilize an interface providing e.g. guiding gadgets to reach the expected best entry point. Nevertheless, any browsing strategy will eventually lead to reading of some text passage (or e.g. picture, if sought) and determining its relevance. After assessing the passage's relevance the user decides either to read another passage or move to the next document (or e.g. perform a new search). If the passage seems to completely fulfill the users information needs, he or she leaves the document. On the other hand the user may be bored and discover that there is no (further) relevant material to be read. In case the user is still willing to continue, he or she faces again the alternatives in how to proceed browsing the document.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

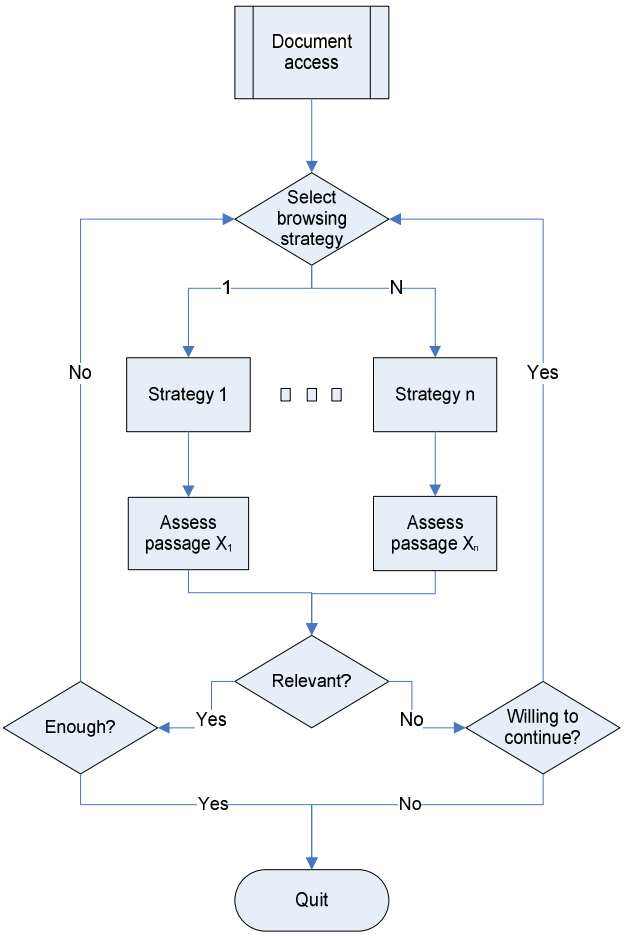


Figure 1. A flow diagram of a user's browsing within a document

2. USER SIMULATION

Next, we will present our consideration of simulating the user within the presented model. The simulation starts when the user accesses a result document and ends when the user leaves it. At this stage we do not offer any exact specification for the simulation, or metrics to measure it. Instead, we define which simulated user attributes affects the costs the user wastes when exploring the document. Accordingly, in Figure 1 the user

attributes (i.e. decisions) to be simulated are marked as diamonds and the costs to be calculated as rectangles.

2.1 Simulated user parameters

Selecting browsing strategy. Right after the document access and later, after assessing a passage[1] and agreeing to continue with the current document, the user decides how to proceed to the next passage to be assessed. Depending on the user's location and the available navigational gadgets, the user has a finite set of meaningful strategies to browse on.

Relevance. The relevance of a read passage is dependent on the user's valuations. However, in simulations, a recall base having the relevant text assessed beforehand [3] could be used. Principally, the degree of relevance may vary. If graded relevance is used, different relevance interpretations are available.

Willingness to continue. In the event of a non-relevant passage the willingness to continue is related to the concept of tolerance-to-irrelevance (T2I) [4], where after the user exceeds some limit of reading irrelevant material, he or she gets bored and leaves the current document. The T2I may vary according to the user profile or user interface (e.g. screen size [1]).

Enough. After reading relevant material, the user may assume that there is still relevant material in the current document and wants to continue exploring. This depends at least on the relation of the document characteristics and the topic type. Namely, in fact finding it is enough to find the first relevant passage and stop browsing, whereas in e.g. information gathering the browsing may continue.

2.2 Calculating costs

Browsing strategy. Simply reading on a document requires no *browsing* costs (instead assessing costs), but using e.g. navigational gadgets or performing a keyword finding requires some effort, amount of which depends on the nature of the selected strategy.

Assessing a passage. A comprehensive reading of a passage requires effort, which is related to the length of the passage and also its relevance.

Quit. After seen enough, the user quits browsing the document. The outcome of a simulated strategy depends on two factors:

Success: Did the user find relevant material from the document and how much?

Cost: How much effort the user wasted while exploring the document?

The amount of costs varies according to the simulated user parameters and they consist of executing a browsing strategy plus assessing the relevance of the consequential passage.

When it comes to the success, a non-relevant document should not be rewarded, but even in the event of a relevant document, the relevant content is not always reached, partially or at all. This is because the user may stop reading the current document because of satisfaction or T2I. The opposite case is that when all the relevant content is read, the user is still seeking for more.

[1] In this study the passage is not predefined, but it seems reasonable that a passage should have a meaningful length (literally).

3. DISCUSSION

A search process is a ternary relationship between the user, the retrieval system and the user interface, which affects the browsing alternatives. Therefore, the user decision driven process described in this study is neither completely coincidental nor completely deterministic and under some circumstances the alternatives can be controlled. For example a small display restricts the number of browsing alternatives; with such a device it is not meaningful or even possible to use a flip through browsing strategy for a long document.

Accordingly, in our previous study [1], we introduced the metrics called localizing effort (*LE*) and character precision/recall (*ChPR*) and justified our approach with a small screen scenario. For those metrics, we considered T2I and reading order to be the modeled user parameters. However, the user simulation was implicit and simplified: the T2I was bond to the screen size and two generic browsing strategies were introduced. The baseline (default) browsing strategy was simply reading the document consecutively from the beginning to the breaking condition, which was either T2I or the end of the document. This strategy was compared to the focused strategy, where the retrieved passages were read first in document order.

Because of the space limitations, the *ChPR* and *LE* metrics are not presented here in detail. In any case, while the metrics do not assume any specific reading order of the documents, they can be developed and adjusted to serve the varying simulated browsing scenarios based on the model presented in this study. In addition the lacking satisfaction attribute (Enough) as a breaking condition can apparently easily, as was T2I, to be included to those metrics.

4. ACKNOWLEDGEMENTS

The study was supported by the Academy of Finland under grants #115480 and #130482.

REFERENCES

[1] Arvola, P., Junkkari, M., and Kekäläinen, J. Expected reading effort in focused retrieval evaluation. To appear in Information Retrieval, Vol. 13, Nr. 4, 25 pages, 2010.

[2] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, and A. Trotman, editors, INEX 2006 Workshop Pre-Proceedings, pages 381–388, 2006.

[3] Piwowarski, B., and Lalmas, M. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of CIKM '04*, 361–370, 2004.

[4] de Vries, A.P., Kazai, G., and Lalmas, M. 2004. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *Proceedings of RIAO 2004*, 463-473, 2004.

Query and Browsing-Based Interaction Simulation in Test Collections

Heikki Keskustalo and Kalervo Järvelin
Department of Information Studies and Interactive Media
FI-33014 University of Tampere, FINLAND
heikki.keskustalo@uta.fi, kalervo.jarvelin@uta.fi

ABSTRACT
In this paper we propose using query and browsing-based interaction simulations to resolve some of the discrepancies between the traditional Cranfield-style IR evaluation experiments and type of searching common in real life. We describe our approach to perform simulations (short-query sequences based on prototypical modifications) in test collections and suggest constructing and utilizing databases containing a facet analysis to aid expressing topical session strategies in simulations.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Selection process

General Terms
Measurement, Performance, Theory

Keywords
Iterative search process, session-based evaluation, simulation

1. INTRODUCTION
Valid study designs should incorporate major factors of the phenomenon explained. In traditional test collection-based experiments the user is not explicitly modeled, and only one query is constructed, typically using a subset of terms available in selected fields of a test topic (e.g., title, description) thus describing a static and well-defined topic. In real life, on the contrary, different kinds of searchers and searching situations exist, and the information needs may be ill-defined, dynamic and difficult to express [4]. Real searchers often prefer using very short queries (only 1-2 keys) [6,14]; try more than one query per topic if needed; cope by making minor query modifications; browse as little as possible (not past the first 10 documents); and stop soon after finding few relevant documents [6, 7, 12, 14, 15]. Also entirely different wordings may be used in real life even by experts encountering an identical search task. To resolve such differences traditional test collection-based IR experiments should be extended towards real life. In this paper we suggest an extension by utilizing (i) query and browsing-based interaction simulations, and (ii) traditional/extended test collections.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

2. QUERY AND BROWSING-BASED SIMULATIONS
Simulation consists of experimentation using a model which represents relevant features of the real world in a simplified form [1]. We assume a model of real life searching where the simulated user will launch an initial query; the IR system returns a ranked list of documents; and the user will browse some of the documents retrieved. At some rank point regarding any particular query the user will stop browsing and instead launch another query (followed by browsing), or give up the session.

We focus on users who prefer short queries [6, 14, 16]; revise their queries by using few term-level moves [17]; use multiple query iterations [4,11]; select different wordings, and avoid extensive browsing [3]. Our research question was to study how effective are *sequences of very short queries* combined with impatient browsing, compared to the traditional scenario of using only one longer query and patient browsing. To study this issue we performed a simulation in [8] based on a user model.

We defined a user model M=<R, B, SS> where R denotes the requirement for document relevance (stringent or liberal relevance threshold), B denotes the user's willingness to browse (at most B top documents) and SS denotes a set of *session strategies*. This simplified model explicates the independent variables we wanted to vary; in addition there are also other attributes which could be included (e.g., the typical size of the query, and the source of query terms). We restrict our attention to impatient users – preferring short queries (in most cases 1-3 words), tolerating *limited browsing* (10 documents per query, i.e., B=10), and quitting after finding *one* relevant document assuming liberal (R=1) and stringent (R=3) relevance thresholds. We used three session strategies (S1-S3) in the experiments. Each strategy determines a specific way to construct the "next query" in case the previous one failed. Terminological moves observed in the empirical query session data by Lykke et al. [10] were used to define the following strategies (the k_i stand for search terms):

S1: Sequence of one-word queries (k_1->k_2->k_3->..)

S2: Incremental extension (k_1 -> k_1 k_2 -> k_1 k_2 k_3 ->..)

S3: Varying the third word (k_1 k_2 k_3 -> k_1 k_2 k_4 ->..)

Having these strategies formulated we turned toward a TREC test collection and asked two groups of test persons (students and staff members of our institution) to suggest search terms for short queries based on the test topics. These terms were used to con-

struct up to five queries for each topical session. The baseline strategy S4 was one verbose TREC-style query, in which case at most top-50 documents were examined by the simulated user, in chunks of 10. The short query sessions turned out to be surprisingly effective. The short queries were inferior if only one query per topic is assumed, but they made sense as sequences. For example, S1 was successful (i.e., P@10>0) in more than half of the topics after only three single-word queries were tried out [8]. We only studied a limited number of prototypical session strategies (S1-S3) while in real life further strategies may be used and different types of terminological moves may be mixed. Another limitation in the simulation was that we used a limited number of search terms. Therefore it would be desirable to (i) recognize further session strategies and (ii) explore their effectiveness based on a larger set of terms. We will describe this issue next.

3. EXTENDING TEST COLLECTIONS

Constructing queries by using only the terms in the fields of the topic descriptions of test collections is inherently limited. In real life different query sequences are expected. The intention of [8] was to expand test collection-based evaluation toward multiple query sessions. However, the idea can be extended further. First, the topic descriptions can be the target of intellectual analysis by test searchers in a (simulated) task context [5] who suggests various query approaches and provide *searcher warrant*. Secondly, to guarantee terminological and conceptual exhaustivity and *literary warrant*, a facet analysis can be performed to relevant documents so that the facets and their linguistic expressions in relevant documents are recognized. In fact, this has been performed by a group of UTA researchers for one test collection. The end result is a structural description of the concepts, conceptual relations, and their expressions in relevant documents. Such a description covers the possible "terminological worlds" reasonably available for the simulated searcher. The data can be used in simulations to systematically study the effectiveness of various session strategies (e.g., how successful is it to move from expressing the most specific facet "[A]" to expressing pairwise facets "[A] and [B]" using available expressions a_1, a_2, b_1, …). The effectiveness of searcher warrant and literary warrant queries and of their intersection queries may also be studied. In [9] we discuss a graph-based method to explore short query sessions.

4. CONCLUSION

Figuring out good query terms, combinations of terms, and identifying which moves are effective between differently behaving queries is an important research task which cannot be answered if queries are not considered as sequences. We suggest that the effectiveness of interactive searching could be studied in test collections by simulating users having variying behavior properties (e.g., different preferences regarding how to construct queries, how to browse, what kind of search result is required, and what kind of limitations the simulated searcher sets regarding the interactive searching process itself, e.g., the number of iterations). This approach may help resolving the disparity of the real life searching observed and some implicit assumptions of the traditional test collection-based evaluation.

5. REFERENCES

[1] E. Adams and A. Rollings (2007) Fundamentals of Game Design, Prentice Hall, New Jersey, 669 p.

[2] L. Azzopardi (2007) Position Paper: Towards Evaluating the User Experience of Interactive Information Access systems. In SIGIR'07 Web Information-Seeking and Interaction Workshop.

[3] L. Azzopardi (2009) Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In SIGIR'09, 556-563.

[4] N.J. Belkin (1980) Anomalous States of Knowledge as a Basis for Information Retrieval. Canadian Journal of Information and Library Science, 5, 133-143.

[5] P. Borlund (2000) Experimental Components for the Evaluation if Interactive Information Retrieval Systems. Journal of Documentation 50(1), 71-90.

[6] M.B.J. Jansen, A. Spink, T. Saracevic (2000) Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, IP&M 36(2), 207-227.

[7] K. Järvelin, S.L. Price, L.M.L. Delcambre, M.L. Nielsen (2008) Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In ECIR'08, 4-15.

[8] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, M. Lykke (2009) Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries. In AIRS'09, 63-74.

[9] H. Keskustalo, K. Järvelin (2010) Graph-Based Query Session Exploration Based on Facet Analysis. Submitted to SimInt 2010.

[10] M. Lykke, S. L. Price, L. M. L. Delcambre, P. Vedsted (2009) How doctors search: a study of family practioners' query behavior and the impact on search results (submitted).

[11] I. Ruthven (2008) Interactive Information Retrieval. In Annual Review of Information Science and Technology, Vol. 42, 2008, 43-91.

[12] C. Smith, P. Kantor (2008) User Adaptation: Good Results from Poor Systems. In SIGIR'08, 147-154.

[13] E. Sormunen (2002) Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In SIGIR'02, 324-330.

[14] D. Stenmark (2008) Identifying Clusters of User Behavior in Intranet Search Engine Log Files. JASIST 59(14):2232-2243.

[15] A. Turpin, W. Hersh (2001) Why Batch and User Evaluations Do Not Give the Same Results. In SIGIR'01, 225-231.

[16] P. Vakkari (2000) Cognition and changes of search terms and tactics during task performance: a longitudinal study. In RIAO 2000, 894-907.

[17] P. Vakkari (2002) Subject Knowledge, Source of Terms, and Term Selection in Query Expansion: An Analytical Study. In ECIR'02, 110-123.

Evaluating a visualization approach by user simulation

Michael Preminger
Oslo University College
michaelp@hio.no

ABSTRACT

This paper briefly describes an approach to IR evaluation by simulation of visual information retrieval. The approach to simulation described here is an extension to the laboratory model of IR evaluation, permitting also the evaluation of usability aspects of a visualization-based IR-system normally not supported by the original model. The core idea is the modeling of a 3D scene as a linear ranked list, and in addition paying attention to spatial locations of documents. Users are modeled through scenarios of navigation in the scene, with test collection queries that represent user needs.

Keywords

Information retrieval, Visualization, Evaluation, User simulation

1. INTRODUCTION

Within the Uexküll approach [4], a corpus of documents is organized as a multidimensional vector space (with dimensionality K) consisting of spatial representations of all documents and index terms. This is brought about by Multidimensional statisitcs (with LSI as an example). The vector space is rotated [3] so that the coordinate axes (dimensions) are interpretable. Simply stated, interpretability means that documents and terms that have high valued coordinates along an axis, are pertinent to this axis, and the higher the value the greater the pertinence. Such a vector space we term *a data organization*.

In a retrieval situation a user is presented with a list of all K dimensions, each represented by a meaningful name. These can be seen as constituting concepts. The user chooses the three dimensions that best represent his information need. A three dimensional (3D) scene, with the chosen dimensions as axes, is extracted. The scene contains spatial representations (droplets) of the terms and the documents that the system finds relevant to the information need. As for the entire space, each one of the dimensions (axes) in the scene also represents a constituting concept, and, ideally, documents pertinent to a concept will have large coordinate values along the axis representing this concept, the larger the coordinate value, the greater the pertinence. This opens for navigation in the directions of concepts. Uexküll supports two levels of navigation:

level 1 navigating within a scene: moving in the scene in meaningful directions, viewing and selecting documents encountered on ones way,

level 2 changing scenes: selecting different combinations of dimensions, downloading new scenes in which level 1 navigation may be pursued.

Choosing among concepts the names of which are provided by the system, and subsequently navigating in the scene, frees the user from typing keywords that may or may not match the vocabulary of the system.

Our approach, described below, attempts to mimic users' pursuit of these navigation levels so as to measure the potential performance of different data organizations.

As the scope of this paper prohibits detailed presentation of models and results, we refer the interested reader to [4] for results as well as details of measures and models.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.
.

2. EVALUATION BY USER SIMULATION

2.1 The main idea

For the evaluation we are transforming a 3D layout (in which documents do not have any inherent order) that represents a retrieval choice made by a user, into a linear ranked list, where documents' RSVs (retrieval status values) are derived from their extensions along the axes. In line with traditional IR evaluation, we seek to measure the quality of the resulting ranking. *As an extension to the traditional approach we also wish to measure the organization's support for the relevant documents to stand out in the visualization, having useful visual separation from the non-relevant ones.*

Our data organizations are multidimensional transformations, possibly rotated, of a term document matrix. The quality of an organization depends on the different steps in the transformation: the decomposition (the statistical reduction methods), the dimensionality of the resulting space, and the rotation algorithm applied. Different combinations of these components will render organizations with various properties. The simulation based evaluation attempts to find out the following about each data organization:

- to what extent it makes relevant documents appear further out on axes than non-relevant documents;
- how clearly it separates relevant documents from non-relevant ones, so that the relevant documents are not shaded or obscured.

The main idea here is the evaluation of downloaded scenes as *best match* systems, using a test collection with a set of queries, each having a "recall base" of documents judged relevant. Here, traditional measures of partial match IR play

a central role. Into this evaluation we wish to incorporate some aspects of the user interaction, doing so through specially devised measures.

2.2 Simulating users through scenarios

As already indicated, our simulation approach is meant to extend the scope of the laboratory model, characterizing some of the parameters affecting the interaction with a visualization based retrieval system. The user simulations are based on letting topics from the test collection mimic "information needs", and retrieving documents in pursuit of such a need. *User simulation* refers to the following process:

1. Representing user needs by queries (topics). Each query has a recall base.
2. Decomposing the query into its constituent terms and finding the *centroid* of all vectors representing the terms in the multidimensional data organization.
3. Generating a 3D representation that draws on the three axes on which the centroid has the largest coordinate values.
4. Using a model to transform this 3D representation into a unidimensional configuration that mimics a ranked list of retrieved documents.
5. Evaluating this ranked list against the recall base of the query, using both traditional and novel measures.
6. Repeatedly substituting axes in scenes to represent each of the terms of the query - mimic level 2 navigation (see Section 1).

We developed two scenarios: *simulation scenario 1* is simple, mimicing an instantaneous interaction with a retrieved scene, corresponding to steps 1 - 5 above, and *simulation scenario 2* which is more elaborate, incorporating also step 6, thereby mimicing a process where the user navigates along axes and shifts scene.

2.3 Models and measures

To pursue step 4 above, two different models for transforming the scene into a ranked list were attempted [4]. Such a model provides us with a *combined axis* that represents all three axes in the scene, along which each document is represented by a *loading* (coordinate value). For step 5, we used traditional ranked list measures as well as the specially devised *Separation Rewarded Exposure* (SRE) and *Separated Rewarded Precision* (SRP). Both are summary measures calculated for each relevant judged document, and averaged to obtain a query score. Though the detailed mathematics of the measures are beyond the scope of this position paper, a summary is provided below. For each relevant document, r_i, SRE equals an *exposure* component (the fraction of the non-relevant documents along the c*ombined axis* that are ranked lower than r_i) multiplied by a *visual separation reward*, $s_{r_i} = f_{r_i}(1 - f_n)$ where f_{r_i} is the relative loading of r_i along the combined axis and f_n is the average loading of the nonrelevant documents. SRP combines, for each relevant judged document, a precision component and a multiplicative *visual separation reward* that, this time, takes *all* documents ranked below r_i (not only nonrelevant as for SRE) into account.

SRE and SRP aim at predicting the actual usability of a data organization that scores high by the ranked list measures above. SRE attempts to model a user in search of very few (possibly known) documents, terminating the search upon finding a relevant document without regarding other (possibly relevant) documents. SRP looks at each relevant document as a part of a ranked list, and rewards "isolation" above neighboring documents, relevant or non-relevant. Inspired by [1, p. 15] The measures are also an attempt at gauging how well the data organizations behind an Uexküll-based system serve two types of users. SRE represents the "casual" user, in need of a factual answer or a single known item, and SRP represents the more "thorough" user, e.g. performing a literature study.

3. CURRENT STATE

As already mentioned, the goal of this paper has not been to present details, and this goes also for results. The effort described here is work in progress. The results so far demonstrate the superiority of rotated organizations in ranking documents correctly along axes, which, being in harmony with what could be expected, serves as a temporary validation of the approach.

The measures described here are still short of emphasizing differences between different data organizations in supporting different kinds of users, so here, additional effort is required regarding measure development.

So far we have been experimenting with dichotomous relevance judgements. Experiments with graded relevance judgments will be a necessary step to see how well the simulation approach brings out differences between highly relevant and moderately relevant documents. Here graded relevance judgments will be invaluable (see e.g. [2]).

4. CONCLUSION

The goal of this paper has been to present a novel evaluation approach to the visualization and navigation within multidimensional data organizations. The evaluation uses a model of a 3D scene, represented as a linear ranked list, combined with user simulations. In this way the evaluation can draw on the traditional laboratory model, augmenting it with some aspects of usability. We have used the evaluation method to demonstrate the superiority of rotated SVD organizations over non rotated in rendering axes interpretable for purposes of visualization / navigation, indicating also the effect of dimensionality. Tests with real users will be necessary for a comprehensive evaluation of an Uexküll-based system. We believe, however, that the presented evaluation approach may be handy in rendering user tests of such systems (and possibly also system based on similar approaches) less resource demanding and more focused.

5. REFERENCES

[1] A. C. Foskett. *The Subject Approach to Information.* Library Association Publishing, London, 1996.

[2] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[3] K. V. Mardia, J. T. Kent, and J. Bibby. *Multivariate analysis.* Academic Press, London, 1979.

[4] M. Preminger. *The Uexküll Approach: Evaluation of Multivariate Data Organizations for Support of Visual Information Retrieval.* Doctoral dissertation, University of Tampere, Tampere, Finland, 2008.

Automatic Evaluation of User Adaptive Interfaces for Information Organization and Exploration

Sebastian Stober & Andreas Nürnberger
Data & Knowledge Engineering Group, Faculty of Computer Science
Otto-von-Guericke-University Magdeburg, Germany
{sebastian.stober, andreas.nuernberger}@ovgu.de

ABSTRACT

Visualization by projection or automatic structuring is one means to ease access to document collections, be it for exploration or organization. Of even greater help would be a presentation that adapts to the user's individual way of structuring, which would be intuitively understandable. Meanwhile, several approaches have been proposed that try to support a user in this interactive organization and retrieval task. However, the evaluation of such approaches is still cumbersome and is usually done by expensive user studies. Therefore, we propose a framework for evaluation that simulates different kinds of structuring behavior of users, in order to evaluate the quality of the underlying adaptation algorithms.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Human Factors

Keywords

User Adaptivity, Evaluation, Information Organization, Information Retrieval, Information Exploration

1. INTRODUCTION

In many domains users are interested in information that they cannot clearly specify, e.g. by some keywords, and therefore they would prefer to use methods that support interactive organization or exploration of the information collections of interest. For example, a journalist might be researching the background of his current article or someone might look for new music that fits his taste. In such everyday exploratory search scenarios, usually large data collections are accessed. Visualization by projection, e.g. using dimensionality reduction methods, or – flat or hierarchical – structuring are means to ease these tasks. These methods are especially beneficial, if the visualization reflects the user's personal interests and thus is more intuitively understandable.

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

However, except for the user's own collections, personalized structuring is not generally available. Meanwhile, several approaches have been proposed to tackle this problem, see for example [1, 3, 4, 5]. Unfortunately, the evaluation of the performance of the different methods is still a quite challenging task, which is usually done by time consuming and expensive user studies. Besides the problems of designing and evaluating a user study appropriately, it is difficult to compare the outcome of different studies for different interfaces, since very often the goals of the study, the study design and the selected user groups differ quite significantly.

In the following, we propose a framework to tackle these evaluation tasks at least for a subset of interfaces: We focus on approaches that combine the visualization of the aggregated or pre-structured content of a collection with a method to interactively change the class/tag assignment and/or the location of the object in the visual interface.

2. SCENARIO

Assume we have a user-adaptive retrieval system that is able to automatically structure a document collection, e.g., using a self-organizing map or a (hierarchical) clustering algorithm. The task of the user is to organize the collection such that it reflects his personal organization criteria. Therefore, first an initially unpersonalized structure is created which provides an overview of the collection. The user can then interact with this representation, e.g., search and access objects and move objects in the visualization that he feels should be located elsewhere. From this user interaction, an individual representation should be learned – e.g., by metric learning or learning of a simple feature weighting scheme for similarity computation [2, 3] – that represents the user's organization preferences. The process is assumed to be continuous, i.e. the collection visualization is iteratively adapted until it meets the user's structuring preferences.

2.1 Evaluation Approach

In order to evaluate such a setting, we have to simulate the way a user is selecting and moving objects. Furthermore, we have to define a target structure or projection – the "ground truth" – that we would like to obtain. One way to do so is to use a structure or projection that would be automatically derived by a (hierarchical or flat) clustering or projection process and to add "noise" features to the objects that disturb the structuring algorithm[1]. As a result of the added

[1] We have to ensure that the "noise" features induce a random structure in the collection that is not ignored as noise by the adaptation algorithm. See [6] for detailed discussion.

noise, the structuring of the objects differs from the ground truth which represents the simulated user's point of view. Changing the amount of noise (e.g. the number of noise features) controls the difference between the initial and the desired structuring and thus the difficulty of the adaptation task. The adaptation process can be described as follows:

```
modify objects by adding random features
compute visualization on modified objects
repeat
  select an object o to be moved
  select new position c (place or cluster)
  move o to c
until o needs or could not be moved
```

Iteratively, the user selects an object and moves it to the best position according to his ground truth similarity measure (i.e. ignoring the artificial noise features). The adaptation algorithm then updates the feature weights. This process is repeated until the selected object could not be moved because it is already at the desired position or due to possible limitations of the adaptation algorithm. Ideally, this should force the adaption process to finally ignore the artificially added noise features. This could later be analyzed, together with the quality of the finally obtained visualization, for which the one obtained without the noise features is the "gold standard". What remains to be done is to make sure that the simulated user behaves like a real user. Since we usually have to consider different kinds of user behavior, we have to simulate different prototypical users by changing selection and moving strategies.

2.2 Example

In the following, we assume a classification scenario, e.g., objects are assigned to categories (cells) obtained by a self-organizing map, which is interactively visualized. Different selection strategies can be applied for user specific selection of changing the assignment of objects to clusters or tags:

1. Greedy selection of cell and object: First, the cell with the lowest average pairwise (ground truth) similarity of the contained objects is chosen for further investigation. Within this cell, the object with the lowest average pairwise (ground truth) similarity with all other objects in the same cell is selected to be moved.

2. Greedy selection of cell and random selection of object: The cell is chosen as in the previous scenario. However, an arbitrary object is selected from this cell.

3. Random selection of cell and greedy selection of object: Here, the cell is chosen randomly whereas the object to be moved is selected from this cell by the greedy selection approach used in scenario 1.

4. Random selection of cell and random selection of object: In this scenario, both, the cell and the object to be moved from the cell, are selected randomly.

Note that scenario 3 appears to be the one that comes closest to the real use case where a user does not look into all cells before picking an object to be moved but within a specific cell tends to select the object that fits least into the cell according to his preferences. An overview of the different selection strategies is given in Table 1. We successfully performed a first experimental evaluation of such a strategy for the evaluation of an interactive retrieval system that makes use of a growing self-organizing map for structuring a text document collection [6]. A second evaluation was done for a prototypical music retrieval system [7]. Both studies could prove the usefulness of the proposed approach.

Table 1: Overview of cell selection strategies.

		cell selection	
		greedy	random
object selection	greedy	scenario 1	scenario 3
	random	scenario 2	scenario 4

3. DISCUSSION

We proposed a framework for evaluation of user adaptive systems for information organization that simulates different kind of structuring behavior of users, in order to evaluate the quality of the adaption algorithm. Even though the proposed framework is quite general, it leaves a lot room for improvement. For example, the so far proposed user strategies for object selection are only object for crisp object classification and do not yet explicitly consider visualizations where the distances between objects in the projection are relevant.

4. REFERENCES

[1] K. Bade, J. Garbers, S. Stober, F. Wiering, and A. Nürnberger. Supporting folk-song research by automatic metric learning and ranking. In *Proc. of the 10th Int. Conf. on Music Information Retrieval (ISMIR'09)*, pages 741–746, 2009.

[2] K. Bade and A. Nürnberger. Personalized structuring of retrieved items (position paper). In *Workshop Personal Information Management (part of SIGIR'06)*, pages 56–59, 2006.

[3] D. Da Costa and G. Venturini. A visual and interactive data exploration method for large data sets and clustering. In *Proc. of Advanced Data Mining and Applications (ADMA'07)*, pages 553–561. Springer, 2007.

[4] P. Fischer and A. Nürnberger. myCOMAND automotive user interface: Personalized interaction with multimedia content based on fuzzy preference modeling. In *Proc. of 9th Int. Conf. on User Modeling, Adaptation and Personalization (UMAP'10)*, pages 315–326. Springer, 2010.

[5] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *Proc. of Neural Information Processing Systems (NIPS'06)*, pages 417–424, 2006.

[6] S. Stober and A. Nürnberger. User modelling for interactive user-adaptive collection structuring. In *Proc. of 5th Int. Workshop on Adaptive Multimedia Retrieval (AMR'07)*, volume 4918 of *LNCS*, pages 95–108. Springer, 2007.

[7] S. Stober and A. Nürnberger. Towards user-adaptive structuring and organization of music collections. In *Proc. of 6th Int. Workshop on Adaptive Multimedia Retrieval (AMR'08)*, volume 5811 of *LNCS*, pages 53–65. Springer, 2008.

Author Index

Albakour, M-Dyaa 23
Alonso, Omar 25
Arvola, Paavo 27

Beresi, Ulises Cerviño 17, 23

Carterette, Ben 13
Chappell, Timothy 9
Clough, Paul 13, 19
Cole, Michael J. 1

De Rijke, Maarten 21
De Roeck, Anne 23

Fasli, Maria 23

Geva, Shlomo 9
Gonzalo, Julio 19

Hofmann, Katja 21
Hou, Yuexian 17
Huurnink, Bouke 21

Jethani, Chandra Prakash 11
Jones, Gareth 5
Järvelin, Kalervo 15, 29

Kando, Noriko 3
Kanoulas, Evangelos 13
Karlgren, Jussi 19
Kato, Tsuneaki 3
Kekäläinen, Jaana 27
Keskustalo, Heikki 15, 29
Kim, Yunhyong 23
Kruschwitz, Udo 23

Lawless, Séamus 5
Li, Wei 5

Matsushita, Mitsunori 3
Mulwa, Catherine 5

Nanas, Nikolaos 23
Nuernberger, Andreas 33

Pedersen, Jan 25
Pirkola, Ari 15
Preminger, Michael 31

Sanderson, Mark 13
Smucker, Mark 11
Song, Dawei 17, 23
Stober, Sebastian 33

Tunkelang, Daniel 7

Zhang, Peng 17

ISBN 978-90-814485-3-6