Proceedings of the SIGIR 2011 Workshop on

"entertain me"

Supporting Complex Search Tasks

Nicholas J. Belkin, Charles L. A. Clarke, Ning Gao,
Jaap Kamps, Jussi Karlgren (editors)

July 28, 2011
Beijing, China

Proceedings of the SIGIR 2011 Workshop on "entertain me" : Supporting Complex Search Tasks. Held in Beijing, China. July 28, 2011.
Published by: IR Publications, Amsterdam. ISBN 978-90-814485-0-5.

Preface

These proceedings contain the research and position papers of the *SIGIR 2011 Workshop on "entertain me" : Supporting Complex Search Tasks*, held in Beijing, China, on July 28, 2011. The workshop consisted of three main parts:

- First, a keynote by Jussi Karlgren that helps frame the problems, and outline potential solutions.
- Second, paper sessions with eleven papers selected by the program committee from twelve submissions (a 92% acceptance rate). Each paper was reviewed by at least two members of the program committee.
- Third, break out sessions on different aspect of supporting complex search tasks 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 made the workshop a success.

We like to thank the ACM and the SIGIR for hosting the workshop, and Luo Si and Noriko Kando for their 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 2011

Nick Belkin
Charles Clarke
Ning Gao
Jaap Kamps
Jussi Karlgren

Organization

Program Chairs

Nick Belkin	Rutgers University
Charles Clarke	University of Waterloo
Ning Gao	Peking University
Jaap Kamps	University of Amsterdam
Jussi Karlgren	SICS

Program Committee

Nick Belkin	Rutgers University
Charles Clarke	University of Waterloo
Ning Gao	Peking University
Gene Golovchinsky	FX Palo Alto Laboratory
Donna Harman	NIST
Jaap Kamps	University of Amsterdam
Noriko Kando	National Institute of Informatics
Evangelos Kanoulas	University of Sheffield
Jussi Karlgren	SICS
Gabriella Kazai	Microsoft Research
Diane Kelly	University of North Carolina
Mounia Lalmas	Yahoo! Research
Birger Larsen	Royal School of Library and Information Science
Ian Ruthven	University of Strathclyde
Falk Scholer	RMIT University
Mark Smucker	University of Waterloo
Anastasios Tombros	Queen Mary University of London
Daniel Tunkelang	LinkedIn
Ryen White	Microsoft Research
Max Wilson	Swansea University

Table of Contents

The Use Case Perspective for Single Query Information Access

Jussi Karlgren
SICS & Gavagai
Stockholm

ABSTRACT

The "entertain me!" workshop is intended to discuss information access for a complex task based on a single query. Such scenarios may occur for many reasons — a framework for a systematic discussion of differences and likenesses based on the notion of a use case is proposed.

Categories and Subject Descriptors

H.5 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces—*benchmarking, evaluation*; H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Search process, Selection process*

Keywords

Use cases, evaluation, validation, benchmarking, information access

1. "ENTERTAIN ME!" — AN EXAMPLE OF SINGLE QUERY INTERACTION

The most obviously interesting aspect of the topic of this workshop is its example of a single query being the nexus of a complex information access task. The simple request for entertainment is the proxy for a complex information need, one which is likely to require domain and task knowledge, awareness of various contextual constraints, knowledge about the user and the user community, and reasoning capabilities with explanatory power.

Discussing this single query allows generalisations to other complex access tasks and usage scenarios — and at the discussions of this present workshop we should try to keep in mind what sorts of family likenesses we are talking about, which parameters of variation we are moving along, and which we are attempting to keep constant.

entertainme 2011 July 28, 2011, Beijing, China

2. A FAMILY OF SCENARIOS

There may be many reasons for a single simple query being the most appropriate initiating action from the perspective of the user. Some examples might be:

lack of knowledge
: Users may not know the domain of inquiry well enough but is seeking enlightenment. If a user learns enough from the first request, the interaction is likely to evolve into a different type of interaction.

lack of commitment or investment will
: Users may not be committed to working towards a successful resolution of a session but is willing to give a system a try, or users may have little energy or attention to devote to formulate queries in view of other constraints on their momentary context.

lack of specificity
: Users may not have a specific need in mind but is willing to indicate readiness to receive some entertaining or diverting material.

lack of bandwidth
: Users may not have access to a high-throughput communication device and interaction is constrained to a substandard keyboard, a slow connexion or high cost. The system will be required to infer some information to enhance the informativeness of the query.

These different interaction situations are likely to require different designs for interaction and different requirements on the information provided by the system. In some of the cases under consideration in this workshop, we are considering cases where the system can provide *short-coding* of user input, where a less knowledgeable or less committed user can reduce their input to the system to acknowledging or rejecting system suggestions. In others, a system geared towards a success metric such as high recall or a system which provides results by facet or aspect analysis might be the most appropriate design.

We should at this workshop try to keep systematic differences between usage scenarios in mind — they will have effects on the solutions we will be discussing!

3. USE CASES — A FRAMEWORK FOR THINKING ABOUT USER-ORIENTED SERVICES

A *use case* is a relatively informal description of system behaviour and usage, which is designed to show how a *system* is

used by *actors* – stakeholders, consumers, other systems who act outside the system being described and which provides some value for the user[4, 5, 3, 8]. A use case is intended to capture all the ways a system is used by its environment, to describe all the services it offers and the entire relevant behaviour of the system and the actors engage in for some specific purpose of value for the actors. The use case is a tool for developing a system, and user actions as formalised in the use case — most often using UML, the Unified Modeling Language — are mapped onto system components and system development objects for the purposes of system development and evaluation.

Scenarios, which often are the inspiration for use cases, are not use cases but *instances* of them: often several scenarios are necessary to track the various paths through a given use case for a system. A scenario describes the actions of a user during the course of an interaction. For instance, one scenario based on the use case `search for a restaurant in a city of interest` for a image search engine could be a description of Marco typing names of foods and cuisines he knows into the query field of a web search interface at a public location in Canton to find a noodle restaurant in the vicinity.

While the notion of a use case has not been explored to any great extent in information access research[1], there is an implicit notion of retrieval being a topical and task-based activity for focussed, active, and well-spoken users. This implicit use case informs both evaluation and design of systems: recall and precision can be worked together to become a fair proxy for user satisfaction in that usage scenario, even when abstracted to be a relation between query and document rather than between need and fulfilling that need. When information access technology moves from its current prototypical domain of topical text retrieval, the implicit information retrieval use case becomes less useful as a backbone for evaluation.

Recent strands in the study of interactive retrieval have begun to move beyond the modelling of sessions as simple retrieval of items from a collection, emphasizing the importance of modelling context beyond the query itself in understanding the goals of the user (e.g. [6]) and during the course of the European CHORUS coordination action a number of Europe-wide and national research projects on information access were polled for their respective view of future usage of the technology solutions they proposed. The responses were aggregated and collated in terms of a *use case space* with the purpose of improving project-to-project cooperation. [2, 1, 7]

Use cases show promise to be a helpful tool to parametrise differences and likenesses between information access scenarios of various types, allowing the information retrieval research field to provide evaluation and benchmarking mechanisms for situations which are similar but not identical to previously known application scenarios.

4. USE CASE MODELS FOR FAMILY LIKENESS

What parameters of variation should we assume cut across the scenarios we will be discussing at this workshop? What distinguishes the scenarios we are discussing from others? How can we provide a framework from which we can generalise the results from our deliberations? Use cases are one potential vehicle to conduct this discussion with — but as they are not intended for this purpose, we will need to provide enhancements to them for this purpose.

[1]The term "use case" is frequently used in papers on information access technology, but usually it is used to refer to informal descriptions of how useful a certain system component might be.

5. REFERENCES

[1] R. Bardeli, N. Boujemaa, R. Compañó, C. Dosch, J. Geurts, H. Gouraud, A. Joly, J. Karlgren, P. King, J. Köhler, Y. Kompatsiaris, J.-Y. LeMoine, R. Ortgies, J.-C. Point, B. Rotenberg, Å. Rudström, O. Schreer, N. Sebe, and C. Snoek. CHORUS deliverable 2.2: Second report - identification of multi-disciplinary key issues for gap analysis toward eu multimedia search engines roadmap. November 2008.

[2] N. Boujemaa, R. Compañó, C. Dosch, J. Geurst, Y. Kompatsiaris, J. Karlgren, P. King, J. Köhler, J.-Y. LeMoine, R. Ortgies, J.-C. Point, B. Rotenberg, Å. Rudström, and N. Sebe. CHORUS deliverable 2.1: State of the art on multimedia search engines. November 2007.

[3] A. Cockburn. *Agile software development*. Addison-Wesley, 2002.

[4] I. Jacobson. Object-oriented development in an industrial environment. *Proceedings of OOPSLA '87: Sigplan Notices*, 22(12):183–191, 1987.

[5] I. Jacobson, M. Christerson, P. Jonsson, and G. Övergaard. *Object-Oriented Software Engineering: A Use Case Driven Approach*. Addison-Wesley, 1992.

[6] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2), 2009.

[7] P. King and Y. Kompatsiaris. Towards a use case ontology for multimedia information retrieval. 2008.

[8] G. Övergaard and K. Palmqvist. *Use cases - Patterns and blueprints*. Addison-Wesley, 2004.

Applying Metasearch Technique to Medical Literature Retrieval for Evidence-Based Medicine

Sungbin Choi
College of Medicine
Seoul National University
Seoul, South Korea
wakeup06@empal.com

Borim Ryu
College of Medicine
Seoul National University
Seoul, South Korea
borim@snu.ac.kr

Sooyoung Yoo
Seoul National University
Bundang Hospital
Gyeonggi-do, South Korea
yoosoo0@snu.ac.kr

Jinwook Choi
College of Medicine
Seoul National University
Seoul, South Korea
jinchoi@snu.ac.kr

ABSTRACT
Evidence-based medicine has been a highly emphasized concept in the medical domain. To facilitate clinicians' practice of evidence-based health care, current best evidence, which is relevant to the clinical question and also have methodologically high quality, should easily be found. We hypothesized that by counting these two different aspects in ranking algorithm, search engine can automatically retrieve articles which are relevant to clinical question; and also have valid evidence. We approached this problem with combining methodologies. After working out document's query-relevance score and methodological quality score respectively, we combined them using various metasearch methods. For correct evaluation, we built a test collection utilizing preexisting reliable database; Cochrane Reviews.

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

General Terms
Experimentation, Algorithms, Performance

Keywords
Evidence-Based Medicine, Ranking, Classification

1. Introduction
EBM has been widely recognized as an important concept in medical domain. Evidence-based health care is the conscientious use of current best evidence in making decisions about the care of individual patients or the delivery of health services. Current best evidence is up-to-date information from relevant, valid research about the effects of different forms of health care.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing, China.

However, practicing EBM in daily clinical care might be challenging though, considering clinician's time scarcity and inadequate search skills. EBM entails appraising step, critically evaluating article's evidence to decide if it is reliable and robust. Searching for relevant article, plus assessing validity of them, must be a complex search task.

We hypothesized that by counting these two different aspects in ranking algorithm, search engine can automatically retrieve articles which is relevant to clinical question; and also has valid evidence. We approached this problem as an information retrieval task with two distinct priorities, finding enough research articles relevant to the clinician's question, and also valid from the perspective of EBM principled methodological criteria. Using various metasearch algorithms, we combined relevancy and methodological quality scores into single ranking.

In this paper, firstly we built test collection using preexisting sources. Secondly, we used a probabilistic retrieval model and a machine learning classifier to work out a document's query relevancy score and a quality score respectively. Finally, we applied various metasearch techniques to rerank documents. Experimental results show that there are significant improvements over baseline (Ranked by quality score only) with our reranking process.

2. Method
2.1 Test collection
We utilized *Cochrane Reviews* to make our test collection. *Cochrane Reviews* publish systematic reviews of primary research in human health care and health policy. On each review article, objectives are described explicitly, for example, "To assess the effects of donepezil in people with mild cognitive impairment but no diagnosis of dementia". Reviewers, who are domain experts, perform a comprehensive search to find all potentially relevant articles for given topic. When reviewing these retrieved articles, they assess methodological quality for each article, excluding studies not satisfying their predefined criteria, to draw sound conclusion.

We utilized 2009 MEDLINE® /PubMed® Journal Citations, having 17 million MEDLINE documents, as our corpus.

Cochrane Reviews' topic (e.g. "Donepezil for mile cognitive impairment") was adopted as a search query. Reference lists included in the review were taken as gold standard for each query. On average, there were 11 target documents for each query. We prepared 145 queries, 100 queries randomly assigned to the training set, remaining 45 queries to held-out test set.

2.2 Design of our ranking strategy

2.2.1 Overall strategy

We organized our ranking strategy as a 3 step process. Our overall strategy is illustrated in Figure 1.

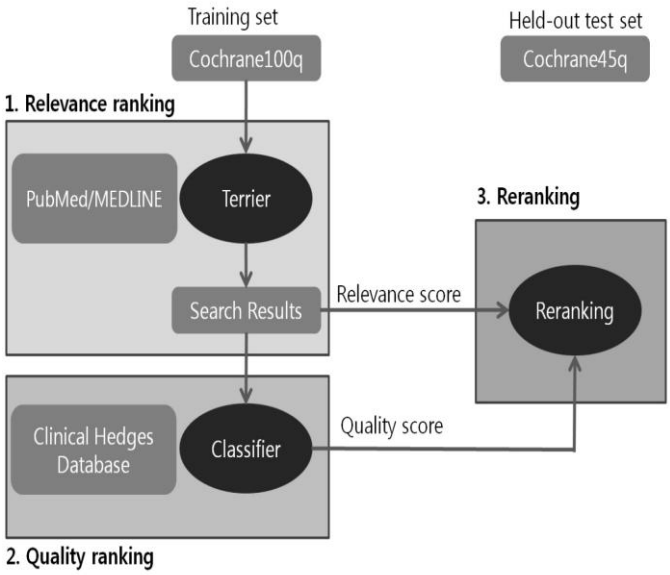

Figure 1. Overall strategy in this study

Firstly, we worked out relevance score using probabilistic retrieval model (Okapi BM25). Title, abstract, Medical Subject Headings (MeSH), publication type fields were extracted indexed.

Secondly, we used machine learning classifier (Naive Bayes, SVM) trained on Clinical Hedges Database, to compute quality score (The value of decision function was used as quality score). We generated various sets of models by trying different classifier and parameter combinations. We tried to find the best classifier model.

Finally, we combined relevance score and quality score using various metasearch methods. Mean Average Precision (MAP) was chosen as our evaluation metric.

2.2.2 Reranking

With relevance score and quality score computed, we combined those two scores with various reranking methodologies.

We used a number of simple combination methods referring to [1], and SVM^{rank} [2], which used SVM algorithms for prediction of rankings.

3. Results

Results on held-out test sets are summarized in Table 1.

Borda-fuse, Weighted-Borda-fuse, Multiplicative combination, Weighted multiplicative combination showed significant increase in MAP (p-value < 0.01) compared to Baseline. Weighted linear combination and SVM^{rank} also showed some improvements (p-value < 0.05).

Table 1. Reranked results on held-out test set

Reranking method	MAP %
Relevance ranking	7.4
Quality ranking (Baseline)	8.2
Linear Combination	13.0
Multiplicative Combination	16.4
Borda-fuse	19.6
Weighted Linear Combination	14.7
Weighted Multiplicative Combination	16.0
Weighted Borda Fuse	16.0
SVM^{rank}	14.5

4. Conclusion

Confronted with difficulty of complex search task in medical domain, we tried to build effective search system, by counting relevance and quality aspects together in the ranking algorithm. Reranking process improved search performance impressively. We hope to make further progress in the future study.

5. ACKNOWLEDGMENTS

Use of the Clinical Hedges database was made possible through a collaboration agreement with R B Haynes and N L Wilczynski at McMaster University, Hamilton, Ontario Canada. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) (No. 2009-0075089).

6. REFERENCES

[1] E.A. Fox and J.A. Shaw. Combination of multiple searches. Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215, pp. 243-252, 1994.

[2] T. Joachims. Training Linear SVMs in Linear Time, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006.

Affective Classification of Large Scale Broadcast Archives

Sam Davies
BBC Research & Development
BBC R&D South Lab
London, UK
00 44 3040 9702
sam.davies@bbc.co.uk

Denise Bland
BBC Research & Development
BBC R&D South Lab
London, UK
00 44 3040 9803
denise.bland@bbc.co.uk

ABSTRACT

In this paper, we present an overview of our framework system for the affective classification of a large scale broadcast archive. Using a combination of video and audio processing we classify programmes according to their affective content, resulting in a mood vector for each programme. This is displayed on a two-dimensional graph, allowing users to select programmes based on mood. We also present an overview of our work on automatic event detection with the initial aim of identifying which sections of a sports match are of interest and ranks these by way of overall interest. This paper forms an overview of the British Broadcasting Corporation (BBC) Research and Development (R&D) department's work on automatically classifying TV programmes for entertainment re-use.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]**:** Abstracting Methods

General Terms

Algorithms, Management, Experimentation.

Keywords

Multimodal, feature extraction, semantic metadata, classification, retrieval.

1. INTRODUCTION

The amount of audiovisual material available to viewers in a digital format is growing rapidly as broadcast transmission capabilities increase and archived content is digitised. With this, it is important that users are able to find not only the media they want but also the segments of media they want. As such, metadata is required for each media asset to allow for inter and intra document searching. Within the (BBC), all media assets that are likely to be reused have manually created metadata, contents of which range from brief synopses to detailed shot and topic listing. However, this is a time and resource expensive process. On average a detailed analysis of a 30 minute programme takes a professional archivist around 8 to 9 hours.

The BBC uses a system called LONdon CLASSification (LONCLASS), an in-house developed extension to the Dewey Decimal System for this classification. Programmes will also have an entry in the BBC's INFAX database, which contains all available metadata about a programme such as LONCLASS number, synopsis, and programme credits. Using this, factual aspects of a programme are classified and then easily found. Currently, BBC Information and Archives (I&A), the section of the BBC which is responsible for the archive, periodically release digitized collections of archived factual programmes themed around specific subjects. However, these are based on expanded INFAX entries. Within these collections, only those parts of programmes that are of relevance and interest are provided. As such, the synopses are all that is required for navigation and selection by viewers. This labour intensive process requires teams of professional archivists to search, tag and segment the archives by hand. As more content is digitised this manual segmentation and metadata generation will become less feasible meaning automation of this process will become more important. Currently only a small fraction of the archives are digitised but this is rapidly growing due to digitisation projects such as [1]. However the main purpose of manually generated metadata is for professional reuse. Frame accurate metadata is designed to allow users such producers and researchers to find stock shots, interviews or other precise sections. However there is currently limited provision for classifying programmes according to entertainment value – if a user doesn't want to find out about a particular event or person, if they just want to be entertained. Our Multimedia Classification project aims to automatically generate metadata that will allow for retrieval of content from broadcast archives when the user wants to be entertained.

Various UK broadcasters now offer 'catch-up' services, such as the BBC's iPlayer, with an industry wide move to integrate these with traditional television set top boxes . This ability to download and view vast amounts of content again presents a requirement for the ability to find and watch desired content. As viewers using these services have the ability to skip forward through the programme, identifying which parts of it are interesting would be of great benefit. This identification would be of even greater use in large scale sporting events, such as world cups or the Olympics, where large numbers of matches or competitions are broadcast either concurrently or in close time proximity. Event detection tools could also be of use in a production environment, allowing for quicker creation of highlights programmes by providing a candidate list of interesting or unusual events during an event.

2. MULTIMEDIA CLASSIFICATION

Our system comprises of three main sections; characteristics extraction, feature extraction and a final machine inference module. We take a multimodal approach analyzing both the audio and the video to create an affective vector for each programme. This is then displayed on a 2D graph with the affective adjective labels, Happy/Exciting and Serious/Lighthearted. These were not chosen as they are diametrically opposed, more that they were initially found to map to extracted features and characteristics. An overview of our system is shown in figure 1.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

2.1 Characteristics and Feature Extraction

The system extracts characteristics from the audio and video signals using signal processing techniques described in [2]. These analyze different temporal and spectral aspects of the audio and video signals. These signals are then used to either identify objects in the audio or video or else in the machine inference modules on their own. Using statistical techniques as described in [2], features such as laughter, motion and shot cut frequency are identified. These are then used in the machine inference module to identify the affective content of the programme.

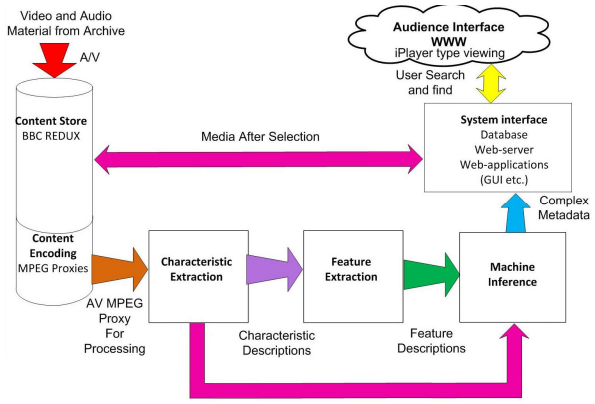

Figure 1. Overview of Multimedia Classification System.

2.2 Machine Inference

This module maps the features and characteristics extracted to the affective scales. Currently this is based on simple observed heuristics described in [2]. Current work is looking to incorporate machine learning methods such Support Vector Machines to increase the accuracy of the system.

3. INTERESTING EVENT TIMELINES

A further area of affective classification we are studying is that of interesting events; any event in a programme which a user may find interesting. In our initial study [3] we examined large scale sporting matches, creating timelines of interest within matches along with an overall 'interestingness' score for a match. Using signal processing of the audio only (to account for the large number of radio only broadcasts made by the BBC of sporting matches), we initially segmented a programme into pitch/studio segments, then analyzed the pitch based segments for interesting events, looking for crowd excitation levels and referee whistles. Events were identified as peaks in these two sonic features. An example of this for one match is shown in figure 2. Ground truth data was taken from the BBC Sports Library, a professional service which identifies interesting events in some modern matches.

4. USER SEARCHING FOR CONTENT

Our current system presents users with content in one of two ways. Our multimedia classification system presents users with programmes arranged on a 2D graph. This shows programmes with similar overall affective content clustered together. This is shown in figure 3. When a user hovers their mouse over the programme marker, programme guide information is displayed.

Using this approach we aim to allow complex searches for content to be broken down into a simple graph.

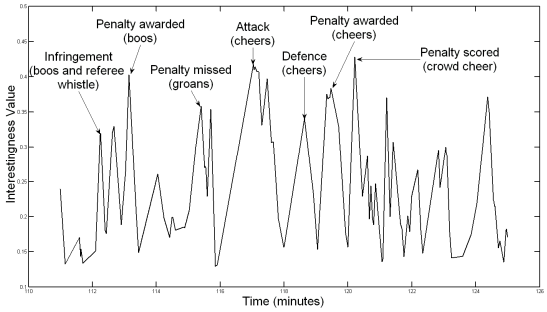

Figure 2. Event detection in sports broadcast.

Users may want to find content via a variety of methods; title, genre or subject or programme mood. Using this approach we allow users to combine these approaches. They can readily identify programme names from the display; genre and subject are contained within the programme guide and programmes with similar moods are clustered together.

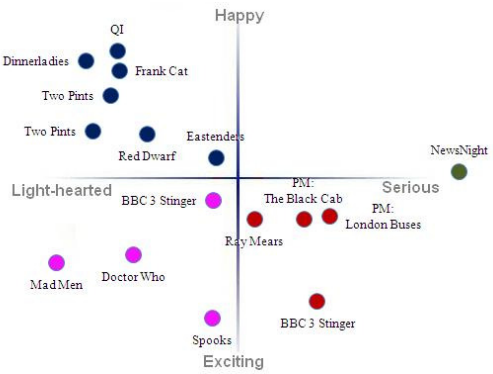

Figure 3. Overview of Multimedia Classification System.

The other approach is to present users with a time line of interesting events in a programme. Users could then select a programme they want using the system in figure 3, choosing only the sections that are of interest.

BBC R&D are investigating these multimodal searches to help user access the content in the BBC archive which they don't know exists. Current methods rely on the user having some idea about what type of content they like. We are trying to solve the issue of when a user doesn't really know what they want. The systems we are developing will allow users to search and browse based on traditional techniques such as keywords or name, but also on new techniques such as affect – or a combination of both.

5. REFERENCES

[1] S. Cunningham and P. de Nier, "File-based Production: Making It Work In Practice," in *Proceedings of the International Broadcasting Convention*, Amsterdam, NL, 2007.

[2] S. Davies, D. Bland, and R. Grafton, "A Framework for Automatic Mood Classification of TV Programmes," presented at the *5th International Conference on Semantic and Digital Media Technologies* Saarbrucken, Germany, 2010.

[3] S. Davies and D. Bland, "Interestingness Detection in Sports Audio Broadcasts," presented at The Ninth International Conference on Machine Learning and Applications (ICMLA 2010), Washington, DC, USA, 2010.

Systematic Reviews: A Complex Search Episode for Evidence Based Policy and Practice

Sarvnaz Karimi
NICTA and the University of Melbourne
Department of CSSE
Parkville, VIC, Australia
skarimi@unimelb.edu.au

Falk Scholer
RMIT University
School of Computer Science and IT
Melbourne, VIC, Australia
falk.scholer@rmit.edu.au

ABSTRACT

Evidence based policy and practice – a paradigm that aims to ensure that decisions are based on consideration of research evidence that meets a high standard – began in the field of medicine, but is becoming widely used in other fields such as economic policy, education, and software engineering. Systematic reviews, the core tools of this evidence based approach, require stringent searching to identify sources of evidence that should inform a decision. We outline the systematic review process, an example of a *complex search episode*, and describe some of the challenges facing information retrieval in this domain.

Categories and Subject Descriptors

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

1. INTRODUCTION

Evidence based practice refers to the use of rigorous evidence, supported by systematic empirical research, to guide decisions. The paradigm was first developed in the field of medicine, aiming to ensure that medical decisions take the best available external evidence into account, rather than resting primarily on the basis of opinions and personal clinical experience [4]. The evidence is focused on rigorous, statistically significant results that are typically the outcomes of randomized controlled trials. Since being embraced in the medical field, the evidence based paradigm has been extended to many other areas of decision-making, from government policy, to software engineering, and product design.

The key tool used in evidence based policy and practice is the *systematic review*, a document which synthesizes available research on the topic of investigation. While most research work involves some sort of literature survey, a distinguishing feature of the systematic review is that it is carried out to agreed standards: using clear protocols in carrying out the process; focusing on specific questions; identifying as much of the relevant literature as possible; critically appraising the quality of the research included in the review; synthesizing research findings from included studies; being as objective as possible to remove bias; and, updating the review so that it remains relevant [1].

A key part of the systematic review, therefore, is the identification of the related literature. Indeed, achieving the objective of an unbiased synthesis of current evidence assumes that *all* relevant related work is identified and considered. In information retrieval terms, therefore, the systematic review process can be characterized as a *recall oriented* task, with the aim of finding all relevant documents that support the current review's underlying question.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

2. SEARCH FOR SYSTEMATIC REVIEWS

We illustrate the challenges in conducing search for systematic reviews in the domain of evidence based medicine as an example of a complex search episode.

A focused research question is first specified by the researchers. An example is: *"Exercise in prevention and treatment of anxiety and depression among children and young people".*[1] Together with the research question, detailed inclusion and exclusion criteria are also specified. For the previous example, these are summarized as: *"Randomized trials of vigorous exercise interventions for children and young people up to the age of 20, with outcome measures for depression and anxiety".* However, we note that as part of the reported search strategy, the criteria are actually fully specified under four different headings: types of studies, participants, interventions and outcome measures.

The search process can then be viewed as consisting of three broad steps:

1. Search experts (e.g., health librarians) formulate complex Boolean queries – also known as *search strategies* – which are run over biomedical databases such as PubMed. The output is a large pool of document summaries consisting of titles, abstracts and authors.
2. The set of *summaries* is scanned by the investigators to identify a short-list of candidate documents that meet the systematic review inclusion criteria.
3. The investigators examine the *full text* of the articles in the short-list, and identify the final set of documents that will be included in the systematic review.

Each step of the process reduces the size of the candidate set drastically. For example, the MEDLINE bibliographic database of life sciences and biomedical information currently indexes around 20 million citations. The search strategy from Step 1 is typically formulated to retrieve a result set ranging from several hundred to a few thousand candidate documents. Triage based on summaries in Step 2 of the process reduced this candidate set to a few hundred items. The review of full text items in Step 3 the leads to

[1] `http://www2.cochrane.org/reviews/en/ab004691.html`

final included papers, typically from ten to a hundred documents [2].

3. COMPLEXITIES IN SYSTEMATIC REVIEWING

The primary *search* complexity in identifying papers that need to be included in a systematic review arises from the specific details of the information being sought. To effectively identify answer documents, it is for example necessary to understand the relationship between various entities in the query (in the example, this might include that patients are suffering from the specified condition, and that the condition could involve anxiety and depression, but only one is a necessary criterion for relevance), as well as the search context (for example, the fact that studies on older people should be excluded, and that only studies reporting specific outcome measures should be considered).

Currently, support for the multiple criteria which need to be considered in order to determine whether a document is likely to be relevant consists of the development of complex Boolean queries in Step 1 of the outlined process. These queries are often of the order of a hundred lines in length, and can take many weeks to develop. In the example systematic review on exercise, the search strategy involved Boolean queries over 7 biomedical databases, and the complex queries ranged from 37 to 79 lines in length. Moreover, these queries include the use of advanced operators for partial string matching, query expansion based on medical subject categories, and the complex manual combination of sub-sets of search results. Despite this intense human-driven effort, it is clear from current practice that specifying inclusion and exclusion criteria is insufficient: human intervention is needed at several steps of the process to remove many thousands of non-relevant items from the candidate set.

A further challenge is presented by the implicit assumption that the initial search strategy identifies all possibly relevant documents. This is fundamental to the evidence based paradigm, which posits that all high-quality evidence needs to be considered. The search task is therefore inherently recall focused: the cost of missing a relevant piece of evidence is high, potentially calling the findings of the final systematic review – a document that may take from 6 months to 2 years to produce – into question. Although the search strategies in Step 1 are typically developed by experts who are familiar with the domain in which the systematic review is being undertaken, it is still likely that some potentially relevant documents may be missed. The problem is further compounded by the fact that the reported search strategies sometimes contain errors, and on re-execution on the same document collection it often transpires that certain included documents in Step 3 are not in the candidate list from Step 1 [3].

Data complexities also exist; the key factor contributing to the difficulty of systematic review search episodes here is that the source collection to be searched over is often not in the form of full-text documents. For example, in PubMed – the most widely-used database for medical systematic reviews – only about 1 million of the 20 million indexed MEDLINE articles include full text, with the remainder consisting only of abstracts and metadata.

4. SUPPORT SYSTEMATIC REVIEWING

We contend that to effectively support search for complex scenarios such as evidence-based policy and practice, next generation information retrieval systems need to incorporate a range of features and technologies.

- To assist in query formulation for an initial search strategy, retrieval systems should aid the user in identifying relevant *entities*, this will relieve the need for searchers to construct long manual lists of synonyms. While attempts at synonym expansion using biomedical dictionaries or taxonomies (e.g., MeSH) are common, the naming conventions should be resolved with reference to the *current collection* that is being searched.
- Automated assistance in formulating the *relationships* between identified entities should be available, such that these accurately and directly map to the inclusion criteria. This is vital in reducing the complex re-combination of answer subsets that is currently required in the Boolean approach.
- While selecting individual documents for further consideration in each of the search steps, automated support for *consistency* is vital. If a reviewer selects one document, but later chooses to ignore a similar one, the system should flag this possible inconsistency.
- A dynamic *relevance feedback* approach that is active during the document selection process could rank the remaining documents based on estimated importance, assisting assessors in focusing their efforts. Moreover, such an approach might identify *additional* documents that exist in the collection but were missed by the initial search strategy.

While many of these items have been proposed and validated experimentally in isolation, we are unaware of a system that comprehensively includes all of these features.

5. CONCLUSION

Systematic reviews are a key tool for evidence based policy and practice, a decision making paradigm that is becoming increasingly widespread. The cost of producing such reviews is a direct function of the quality of the search used to identify relevant evidence. While there are a number of challenges that need to be resolved to allow the easy formulation of comparative search experiments in this paradigm, we believe that working to resolve these can offer significant benefits for information retrieval in evidence based policy and practice.

Acknowledgments NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence program. Thanks to our collaborators at the Global Evidence Mapping Initiative.

6. REFERENCES

[1] A. Boaz, D. Ashby, and K. Young. Systematic reviews: What have they got to offer evidence based policy and practice? *ESRC UK Centre for Evidence Based Policy and Practice*, 2002.

[2] A. Cohen, W. Hersh, K. Peterson, and P. Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *JAMIA*, 13(2):206–219, 2006.

[3] S. Golder, Y. Loke, and H. M. McIntosh. Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *J. Clin. Epidemiol.*, 61(5):440–448, 2008.

[4] D.L. Sackett, W. Rosenberg, JA Gray, R.B. Haynes, and W.S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71, 1996.

Articulating Information Needs by User Profile Enrichment

Chen Chen*, Tiejun Zhao*, Muyun Yang*, Sheng Li*, Haoliang Qi+

Harbin Institute of Technology*
Harbin, 150001, P.R.China
{chenchen, tjzhao, ymy, lisheng}@mtlab.hit.edu.cn

Heilongjiang Institute of Technology+
Harbin, 150050, P.R.China
haoliang.qi@gmail.com

ABSTRACT

In this paper, we tentatively study methods of articulating information needs by user profile enrichment. The main idea is to exploit click-through data of the similar users or the same query to enrich the current user profile. The experimental results show that the user profile enrichment can produce much better search results. We also find that user profile enrichment first emphasizes the topical relevance, and then considers user relevance.

Categories and Subject Descriptors

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

General Terms

Measurement, Experimentation, Performance

Keywords

Information Needs, User Profile, Query Logs

1. INTRODUCTION

One of the fundamental problems of information retrieval (IR) is to search for documents that satisfy a user's information need. Often a query is too short to describe the specific information need clearly [1]. If the search engines is able to effectively identify the user interests and build a profile for every individual user (i.e. user profile), search engine can articulate information needs and thus improve user experiences.

User profiles can be distinguished by the timeframe of their construction and usage. Two of the most typical are short-term profiles, which model immediate information needs, and long-term profiles, which attempt to model user general interests and preferences. Long-term search history is unlimited in time scope and may include all search activities in the past.

Although the long-term search history seems promising, the full potential of long-term search history cannot be reached easily. This is because long-term search history inevitably involves a lot of noisy information that is irrelevant to the current search. Only searches that are related to the current one should be considered as useful context. To address this problem, some studies on long-term search history apply similarity between past queries and the current query [2], or introduce the notion of tasks and match the current user information need with the relevant past user tasks [6].

However, their approaches require enough relevant queries or tasks in user's search history. Moreover, a user issues a variety of queries to the Web search engine and most of them are fresh queries [2], and it is hard for user profile to improve their retrieval performance. With respect to these issues, one natural idea is to mine the similar users' profile from query logs and enrich the current user's profile.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

2. METHODS

We use the Kullback-Leibler (KL) divergence method [3] as our retrieval method. We follow Tan et al. [2] and investigate the smoothing strategy for combining user profile with the query itself. Given user profile P, we can estimate query model as

$$p(w|\theta_k) = (1-\lambda_q)p(w|Q) + \lambda_q p(w|P) \quad (1)$$

Where $p(w|Q)$ is the language model estimated using query text only, $p(w|P)$ is a user profile (language model) learned from long-term search history, and λ_q is the interpolation weight. We estimate the document model using Bayesian smoothing with Dirichlet prior [4]. The user profile is based on the content of previously clicked pages (also known as click-through data).

The relevance can be categorized into two classes: topical relevance and user relevance. We will examine their effectiveness to user profile enrichment. User relevance takes the individual user's information into account. If there is small account of user's search history, we can find the most similar user and put his user profile into the current user's profile. This kind of enrichment is called *similar-user enrichment*. A document is topically relevant to a query if it is on the same topic. Generally, the clicked pages for the same query are topical relevant to this query. The enrichment with the click through for the same query makes user profile topical relevant. We call this method *same-query enrichment*. In order to consider both of topical relevance and user relevance, we propose a kind of *mixture enrichment*. We first find all users who issued query Q_k and select the most similar user to the current user among them. The clicked pages by the most similar user for query Q_k are put into the current user's profile. If the most similar user issuing the same query is not found, all click-through of the same query will be added into the current user's profile.

The above enrichment approaches need to be dealt with how to find the similar user, we propose resource allocation algorithms on a weighted User-Query-URL tripartite network from query log. Initially, we assign User u with resource r indicating the user search interests behind him. Then the resource-allocation is processed which consists of four steps. First, the resource in user nodes (Initially only u has the resource.) is distributed proportionally according to the link weights to their neighbor query nodes. Second, the resource in query nodes is proportionally distributed to their neighbor URL nodes. Third, analogously the resource in URL nodes is distributed back to query nodes. Fourth, the resource is transferred back to the user nodes from query nodes. The final resource located in the user nodes denotes the distribution of interests and performance behind the user u and thus indicates the similarity strength of each user for u. It is easy to find clicked pages of an issued query by tripartite network.

Once the augmented user profile is obtained, the following formula is used to calculate new user profile.

$$p(w \mid P) = \frac{\sum_{Q_i \in E \cup H_k} \lambda_i p(w \mid \theta_i)}{\sum_{Q_i \in E \cup H_k} \lambda_i} \quad (2)$$

where E and H_k is augmented user profile and the current user's profile. $p(w \mid \theta_i)$ is the probability of term w in the clicked page θ_i .

3. EXPERIMENTS

In order to create a document collection, we download all content of clicked web pages in the Sogou query logs [7] during March 2007, resulting in a collection of 1,503,150 documents (after badly-formed URLs, binary data, non-existant web pages, etc. had been removed from the collection), equating to about 10GB of data. We index all the downloaded pages using the indri search engine [5]. Preprocessing includes content extraction from web pages and Chinese word segmentation with Bigram.

To create tripartite graph efficiently, we process query logs as follows. First, we look at a large sample of users issuing more than ten unique queries to ensure enough data for modeling user profile. Second, the last query issued by each user and its clicked pages are moved out of the sample as candidate logs and the rest (i.e. training logs) is used to produce the tripartite graph.

Two types of queries are distinguished due to their different property and retrieval performance. If a query has occurred before in the search history (exactly matching or keywords' order changing), it belongs to the category of recurring queries. Otherwise, we call the query fresh.

From candidate logs, we extract queries that at least ten unique people issued in training data to ensure that (1) we have sufficient relevant judgments to evaluate personalized information retrieval and (2) we need enough topical relevant profile to examine its effectiveness. We call these extracted logs as testing logs. We select 100 users and their 100 fresh queries randomly from the testing logs and each user has one query. Table 1 shows statistics of training logs, candidate logs and testing logs.

Following [2], we utilize click through data and the cosine weighting for all experiments. Sogou query logs include URLs of clicked pages when issuing a query by a user. For a query issued by a user, the clicked pages are regarded as relevant and others are irrelevant. So the two-level relevance judgment is used to calculate all evaluation metrics. Precision@10 is usually evaluation metric to indicate the performance of a few top-ranked results while MAP and NDCG@100 measure the performance of many retrieval results. The interpolation weight λ_q is set to 0.1.

Significance is tested using a 2-tailed paired t-test on 100 queries with the click-though approach as baseline.

Table 2 presents the effectiveness of different user profile enrichment. User profile enrichment by the similar user does not improve retrieval. The possible reason is that although this approach puts more profile into the current user's profile, it just brings noisy information which is not topical relevant to the current query. Enrichment by the same query and the mixture significantly outperform others. These two methods first guarantee that the added profile is topical relevant by extracting relevant profile to the current query. The mixture is a little bit better than the same query one. Since the mixture takes user relevance into account, enrichment can articulate information need better. Thus user profile enrichment first emphasizes the topical relevance, and then considers user relevance.

Table 1. Statistics of the training, candidate and testing data.

	#unique users	#unique query	#unique URLs
Training logs	164,595	1,530,549	4,811,933
Candidate logs	164,595	105,784	122,911
Testing logs	71,001	15,320	34,756

Table 2. Effectiveness of different user profile enrichment.

	NDCG@100	MAP	Precision@10
click-through	0.1276	0.0743	0.0270
similar-user	0.1255	0.0734	0.0260
same-query	0.1365*	0.0828*	0.0270
mixture	0.1378*	0.0832*	0.0290*

4. CONCLUSION AND FUTURE WORK

Long-term user profile is based on the long-term user's search history, which seems to be promising to improve retrieval. However, we investigate Sogou query logs and find fresh queries make up the majority of the last query issued by each user. In this paper, we focus on the fresh query and try to mine user profile from query logs to enrich the current user's profile. We mine user profile from the similar user, the same query and their mixture.

The experiments show that the user profile can articulate information needs over the case without user profile. The user profile enrichment based on the same query is better than on the similar user and the mixture one works best. Meanwhile, we also find that user profile enrichment first emphasizes the topical relevance, and considers user relevance. In the future, we test the sensitivity of methods of user profile enrichment to the interpolation weight.

5. REFERENCES

[1] B.Jansen, A.Spink and T.Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. Information processing & management (2000) 36: 207-227.

[2] B.Tan, X.Shen and C.Zhai. Mining long-term search history to improve search accuracy. In Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining. 2006, 718-723.

[3] C.Zhai and J.Lafferty. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of the tenth international conference on information and knowledge management. 2001, 403-410.

[4] C.Zhai and J.Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) (2004) 22: 179-214.

[5] Indri. http://www.lemurproject.org/indri/.

[6] J.Luxenburger, S.Elbassuoni and G.Weikum. Matching task profiles and user needs in personalized web search. In Proceeding of the 17th acm conference on information and knowledge management. 2008, 689-698.

[7] Sogou Query Logs. http://www.sogou.com/labs/dl/q.html.

Show and Tell: Supporting Children's Search by Interactively Creating Stories

Andreas Lingnau, Ian Ruthven and Monica Landoni
Department of Computer and Information Sciences
University of Strathclyde
United Kingdom
andreas.lingnau,ian.ruthven,monica.landoni@cis.strath.ac.uk

ABSTRACT
In this paper, we describe the Show and Tell system for childrens' interactive search. This encourages children to conduct searches by creating an entertaining digital artifact.

Categories and Subject Descriptors
H3.3 Information search and retrieval

General Terms
Human Factors

Keywords
Children, interfaces, complexity, search

1. INTRODUCTION
Online information is now a standard component in most childrens' information worlds. Children are encouraged to use the Internet for education, have specialized online resources created for their entertainment and increasingly have Digital Libraries created specifically for their use [3]. Most schools, at least in affluent Westernised countries, have computers in the classroom and many nurseries have computers for use by pre-school children.

However, the majority of research on search interface and interaction design has been on software intended for literate, adult users. Whilst this research has led to many successful and popular systems, the increased use of computers by children has focused attention on information access tools for younger computer users, e.g [1, 2, 3]. Studies of children's search behavior and interaction styles, notably those by Bilal et al. [1,2], Druin et al. [3, 6] and Large et al. [5] have shown that there are differences in how children interact with information systems and that these differences can be exploited to provide child-appropriate information systems.

However, what these studies have also shown is that, beyond a few basic design principles, we don't yet know what are appropriate interface models for childrens' search systems. The response by most system developers to childrens' design needs is often to simplify content, to add visual content or to simplify the interaction to a few basic interactions. This approach sees children as simple versions of adults rather than responding to the specific needs of children using search systems [7].

SIGIR Workshop on "entertain me: Supporting Complex Search Tasks, July 28, 2011, Beijing

2. DESIGNING FOR CHILDREN
In this paper we focus on young children around the ages of 6-9. These children are developing cognitive and computer skills, are developing their vocabulary, their ability to read information and are learning to interact cooperatively. This group of children as information seekers faces three core problems:

1. Young children often struggle with the complexity of information seeking. Children do engage in complex thinking about searching [7] but can struggle in creating appropriate strategies to perform complex searches. This is particularly true for actions such as querying; although children like to issue queries and have many definitional search requests they can have problems with creating queries and are less able to generate a good search request [5].
2. Children can also struggle with complex information displays and are more susceptible to lose their way in interfaces with too many special features [4]. So although we want features that engage children in their natural interaction we also want the system to help children structure their information search and provide external motivation for completing a search.
3. A particular feature of children's information seeking is that they often engage in non-linear information search behavior [1], following interesting information rather than information that is useful for completing a task. This is not an issue if the search is simply for pleasure; in other settings where there is often a particular defined task (e.g. writing a report for school) then search systems should help the child keep focus.

The system we describe in this poster is an attempt to help children with complex parts of searching (in particular query creation and reformulation and task structure) through interface design.

3. SHOW AND TELL
In Figures 1 and 2 we present a prototype called Show and Tell (SAT). In SAT the child *shows* an object to the system and the system responds by *telling* the child something about the object.

SAT operates a book metaphor, a familiar concept for children. The child initiates a search by showing the system an object which they want to learn more about. This decision is based on many scenarios we have encountered in our work with school-age and nursery-age children in which children either present an object to an adult to initiate a discussion or in which children are given objects (or object representations such as images) to

learn more about. The latter we found common in school projects.

In this version of SAT the child gives an image of the object to start the interaction. This image may be from an existing source, such as a website that they have found interesting, an image that an adult or friend has given them, or may be from a digital camera, e.g. as the result of a school or family trip. A second version of SAT, under development, uses GPS information associated with images as additional sources of information for images from children's cameras.

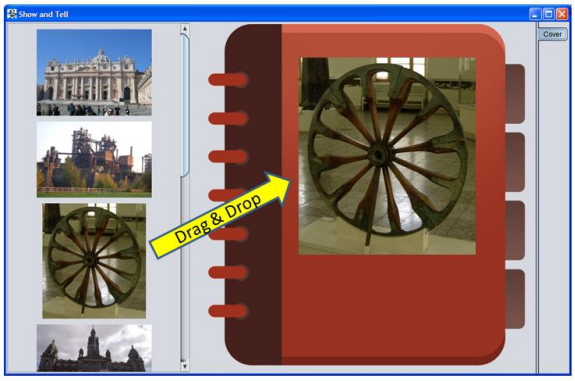

Figure 1: Show and Tell initial interface

The child's image becomes the front cover of the book and the focus for the searching task, Figure 1. Using an online tagging service the image is tagged with simple concepts which are used as a query to initiate searches on various search engines, returning a mixture of child-appropriate text, images and video.

On opening the book SAT provides a selection of these search results on the left hand page, Figure 2. The right hand pages of the book are where the child selects those objects to create their own story: using drag and drop the child can move the useful result to their own page. Text can be read aloud by SAT if the child clicks on the speaker icon.

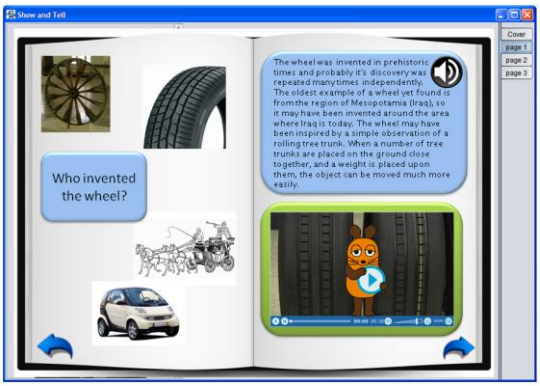

Figure 2: Show and Tell story-building interface

The default book has 3 pages (a variable parameter which can be adapted for older children who may be tackling more challenging tasks) which the child should complete to finish a story. As the child fills each page SAT uses selections from the previous pages to select new results, using a form of relevance feedback to modify the query and information on the types of media selected to determine how many of each type of object to show in following pages. SAT, therefore, adapts subsequent results based on previous interactions.

After 3 pages the child can continue his book by requesting more pages or save his book in a virtual bookshelf so he can continue, update or reference the book later. SAT can be used in two modes: independently by older children or in mediated used with an adult helping the child. With this age range mediated use of systems is often common [7] as is group work within classrooms.

SAT is an attempt to work with what skills children do have – the ability to identify interesting material and connect information through telling stories – and allow the system to make difficult decisions – how to create queries and select what information to show children. The work is an attempt to help children with task structure and the maintenance of a task using a familiar metaphor to children, as they know books have a main topic and consist of a series of pages. For children, books are designed to be entertaining and in SAT the task of searching for information is translated into the task of creating an entertaining object for other people.

4. REFERENCES

[1] Bilal, D. 2001. Children's use of the Yahooligans! Web search engine: II. Cognitive and physical behaviors on research tasks, Journal of the American Society for Information Science and Technology, 52, (2), 118-136.

[2] Bilal, D. and Bachir, I. 2007. Children's interaction with cross-cultural and multilingual digital libraries: I. Understanding interface design representations. Information Processing & Management**,** 43, (1), 47-64.

[3] Hutchinson, H., Bederson, B. B. and Druin, A. 2006. The evolution of the International Children's Digital Library searching and browsing interface, In Proceedings of Interaction Design and Children [IDC'2006], Finland.

[4] Jochmann-Mannak, H. and Lentz, L. 2010. Children searching information on the Internet: Performance on children's interfaces compared to Google, ACM Workshop on Accessible Search Systems at ACM Sigir 2010.

[5] Large, A., Beheshti, J., & Moukad, H. 1999. Information seeking on the Web: Navigational skills of grade-six primary school students. Proceedings of the 62nd ASIS Annual Meeting.

[6] Reuter, K. and Druin, A. 2004. Bringing together children and books: An initial descriptive study of children's book searching and selection behavior in a digital library. In Proceedings of American Society for Information Science and Technology Conference (ASIST).

[7] Spink, Amanda H. and Danby, Susan J. and Mallan, Kerry M. and Butler, Carly. 2010. Exploring young children's web searching and technoliteracy. Journal of Documentation, 66(2). pp. 191-206.

Towards Interactive QA: suggesting refinement for Questions

Yang Tang
Tsinghua University
tangyang9@gmail.com

Fan Bu
Tsinghua University
bufan000@gmail.com

Zhicheng Zheng
Tsinghua University
zhengzc04@gmail.com

Xiaoyan Zhu
Tsinghua University
zxy-dcs@tsinghua.edu.cn

ABSTRACT
The user's intent can be understood better in a question answering system if there are interactions between the user and the system. Consequently, more accurate answers may be served. In this paper, we propose a method to recommend refinement keywords for an input question to facilitate users to make themselves clear in questioning: an important step for interactive question answering when the question is unclear or unspecific. In this paper, we utilize similar questions to explore the refinements of the input question. We show these refinements in rhetorical questions based on Hownet (Chinese WordNet), which help the user identify the unclearness in the question. Experiments show that the precision of recommendation is about 80%.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Question Answering

General Terms
Algorithm, Experiment

Keywords
Question Refinement, Question Answering, Hownet

1. INTRODUCTION
Nowadays, the major way of interacting with the search engines is typing several keywords into the search box. In order to help the user express their inquiry intention more clearly, some researchers proposed to use natural language questions instead of keywords queries[1]. Compared with keywords, natural language question is able to express detailed relationship among the inside words. For example, when a user types *Flight Beijing New York*, it's impossible to know whether he wants to find *the flight from Beijing to New York* or *the flight from New York to Beijing*. If he types the exact question *what is the flight number from New York to Beijing*, then his inquiry intention will be much more clearly. Although natural language question is more powerful at expressing user's query intention, it faces a similar problem with the keywords query: the user may forget to supply some constraint context. For example, when a user asks for the question *what is the best restaurant*, he may forget the location constraint for the restaurant. Without the constraint, the system is not able to give a useful answer to him. In such case, question refinements suggestion is quite helpful.

We define the question refinements task as follows: Given an input question, question refinement suggests a list of keywords which provide more specific or restrictive context for the original question.

Suggesting question refinements faces two main challenges: (i) Question style queries are far less than keywords queries. Question refinement is similar with query refinement[2]. However, since the query refinement algorithms make heavy use of query logs, they are unsuitable to be directly applied to question refinement. (ii) There are too many possible constraint context words. These context words need to be well organized so as to be user friendly.

In this paper, we utilize similar questions to explore refinements and show them in rhetorical questions based on Hownet (Chinese WordNet). First, we retrieve a set of similar questions by searching the initial question in a question set. Then, we extract the refinement words (the refinment words reflect the subtopics of the initial question) of these similar questions and map the refinement words to Hownet. Finally, the refinement words are clustered and shown with the form of rhetorical questions. The experiment shows that the algorithm generates accurate question refinements.

The rest of the paper is organized as follows. We first present our approach to suggest question refinements. Then we show some case studies to evaluate our approach. Finally, we conclude the work by summarizing our contributions and outlining future work.

2. Methodology
Question refinement is to suggest some refinement words to the initial question. Different from query refinement, we can't get sufficient refinement information from the query log. In this paper, we propose to find the refinement words in a set of questions which are similar to the initial question.

We denote the initial question as *Qr*, and it can be represented as a set of words: $Qr = \{w_1, w_2 \ldots w_n\}$. With the vector space model, we retrieve a set of similar questions. We use *V* to denote the word set consisting of the words that come from the similar questions. Notice that we aim to suggest refinement words from *V*, so we filter out those words already existing in *Qr*. However, even after filter, not all of the words in *V* are suitable to be refinement words. There are two kinds of exceptions: (i) the words which are tightly related to the words in *Qr*; (ii) the words which are too common to be a refinement words. We illustrate the two exceptions with an example. A user input the initial question *where is the KFC*. Then we retrieve several similar questions, such as *where is the KFC restaurant in the Beijing Airport*, *where is the KFC restaurant near Beijing Zoo*[1], etc. Besides the words in the initial question, *V* still contains the following words: *restaurant*, *near*, *in*, *Beijing Zoo*, *Beijing Airport*. Among these words, *restaurant* is not a refinement word, since when we mention *KFC*, it implies *KFC restaurant*; *near* and *in are* not refinement words either, since they are both common words. To filter the two kinds of words from *V*, we use two metrics: PMI[3]

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

[1] The questions are Chinese questions in Baidu Zhidao, see in: http://zhidao.baidu.com/question/94389482.html?an=0&si=1 http://zhidao.baidu.com/question/61613739.html?an=0&si=1

and IDF[4]. PMI(w_1, w_2) reflects the relatedness between w_1 and w_2. The higher the PMI is, the stronger the relatedness between the two words is. In order to filter out the words of first exception, we set an upper bound of PMI (denoted as u). For each word v in V, if there is a word w in Qr satisfies that PMI(v, w) > u, then v is excluded from V. IDF(w) reflects the commonness of a word w. The smaller the IDF is, the more common the word is. In order to filter out those common words, we set a lower bound of IDF (denoted as l). For each word v in V, if it satisfies that IDF(v) < l, then v is excluded from V.

After filtering out the words which are not refinement words to the initial question, we get a set of refinement words (denoted as C). To well organize these refinement words, we use ontology to cluster these words. In practice, we handle with Chinese questions, hence we use Hownet as ontology. Similar with Wordnet, Hownet organize the Chinese words in an ontology form. We cluster the refinement words with following steps. (i) For each constraint word c in C, we map it to a node m in Hownet. Since Hownet is ontology, we can find a path from the root of the ontology to the node m. We denote the path as {m_{c1}, m_{c2} … m_{ck}}, where m_{c1} is the root and m_{ck} is the node which c is mapped to. (ii) With the paths, we can build a sub tree of the ontology. For example, assume there are three words in C, denoted as {w_1, w_2, w_3}, and we find three paths: {a, b}, {a, c, d}, {a, c, e}. Then we form a sub tree as shown in Figure 1. Since each node exists at least in one path, each node corresponds to a set of constraint words. In the example, node c is associated with w_1 and w_2. (iii) Scan the nodes in the sub tree in a bottom-up way. If the node contains more than t refinement words, the node is extracted to represent the cluster consisting of the refinement words in it. Then we remove the node and its refinement words from the sub tree. After we remove all the refinement words, we end the clustering process.

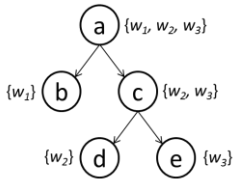

Figure 1. Sub tree of the illustration example.

We can organize these words by clusters instead of simply listing all the refinement words. We design some rhetorical question templates for different clusters. We still take the example *where is the best restaurant*. Different from simply listing all the refinement suggestions, such as *Beijing, New York, Seafood, Steak*, we show the suggestion in a rhetorical question form. With the clustering algorithm, we obtain two clusters from these constraint words. The two clusters are mapped to the nodes *City* and *Food-Type* in Hownet. With manual created templates, we suggest the following text: *Which city it locates? Beijing, New York, etc.* and *Which food type? Seafood, Steak, etc.* In this way, we can help the user identify the missing constraints easier.

3. Experiment

We set up an experiment on a set of Chinese questions. We sampled 100 questions as the initial question set. These questions satisfy the following conditions: (i) The length of the question is less than ten; (ii) We find at least 20 questions which are similar to the question in our question set.

We apply the question refinement suggestion on the initial question set. We get 138 rhetorical question suggestions. Manual judgments show that 111 of them are correct. The precision of question suggestion is 80.4%. We also ask users to manually write question refinements first. Then we compare the system refinement results with manual results. The recall is 56%.

We also show some suggestion examples in Table 1. In the first example, our algorithm generates two rhetorical question suggestions. One implies that the initial question requires constraint of location; the other implies that the initial question requires constraint of the disease being treated in the hospitals. In the second example, the algorithm also generates two rhetorical question suggestions. The first one well implies the requirement for constraint of the place of origin. However, since the cluster is not well explained by Hownet, the second one implies the requirement for constraint of the movie in a weird way.

Table 1. Two Examples of Question Refinement Suggestion

Initial Question	Question Refinement Suggestion
哪家医院比较好 *(Which hospital is better)*	哪个地方(市)?是上海,苏州, …, 还是深圳? (Which city? Shanghai, Suzhou, …, or Shenzhen?) 哪些疾病?是皮肤病,白癜风, ...,还是心脏病? Which kind of disease? Skin disease, vitiligo, ..., or heart disease?
有什么好看的电影推荐 *(Can you recommend some good movies)*	哪个地方?是欧洲,美国, …, 还是香港? (Where? Europe, America or Hongkong?) 什么人?是死人,匪,间谍还是吸血鬼? (Which kind of person? Dead man, robber, spy, or vampire?)

4. Conclusion

In this paper, we utilize similar questions to generate question refinements, and show these refinements in rhetorical questions. We use PMI and IDF to extract refinement words. Then we cluster the refinement words with Hownet. According to the cluster notation given by Hownet, we apply different rhetorical question templates to different clusters of constraint words, and generate different kinds of rhetorical question suggestions. The experiment shows that the algorithm generates accurate suggestion results. However, since Hownet is not able to well explain all the clusters of constraint words, the algorithm sometimes generates some weird suggestions.

The performance of the algorithm is limited by the coverage of Hownet. In the future work, we aim to explore other resources to cluster and explain the constraint words.

Acknowledgement

This work is supported by Canada's IDRC Research Chair in Information Technology program, Project Number: 104519-006.

5. REFERENCES

[1] Voorhees E M. *The TREC-8 Question Answering Track Report*. Proc. of TREC-8, 1999. 77–82.

[2] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. 2010. *Clustering query refinements by user intent.* In Proc. of the WWW. ACM, New York, NY, USA, 841-850.

[3] Manning C D and Sch ¨utze H. *Foundations of Statistical Natural Language Processing.* The MIT PRESS, Cambridge,Massachusets, 1999.

[4] Church K W and Gale W A. *Inverse document frequency: a measure of deviation from Poisson*. In A. et al. (Ed.), NLP using Very Large Corpora. Kluwer Academic Publishers, 1999

Supporting Complex Tasks in a Spoken Language Interface

Xiaojun Yuan
College of Computing and Information
University at Albany, SUNY
Albany, NY 12222, USA
+1 518 591 8746
xyuan@albany.edu

Nicholas J. Belkin
School of Communication & Information
Rutgers University
New Brunswick, NJ 08901, USA
+1 732 932 7500
belkin@rutgers.edu

ABSTRACT
Current search engines do a fine job in assisting users with simple and direct tasks, but need more improvement in coping with difficult user tasks. Users of information systems typically carry out searches with very short queries, on the order of two words or so. This makes it very difficult for the systems to disambiguate their queries and identify potentially relevant documents, and leads to sub-optimal retrieval performance. We hypothesize that users will provide better and more useful descriptions of their information problems if they are able to speak to the system and to easily indicate through speech and gesture, those documents and aspects of documents which they find useful, and not useful. Therefore, spoken language interfaces would be able to better assist users with difficult tasks.

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

General Terms
Design, Human Factors.

Keywords
Searching, spoken language interface, user performance

1. INTRODUCTION
There is no doubt that when dealing with simple and easy tasks, the existing search engines do a fine job. For example, "Where is the capital of China?", users can simply go to a search engine site, and type in "capital China." The answer can be found out from the snippet of the top ranked search results. However, current search engines do not do a good job on the complex situations because of the complexity of human information behavior and needs.

Situation 1 -- "Supporting simple and common requests that express complex and dynamic needs."
Assuming an attendee of SIGIR 2011 would like to find some social events or activities to enjoy in a night of the stay in Beijing, He types in the keywords "Entertain me in Beijing" in his favorite search engine. This task could be complex and challenging because the task itself being not specific but ambiguous and amorphous in goal, the language and culture difference, and required knowledge with China.

Situation 2 -- "Doing a task through a mobile environment." Assuming a SIGIR USA attendee is driving to the airport for SIGIR 2011 conference in Beijing, and needs to find a reasonably rated parking lot near the airport. This task itself is not complex, but the information system needs to provide good support to accommodate the user's information needs in a mobile environment.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

Users of information systems typically carry out searches with very short queries, on the order of two words or so [6]. This makes it very difficult for the systems to disambiguate their queries and identify potentially relevant documents, and leads to sub-optimal retrieval performance. Instead of simply returning a ranked list of documents to respond to this simple query, a better search system or interface is needed to assist users locate needed information in completing complex tasks. This system should assist users during their entire search process and reduce the degree of user perceived task complexity, by iteratively constructing a complex query or search strategy in each searching stage, and by progressively integrating the partial answers into a coherent one at the later stages. To achieve the above-mentioned goals, a spoken language interface which guides users in the user-computer dialogues, and iteratively accepts and aggregates the accumulated query results is necessary and appropriate. More specifically, in response to situation 1, a spoken query interface would allow the user to further extend the original query, and talk more about what the user would like to do to be entertained in Beijing. Through the user's spoken queries, the system would be able to elicit detailed and clear information needs from the user and to produce meaningful retrieval results for the user to choose. In situation 2, a spoken language interface is very important because the user may not be able to type in queries while driving the car. Such an interface would enable the user to articulate spoken queries by talking to the interface, and to respond to retrieval results or reformulate new queries without worrying about typing. Again, the system would respond more satisfactorily by the iteratively collected spoken queries.

In this paper, we propose that a spoken language interface or system that can allow users to talk about their information needs and use gesture to point out what they would like to view is appropriate and effective in supporting difficult tasks with complex and dynamic needs and should be addressed in the related field.

2. PROBLEM BACKGROUND
The tendency of users of information systems to begin their searches with brief queries is probably due to two factors: the general inability of people to specify precisely what documents they require in order to resolve their information problems (cf. [1]); and, the difficulty that people have in finding terms appropriate both for describing their information problems, and matching the terms which have been used to describe the documents in the database with which they are interacting.

To address these two problems, a variety of ways have been proposed and investigated. One approach has been to devise interface techniques which encourage searchers to input longer

queries (e.g. [7]); another has been to automatically enhance the initial query without the searchers' intervention, or through query expansion based on thesauri or similar tools (cf. [5]); a third to offer to searchers, based on their initial queries, terms which could be used to enhance their initial queries (e.g. [2]). Although each of these approaches has been shown to afford some benefit in retrieval effectiveness, none of them has involved searchers in developing and understanding their information problems, finding better ways to express their information "needs", nor succeeded in substantially improving either retrieval effectiveness or searcher satisfaction with the interaction [7].

We propose to address the problem of encouraging effective interaction of the searcher with information systems by moving from keyboard-based interaction to spoken language and gestural interaction of the searcher with the information system.

The origins of this approach are based on Taylor's research on question negotiation between user and librarian in special libraries [10], and on the experience of elicitation of verbal descriptions of searchers' information problems in studies of Anomalous State of Knowledge (ASK)-based information systems (e.g. [1]). Taylor found that, in the types of interactions that he studied, librarians engaged in conversations aimed at eliciting a number of different aspects of the searchers' information problems, and that the searchers were indeed able to address these different aspects. Belkin and his colleagues found that, when suitably prompted, searchers were able to provide search intermediaries with extended verbal descriptions of their information problems. Subsequently, Saracevic et al. [8], in their analysis of searcher and intermediary interaction with information systems, showed that there was substantial direct commentary by both searcher and intermediary on results retrieved with respect to a query put to the system, and, more recently, Crestani & Du [4] have shown that asking for expression of search need in verbal terms results in significantly longer queries than those expressed through a keyboard interface.

Crestani has led a group which has considered spoken language queries and their effectiveness in a variety of contexts [4]. Some of this work has investigated the effectiveness of spoken queries, as well as their length, but in simulated rather than real interaction. Zue, et al. [11]'s work is perhaps the most complete in terms of spoken language query understanding, but it has been applied in limited domains.

The main arguments against taking the spoken language and gesture approach to query input and interaction have been that: there has not been strong evidence that such interaction will actually result in more effective results; it is unclear that searchers will willingly engage in such interaction; and, most importantly, that speech understanding technology is not robust enough to support such interaction. Our position is that: there is some evidence that longer queries and more extensive response to search results that would be afforded by this mode of interaction does improve retrieval effectiveness (e.g. [7]); that when encouraged to describe their information problems more fully, searchers will do so ([3]; [7]); and, that spoken language interaction with information systems appears to be either doable right now [11] or in the very near future, with commercially available speech understanding systems (e.g. Dragon). There is also evidence that speech recognition technology is already in place in a mobile environment [9]. For instance, Google outlined developments in voice search, which allows users to search the Internet from a mobile phone by speaking their requests or queries to Google in Japanese, in addition to Chinese and English. Google is planning to add new languages next year.

3. DISCUSSION AND CONCLUSIONS

How to design a system that can iteratively provide assistance to users in different information-seeking stages during their complex task searching process is an important issue in interactive information retrieval. We believe our proposal makes an important step toward better understanding users' information needs, and investigating different ways to elicit users' information needs and thus in turn improve user performance and satisfaction, as well as reducing the perceived user task complexity.

4. ACKNOWLEDGEMENTS

This research was sponsored by Institute of Museum and Library Services (IMLS) grant RE-04-10-0053-10.

5. REFERENCES

[1] Belkin, N.J. (1980) Anomalous States of Knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, v. 5: 133-143.

[2] Belkin, N.J., Marchetti, P.G. & Cool, C. (1993) BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, vol. 29: 325-344.

[3] Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., Yuan, X.-J. (2003) Query Length in Interactive Information Retrieval. In SIGIR '03. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 205-212). New York: ACM.

[4] Crestani, F. & Du, H. (2006) "Written versus spoken queries: A qualitative and quantitative comparative analysis." Journal of the American Society for Information Science and Technology, 57(7): 881-890.

[5] Efthimiadis, E.N. (1996) Query Expansion. In: Williams, Martha E., ed. Annual Review of Information Systems and Technology, v31, pp 121-187, 1996.

[6] Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management. 36(2), 207-227.

[7] Kelly, D., Dollu, V. J., & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. In Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil, 457-464.

[8] Saracevic, T., Spink, A. & Wu, M-M. (1997). Users and intermediaries in information retrieval: What are they talking about? User modeling. Proceedings of the Sixth International Conference, UM97. New York: Springer, 43-54.

[9] Stone, B. (2009). Google Adds Live Updates to Results. New York Times, December 8 Issue. Retrieved from http://www.nytimes.com/2009/12/08/technology/companies/08google.html

[10] Taylor, R.S. (1968) Question negotiation and information seeking in libraries. College and Research Libraries, v. 29, 178-194.

[11] Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J. & Hetherington, L. (2000). JUPITER: A Telephone-Based Conversational Interface for Weather Information," *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, January 2000.

Searching For Unlawful Carnal Knowledge

Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
Leif.Azzopardi@glasgow.ac.uk

ABSTRACT

Search engines are often used for leisure related search tasks, to find online shops, games, music, movies, celebrity gossip and even sex. While these activities can be broadly considered as entertainment, I shall focus on discussing the different *Sexual Information Needs* (SINs) of users. This unexplored area of Information Retrieval (IR) research considers a variety of search tasks related to sex: from looking for romcoms, to finding a date, to downloading pornography. Here, I outline seven *not-so-deadly* SINs that users try to satisfy on the web. I then discuss how addressing these SINs as part of a response to the query "entertain me" would maximize user satisfaction.

1. INTRODUCTION

According to Van Halen (1984), "everybody wants some, how about you?". While this song and its lyrics are entertaining, it reminds us that sex is an underlying carnal need. As a result, sex is used to grab our attention [4] and often features in entertainment from titillation to stimulation. So if a user were to type in "entertain me" to a search engine or goes online to be entertained, then there is a high likelihood that they would be interested and entertained by something sexy and sex related [1].

When it comes to searching the web, numerous query log studies have shown that query terms related to finding content of an illicit and sexual nature occur with relatively high frequencies. In [2, 3], it was shown around 8-10% of queries were sex related; indicated by query terms such as "sex", "free", "pictures", "naked", "nude", etc. This shows that many web search engine users are interested in being *entertained* by sexual content, *a priori*. The response of the web search engines to such queries is usually web sites that are predominately pornographic in nature and content (i.e. sites that display explicit x-rated multimedia content). However, despite these observations, little research has been conducted that considers these types of sexual information needs[1]. But, with so many searches of this type, it is clear that users are interesting in finding such content, so it is time to abandon the taboo status associated with discussing sex and sex related search topics. And, to consider such needs in a scientific and objective manner within the remit of Information Retrieval. Thus, this paper aims to start the discussion on searching for sex.

[1]However, at the SIGIR 2006 Workshop on evaluating exploratory search systems, Marshall suggested that it was only a matter of time before a "porn" based evaluation track was proposed and run at a forum like TREC. Perhaps, the time is now?

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

2. SEXUAL INFORMATION NEEDS

Typically queries which contain terms indicative of sex or sexuality are considered to have one kind of intent, i.e. to find pornographic material. While this is perhaps the most dominant sexual information need. There are, however, many other types of sexual information needs that users may have - these range from satisfying curiosity, fantasies and romance, to fulfilling basic, carnal desires. Consequently, the classification or rating of such material will range from parental guidance (PG) and general audience (12+) to adult and X-rated (18+). Also, the types of resources required to fulfill the different sexual information needs will vary considerably, from multimedia content (video, picture, audio, dvds), to text (i.e. books, stories, etc), to products and paraphernalia, and invariably to people (either in real life or via live video links). To try and distinguish between the different types of sex based searches, I have formulated a number of different types of sexual information needs that users may have - and then discuss how they related (or not) to being entertained.

Titillation - The suggestion of sex is often too alluring to dismiss and advertisers often take advantage of this desire. Content that hints at sex, beyond advertisements, is usually music and the associated lyrics, and in particular, the related music videos. For example, the Britney Spears video clip, "Hit me baby one more time" is a prime example of sexual innuendo, which resulted in innumerable queries being submitted to search engines so that users could see a scantily clad Spears dressed as a school girl performing in a suggestive manner. Content suggestive of sex may seem harmless, but it is likely to lead to other types of SINs.

Awareness - This SIN stems from a curiosity about finding out about one's own body, about the bodies of the other sex and learning about sex. For many teenagers (and nowadays even younger children) the desire to find out about such things is part of growing up. To satisfy this need, educational content is often created and supplied. It is usually drawn, described and discussed appropriately for the particular age ranges (as supported by sites like `http://www.sexetc.org/` which is a magazine about sex for teenagers) - but sometimes curiosity will lead users to other darker SINs.

Romance - The search for romance is often undertaken by females, though not exclusively, and is generally related to escapism and fantasy (i.e. the need or want of an ideal

love affair or happily ever-after story). The kinds of content which aims to satisfy such a need is usually romance novels (from vendors like `millsandboon.co.uk`) where suggestive proses titillate the reader (i.e. "her loins were burning with desire as she caressed his throbbing member...".). Other types of content that also try to address this need are movies that are of the romantic comedy (or rom-coms) genre. These movies aim to entertain and try to satisfy the needs of both female and male viewers.

Erotica - While erotica is often literature or art that is intended to arouse sexual desire, here we consider erotica in the context of products. Specifically, this SIN relates to the devices and products often used to indulge in some fetish or fantasy and/or to stimulate, arouse or enhance sexual desires and pleasures; and so this need ranges from the desire to feel sexy to increasing the sexual desires through fantasy to being sexually stimulated and gratified through some device. So site selling merchandise such as lingerie and sex toys like the infamous rabbit to costumes and devices available from vendors (see `bravissimo.com`, `lovehoney.com` or `annsummers.com`). Of course, nothing says "entertain me" more than whips and chains.

Love - An increasingly common phenomena is to find a partner online to satisfy the need for love and companionship. So rather than recommend videos or products, the resource required is a service to help users find the love of their lives. Sites like `match.com` and `eharmony.com` enable users, usually singles, to meet others based on their profiles, where they are matched "on the deepest levels of compatibility". Core to these sites are recommendation and matching algorithms to find and narrow down the possible partners to a set of potential or ideal partners. Such sites help fulfill a basic desire of most, i.e. to find love. Though often it is used to have fun on the dating scene (and thus to be entertained), without the connotation of being particularly sleazy, or as direct as the next SIN.

Lust - Like the love SIN, the need of the user, here, is more carnal and the desire is to fulfill their underlying basic needs. Sites like `sexbook.com` and `fbook.com` are specifically dedicated to help users find others to engage in various kinds of activities. These range from sending naked photographs to online sex via a web cam to meeting in real life and participating in sexual acts.

Stimulation - Users wishing to be aroused or stimulated by sexual content fall into this last SIN. Thus, pornographic sites are designed to provide illicit and X-rated content for the pleasure of their adult users (assumed to be 18+, and usually male). Such sites provide hardcore pornography including images and shots of people participating in various sexual activities - usually the participants are semi-naked or naked, and may be wearing various outfits or costumes (i.e. stockings, cowboy hats, boots, masks, etc). And will generally include very explicit and close up shots of genital regions, including penetration shots and money shots. There are a large variety and range of types of hardcore pornography, usually classified at the higher level as straight, gay/lesbian, animal, etc. Then more specifically to describe the particular sexual acts or activities (such as anal, blowjob, handjob, etc) and/or the particular participants (such as amateur, blonde, coed, etc)[2].

[2]For example, see sites like `www.youporn.com` or `www.redtube.com` for detailed classification schemes.

3. SUMMARY AND DISCUSSION

In this poster, I have presented a number of different SINs. These are very real needs, stemming from carnal desires, that are often either implicitly or rather explicitly posed to search engines to satisfy. However, it is clear that a significant amount of further research needs to be conducted to explore this research area in detail. For instance, different users will have various underlying SINs at different times and the level of complexity required to fulfill the different SINs will also vary. For example, sometimes returning an item within one of these broad categories will be enough to entertain, at other times only a very specific item will do. Also, the demographics of users, i.e. their age, gender, mood and sexual preferences is likely to impact on what is relevant and entertaining. So for a query as broad as "entertain me", it is difficult to satisfy all users, but I would argue that providing items that aim to satisfy at least some of these SINs would be a good starting point. For example, returning items like the latest and most popular titillating music video clips and the latest rom-com movies are likely to be entertaining, relevant and acceptable to most users. However, the latter SINS become significantly more complex and challenging to fulfil, i.e. finding the right erotic product, finding the love of your life, or finding the right kinds of stimulation. This is likely to require the recommendation of dedicated search verticals or portals (like the ones previously mentioned), and for users to be more specific about what will entertain them. Other issues that needs to be examined further are the ethical, legal and moral implications of investigating and supporting SINs. However, these issues are largely down to one's personal preferences, the culture within society, and the laws of one's country. But, one thing is for sure, these issues do not stop the supply, nor the demand for items that satisfy these SINs. One issue particularly relevant to IR research is the trade-off between privacy and personalization. Personalization requires tracking the history of interactions of a user, and building up a profile of their likes and dislikes. However, users are often quite sensitive when it comes to their SINs, and would like to avoid any potentially embarrassing situations where the search engine returns or suggests sex related items at an inappropriate time (i.e. when searching in front of others). In conclusion, SINs have been largely ignored by the IR research community, despite the high volume of queries related to some of the more notorious SINs. However, as I have outlined there are a range of SINs, which present a new set of research challenges that are interesting, complex and important[3].

4. REFERENCES

[1] P. Goodson, D. McCormick and A. Evans. Sex and the Internet. *CyberPsychology & Behavior*, 3(2):129-149, April 2000.

[2] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real Life IR. *SIGIR Forum*, 32:5–17, April 1998.

[3] B. J. Jansen, A. Spink, and J. Pedersen. An analysis of multimedia searching on altavista. In *Proc. of the 5th ACM SIGMM*, MIR '03, pages 186–192. ACM, 2003.

[4] T. Reichert. Does sex in advertising work? *Advertising and Society Review*, 8, 2007.

[3]Many thanks to the anonymous reviewers for their expert opinions, helpful comments and positive feedback on this topic.

Why is this restaurant different from all other restaurants? (Captioning for contextual suggestion)

Charles L. A. Clarke William Song
University of Waterloo, Canada

ABSTRACT

In this position paper, we view queries such as "entertain me" as representing an entirely new class of problems, requiring the creation of new information retrieval applications that fall somewhere between traditional recommender systems and traditional search engines. We call these new applications *contextual suggestion* systems. The effectiveness of these systems depends on their ability to exploit context when selecting suggestions, their ability to provide novel suggestions, and their ability to contrast one suggestion against another. In this paper, we outline requirements for contextual suggestion and provide an example directly related to the primary goal of the workshop.

1. CONTEXTUAL SUGGESTION

To answer complex and incomplete queries such as "entertain me", an information retrieval system must take into account the underlying context, including the location, weather, time of day, date, friends, personal taste and many other factors. When considered in a vacuum, the query "entertain me" is nearly meaningless. When considered in the light of person, place, and time, the system might reasonably respond with a rich selection of suggestions, ranging from videos, books, and other solitary pursuits, to restaurants, concerts, and other social activities. Ideally, the system might even offer to invite available and appropriate friends and family.

When presenting its suggestions, the system must avoid both overwhelming and underwhelming the user. Often, users will seek suggestions on mobile devices, where the interface is constrained by both network bandwidth and screen real estate. The system must clearly and concisely communicate its suggestions, allowing the user to retain or reject them through simple interaction mechanisms. As suggestions are reviewed, the system must accommodate this implicit feedback when making further suggestions, providing a continuous stream of novel and interesting ideas.

To answer a query like "entertain me", we imagine a new class of information retrieval applications, which we call *contextual suggestion* systems. The services provided by contextual suggestion systems fall somewhere between those of traditional recommender systems and those of traditional search systems. Unlike traditional recommender systems, the domain is open and the system can suggestion almost anything. Unlike traditional search systems, the information need is poorly specified, with the system depending heavily on context to clarify this need. By writing this position paper, we hope to encourage a research agenda explicitly directed towards contextual suggestion.

A contextual suggestion system must describe a suggestion with a caption[1] that contrasts it against similar suggestions and also reflects its particular appeal to the user. To communicate salient aspects, the system must first determine what makes an suggestion unique (or at least unusual) and if this uniqueness might be of particular interest to its user. For example, when suggesting a restaurant, the system might emphasize elements of the menu or ambience that might be particularly appealing. Later in this paper, we provide an example of how contrastive summarization methods might provide one route to this goal, although unfortunately without considering personalization, which we leave to future work.

2. RELATED IDEAS

Many review sites, such as Yelp and Google Places, incorporate extractive summarization of reviews in their captioning for businesses and other entities, but it is not clear to what extent these sites attempt to identify unique aspects of these entities or to personalize their results. A small but growing body of work explores the generation of contrastive summaries, work which is directly applicable to the problem of creating captions that highlight the unique aspects of entities [3, 6]. Researchers such as Teevan et al. [4] explore methods for personalization, which might be applied to contextual suggestion. Clarke et al. [1] examine the impact of captioning on Web search, demonstrating the importance of clear and useful captions.

3. AN EXAMPLE

As an example, we attempt to answer the question posed in our title by applying a simple contrastive summarization method to a collection of Beijing restaurant reviews taken from the site `localnoodles.com`. Our method is a variant of a simple approach that dates back to the earliest days of information retrieval, but which consistently provides reasonable performance and trivially extends to multi-document summarization by taking steps to minimize redundancy [2, 5]. We stick with a simple approach to meet our

[1]We use the term *captions*, rather than *summaries*, to suggest their lightweight, dynamic, and flexible nature, as well as to reflect the requirement that they include structured information (e.g., addresses and prices).

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

The Saddle Cantina	The menu has American, tex-mex and true Mexican food... * 15 rmb off tacos on Tuesdays * Daily Happy Hour from 6 pm - 8 pm * Cinco de Drinko every 5th day of the month where all drinks (except for bottles) are half priced...
Tube Station Pizza	True, Kro's Nest pizza is ridiculously big... We ordered Garlic Bread, a salad, Pizza(Medium variety),local beer ,onion rings.As were were 3 of us... It was enormous and handle ample crust, cheese, sauce and toppings in the right proportions...
Bookworm	Library: borrow all the books you wish 7... Events: interesting authors, book talks, musical evenings, open mic night, etc... Who else has the International Literary festival... This place is a haven for people watching, and having the world go past you...
Ganges Indian Restaurant	Chicken Tikka Masala,The Butter Chicken and Cottage cheese Spinach curry with Rice and Nan Breads... Go here for the lunch buffet... My staple Indian dishes - Garlic Naan, Lamb Curry- I can't recall the exact name of it, but it is AMAZING and Tandoori Chicken...
Blue Frog	burgers are yummy, especially the blue cheese burger... The food and drinks are a little over priced however the happy hour and two for one burgers on the Monday are good value... The hamburgers are tasty but the fries are almost better...
The Tree	Proper pizza - thin crust, not too many toppings... Amazing wood fired pizzas with a bevy of beers at your beck and call... Their beer menu is longer than their food menu and offers a huge range of Belgian beers, some that you won't find anywhere else in Beijing...

Figure 1: Some suggestions for dinner in Beijing.

requirement for lightweight and dynamic captioning. Our primary innovation is our use of a background model to help identify the unique characteristics of a target entity, such as a restaurant.

We assume the existence of two document collections. The first collection $\mathcal{C}_x$ provides information regarding an entity x, which forms the target of our captioning efforts. For our example, we use a collection of reviews about a specific restaurant. The second collection $\mathcal{C}$ provides information regarding other entities in the same class as x, providing a background model against which we may contrast x. For our example, we use reviews for a wide range of restaurants (including the reviews for x).

From these collections we estimate two probabilities for each term t appearing in the collections:

$$p(t) = \text{probability a document from } \mathcal{C} \text{ contains } t,$$
$$p_x(t) = \text{probability a document from } \mathcal{C}_x \text{ contains } t.$$

For both probabilities, we use maximum likelihood estimates with additive smoothing. From these probabilities we compute a score for each term based on its contribution to K-L divergence, ignoring values below zero.

$$\text{score}(t) = \begin{cases} p_x(t)\log(p_x(t)/p(t)) & \text{if } p_x(t) > p(t), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Rather than the contribution to K-L divergence, which provides the desired contrast, most variants of this approach use $p_x(t)$ only.

We then apply a four-step algorithm to extract sentences from $\mathcal{C}_x$, which together form the caption.

1. Compute an overall score for each sentence s in $\mathcal{C}_x$:
$$\frac{\sum_{t\in s}\text{score}(t)}{length(s)+l},$$
where $l > 0$ is a constant intended to encourage sentences of reasonable length. We use $l = 25$.
2. Add the sentence with the highest score s' to the caption.
3. Set $\text{score}(t) = 0$ for all $t \in s'$, to avoid redundancy.
4. Repeat steps 1-3 until the caption is complete.

Figure 1 provides results for the six restaurants having the most reviews on `localnoodles.com`.

4. CONCLUDING DISCUSSION

Context, contrast, and novelty are the keys to contextual suggestion. When making a suggestion, a system must clearly communicate how it differs from similar suggestions, and how it might have particular appeal to the user. We explore one simple method for lightweight, dynamic and flexible captioning, providing an example directly addressing the primary goal of the workshop. Future work might extend the method to include personalization, develop evaluation methodologies, and adapt other summarization approaches.

5. REFERENCES

[1] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in Web search. In *30th SIGIR*, pages 135–142, 2007.

[2] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[3] M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*, pages 66–76, 2010.

[4] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *28th SIGIR*, pages 449–456, 2005.

[5] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, 43:1606–1618, November 2007.

[6] D. Wang, S. Zhu, T. Li, and Y. Gong. Comparative document summarization via discriminative sentence selection. In *18th CIKM*, pages 1963–1966, 2009.

A Palette Mixing Model of Information Seeking for Complex Queries

Miles Efron, Peter Organisciak
Graduate School of Library & Information Science, University of Illinois
501 E. Daniel St., Champaign, IL, 61820
{mefron, organis2}@illinois.edu@illinois.edu

ABSTRACT
This position paper offers a theoretical approach to considering how information retrieval (IR) systems can support highly contextual queries, such as *entertain me.* We argue that a natural way to pursue this query is by relying on multiple information sources—what we call micro-information interaction services (micros-IIs). In the course of a complex search, people sample from a subset of IR services that they deem relevant. This sampling and combining of services is analogous to the way artists organize and use their palettes. This paper contributes a definition of micro-IIs and an introductory treatment of a model of information seeking that we call *palette mixing.*

Keywords
Information seeking, complex queries, palette mixing model

1. INTRODUCTION
This paper proposes a framework for understanding how people interact with information systems as they pursue a complex query. In this case we focus on a single query, *entertain me.* Specifically, we imagine a use case where a traveller is planning an evening in Beijing and would like his or her evening to be fun.

We argue that a searcher with the query *entertain me* is likely to rely not only on iterative sub-queries to a search engine, but also on multiple, highly specialized *micro-information interaction systems* (micro-IIs). Each of these micro-IIs supports a single implicit or explicit query.

To understand how users employ these micro-II services, we introduce an information seeking model based on the metaphor of an artist's palette. Artists array colors on a palette in ways that are idiosyncratic, expedient, and often geared towards a particular type of painting (i.e. a particular task). The palette mixing model presented here complements established information seeking models to account for contemporary settings where search systems are distributed over multiple "apps" and multiple devices.

2. MICRO-INFORMATION INTERACTION
The *entertain me* query is inseparable from the context in which it is issued. Criteria for results' usefulness would be different if a person is bored at work or making plans from his or her hotel lobby in a foreign city. The complexity of planning an evening out invites us to consider querying as a set of sub-queries such as those in Table 1.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

Table 1. Sample Sub-Queries for *entertain me.*

Where are SIGIR attendees meeting for drinks tonight?
What are good restaurants near my hotel?
I hate the theatre. What else can I do tonight?
What bus do I take to get from here to Chaoyang?

Traditional Web search engines and verticals have a role to play in these queries. But specialized services may be more helpful. Services such as Twitter, Facebook, Yelp, Google Latitude, and Foursquare integrate context and information structure into information interaction in a way that is difficult for a more broadly scoped search engine.

We refer to the act of using specialized services like these as *micro-information interaction*. Whereas standard IR systems field diverse queries, a micro-II system exists to handle a narrowly constrained problem. By virtue of this constraint, micro-IIs are able to (1) capitalize on context, (2) impose intuitive structure on results, and (3) utilize past user patterns in specialized ways.

With respect to context, the simple act of choosing to use a narrowly focused system is informative. Opening Latitude, a location-sharing application, expresses a user's interest in the geographic location of his or her friends at a given time. That is, choosing to use the service implies a type of query.

In addition to the contextual expressiveness of system selection, many micro-II services benefit from device-specific affordances. A person using a location-aware mobile phone can automatically transmit geo-location information, rather than manually specifying it. Affordances such as compasses and cameras inform the query representation in services such as Yelp's Monocle feature and the augmented reality browser Layar.

For complex searches such as *entertain me,* "macro" search engines and verticals can be of help, especially off-line. Bringing many sources to the problem, though, has been shown to be helpful [1]. Given the ubiquitous, lightweight computing increasingly enabled through mobile devices, it is likely that a user will approach *entertain me* with a smattering of focused apps—micro-II services.

3. PALETTE MIXING
Most information seeking models emphasize the temporal dimension of search. These models rightly account for the way queries, and indeed information needs, evolve during the search process. Models such as berry-picking [2] and information foraging [3] take different approaches with respect to the temporal nature of search. But time is central to both of these models.

We suggest another lens for considering information seeking: the artist's palette. The palette is highly personal. Individual artists are known for using idiosyncratic palettes. Additionally, a single artist may create a different palette for paintings of different types (e.g. still life, landscape, figure). In all cases, the artist arrays colors, chosen from a larger collection of paints, spatially. With this arrangement in place, he or she then uses color strategically, drawing on each hue as needed and mixing them to achieve the desired results. If he or she finds that the palette is lacking, he or she supplements it with additional colors.

We argue that using micro-IIs to solve complex problems is analogous to applying colors mixed from a palette. The user chooses services that might be of interest (loading these onto his or her phone, or simply keeping URLs in mind). As sub-problems arise, the searcher turns to the services that will be useful, using combinations of services to solve the problem. A complex query such as *entertain me* is does not necessarily entail a single mode of response. Instead, a user negotiates a variety of specialized response types, from transportation directions to food recommendations, to places best avoided.

To the best of our knowledge, the only prior consideration of information seeking in this vein was proposed by Foster [4]. But in Foster's paper, the palette is ancillary; colors are compared to *activities* such as browsing rather than services such as micro-IIs.

Figures 1 and 2 schematize the metaphor of information interaction as palette mixing. N.B. The three columns in Fig 1 and 2 are not meant to be directly comparable. Fig. 1 shows photos of three artists' palettes[1] downloaded from Flickr, each one unique in in its arrangement and the colors it contains. On Flickr, each photo is described by text articulating artists' motivations for mixing a particular palette.

Fig. 2 demonstrates an ad-hoc pallet mixing approach to a sample scenario: considering things to do from a Beijing hotel lobby. The user (one of the authors) has organized applications in a way that is personally useful. He traverses possible leads between them, as denoted by the yellow arrows. The user organizes found information in the notes of a final app, Evernote, which syncs among his devices (phone, tablet, laptop).

Micro-IIs are effective due to their specialized, niche uses, which are at once strongly tuned to a given task and easily grasped by users. To use a micro-II, a user need not engage in Pallet-mixing (it could be used in isolation). However, with respect to design principles, a micro-II's usefulness can be extended if people can integrate it into diverse interactions with the goal of forming a cohesive understanding of the overarching query.

A pallet mixing-influenced service could be a micro-II in itself, acting as a broker between a user's context and other relevant micro-IIs. This system would specialize in understanding the context of the query, pairing the right services for the user's need and perhaps helping organize the user's preferred results from each service. With respect to IR, *relevance* in this setting would involve presenting a coherent palette of micro-IIs. Here the unit of retrieval would be a micro-II. The result set would be a coherent palette or set of palettes that address the information need. While a query in this context might be as terse as *entertain me*, contextual considerations would be paramount to inducing a viable model of information needs. Such a model would of necessity include contextual cues (implicit and explicit). Affordances of mobile devices would be useful in this regard. But in this brief treatment we remain device agnostic, leaving query design for micro-II retrieval as a future challenge.

Contexts other than a traveller abroad are easy to imagine. A user at home during bad weather could see Internet browsing services such as StumbleUpon, TV schedules, and Netflix instant streaming recommendations. A worker in an office cubicle might appreciate a chain of apps where output from StumbleUpon is piped as a query to youtube and Wikipedia, with their output piped to a final micro-II that arranges the results for browsing.

There is more to consider in the palette mixing model. But we believe that a piecemeal, fragmented information interaction is realistic for a complex, evolving goal and that understanding the creativity that goes into this process offers promise for studying real-world information needs such as *entertain me.*

4. ACKNOWLEDGMENTS

This research was supported in part by a Google academic research award.

5. References

1. Teevan, J., et al. The perfect search engine is not enough: a study of orienteering behavior in directed search. 2004. ACM.

2. Bates, M.J., The design of browsing and berrypicking techniques for the online search interface. Online Information Review, 1993. 13(5): p. 407-424.

3. Pirolli, P. and S. Card. Information foraging in information access environments. 1995. ACM Press/Addison-Wesley Publishing Co.

4. Foster, A., A nonlinear model of information seeking behavior. Journal of the American Society for Information Science and Technology, 2004. 55(3): p. 228-237.

Figure 1. Three Artists' Palettes.

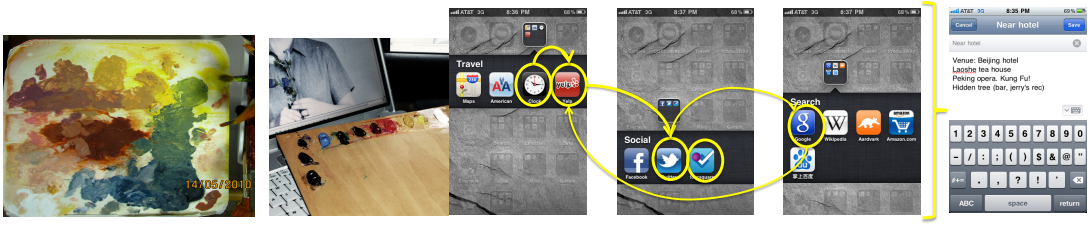

Figure 2. A Sample Micro-Information Interaction.

1. Palette photo URLs, from left to right: http://bit.ly/jDHH8Y, http://bit.ly/inGBWD, http://bit.ly/j8lJt9, all Creative-Commons licensed

How to Evaluate Exploratory User Interfaces?

Tatiana Gossen, Stefan Haun, Andreas Nürnberger
Data & Knowledge Engineering Group, Faculty of Computer Science,
Otto-von-Guericke-University Magdeburg, Germany
{tatiana.gossen,stefan.haun,andreas.nuernberger}@ovgu.de

ABSTRACT

Usability evaluation is an integral part of user interface software development. We discus how to apply existing evaluation methods to exploration tools supporting complex information needs. Evaluation of such complex systems is very challenging and requires collaboration with domain experts for creating scenarios and participation. Furthermore, complex information needs are usually vaguely defined and require much user time to be solved. In order to evaluate these tools more efficiently four components are essential: a standardized evaluation methodology, benchmark data sets, benchmark tasks and clearly defined evaluation measures. As an outlook of this position paper, we propose a method which can serve as a starting point to develop a methodology for evaluation of exploration tools supporting complex information needs.

Keywords

usability evaluation, benchmark scenarios, exploratory search

1. INTRODUCTION

In this paper we discuss issues related to evaluation of exploration tools supporting complex information needs (*CIN-ET*). Our starting point are systems designed for exploration of large, high-dimensional and heterogeneous data sets. The *Jigsaw* [4] system for investigative analysis across collections of text documents, the *Enronic* [7] tool for a graph based information exploration in emails and the *CET* [5] for efficient exploration and analysis of complex graph structures are some examples of exploration tools. The research question which we targeted is how to evaluate such systems.

The most important functionality of exploration tools supporting complex information needs is to support users in the creative discovery of information and relations that were overlooked before in data sets (e.g. document collections). With an evaluation it should be proven that—using the tool—users are able to satisfy their complex information needs effectively, efficiently and with positive attitude.

Evaluation methods which can be used vary and consist of formal usability studies in the form of controlled experiments and longitudinal studies, benchmark evaluation of the underlying algorithms, informal usability testing and large-scale log-based usability testing [6]. There is also some research in the area of automatic evaluation of user interfaces [12]. We consider an automatic approach, but it is not clear if this would work for CIN-ET evaluation.

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

2. EVALUATION CHALLENGES

Since CIN-ETs are complex systems [10], evaluation of them is very challenging. The first challenge is to create an appropriate scenario for evaluation. The tasks must be complex enough to represent a realistic situation. Such realistic exploratory tasks require much time (weeks or even months) to be solved. Lab experiments are limited in time, therefore a "good balance" between time and the right level of complexity is crucial for lab user studies. Longitudinal studies overcome lab experiments drawbacks like strong time limitation and artificial environment. Researchers motivate the community to conduct long-term user studies because they can be well applied for studying the creative activities that users of information visualization systems engage in. [11]

CIN-ETs are often designed to be used by experts with domain-specific knowledge, e.g. molecular biologists, who are more difficult to find than participants without special skills or knowledge. Thus, the second challenge is recruiting the participants. This should be a group of people which represents the end users. It requires either collaboration with scientific institutions or some incentive (like money) to engage their participation [10]. In the study preparation step collaboration with domain experts is also needed to help the researchers in creation of appropriate scenarios.

Controlled lab studies and longitudinal studies require an involvement of target users. The well established usability aspects which are evaluated in these studies, are *effectiveness*, *efficiency* and *satisfaction* [1, 6]. In the context of CIN-ET evaluation, one can express effectiveness in the amount of discovered information, efficiency in time to find new facts or in importance of the made discovery and satisfaction in the user's rating of the tool's comfort and acceptability [3].

3. METHODOLOGICAL SHORTCOMINGS

By evaluating CIN-ETs we can either focus on the tool examination or carry out a comparative evaluation. Most researchers concentrate on evaluating their own tool to gain a deeper understanding of user interactions with it. However, the results do not provide such important information if or under what conditions their tool outperforms alternative tools for the same purpose. We found only one publication [8] that proposed an experimental design and a methodology for a comparative user study of complex systems.

To be able to compare and rank a CIN-ET among similar ones, benchmark data sets and tasks for user studies

are essential [9]. Suppose we wanted to repeat the study conducted in [8] to compare our tool to theirs, we would need the document collection and the task solution used by the authors. However, this data is not available to the public, so we cannot compare the results. A promising direction here is the *Visual Analytics Science and Technology* (VAST) contest[1] which offers data sets of different application domains with description and open-ended domain specific tasks. These tasks should be solved with the help of specific software within the contest. After the contest the solutions are made public, making the data available to evaluations.

Additionally, clearly defined evaluation measures are also important in order to evaluate exploration tools more efficiently. These could be measures from different domains, e.g. information retrieval and human computer interaction, but new measures are still necessary in order to capture the amount of discoveries in document collections or how creative a solution is. The task solution itself can be very complex, so we need a way to account for answers which are only partially correct or complete.

One can draw an analogy between user evaluation of exploration tools and IR automated evaluation of ranking algorithms. The latter requires a set of test queries, a document collection with labels according to relevancies (e.g. TREC) and a measure (e.g. Average Precision) [6], while CIN-ET user evaluation requires a benchmark data set, a benchmark task with a standard solution and an evaluation measure.

4. BENCHMARK EVALUATION

In the following we propose an evaluation method for discovery tools, consisting of two parts: The first part is a "small" controlled experiment with about 5–10 participants. The purpose of this is to collect qualitative data using user observations like audio/video recording and interviewing the participants afterwards. We actually do not need a special task to be solved by the participants. The assignment can be to discover new information using the software. From this study we collect data about learnability improvements and user satisfaction.

The second part is an online study, in which the software is provided to the participants as an online application. The participants can access the tool from their own working environment and spend as much time as they like with the tool, even working discontinuously. After that they can use an online questionnaire to provide the task solution and usability feedback. Participants are motivated to solve a thrilling task using the tool. We assume that the VAST benchmark data with an investigative task (from IEEE VAST 2006 Contest) can be used as a benchmark data set and a benchmark task. The tool interactions of each participant are logged on the server side. We can analyze them to get the time spent by participants to get the solution and interaction patterns. The outcome of the study also contains the number of participants who succeeded in solving the task in comparison to all participants who tried.

The described method is only the first step in the creation of a good methodology. It still has several drawbacks. The first problem is to get an appropriate participants' number. It is not easy to stimulate the participation even with money and if it would work the study becomes cost consuming. One possible solution lies in automatic evaluation (see, e.g., [2]). We could simulate exploration process on different levels and for diverse tasks. However it is not clear how to model a *creative* exploration process, which is important in the case of CIN tasks like creative information discovery. We also do not have a clear understanding how to judge the success of the search given a complex information need. Thus, the question about evaluation measure remains.

[1] http://hcil.cs.umd.edu/localphp/hcil/vast11/

5. CONCLUSIONS AND FUTURE WORK

We proposed a method which can serve as a starting point to develop a methodology for CIN-ET evaluation. However, several aspects are yet unclear. This applies to evaluation methodology, in particular the possibility to evaluate the CIN-ET automatically, and evaluation measures. We would like to motivate the community and make the researchers pay attention to the fact that evaluation of CIN-ETs should be carried out using a standardized evaluation methodology in combination with benchmark data sets, tasks and measures. Only then CIN-ET designers can evaluate their tools more efficiently.

6. REFERENCES

[1] *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs). Part 11 - guidelines for specifying and measuring usability.* 1998.

[2] L. Azzopardi, K. Järvelin, J. Kamps, and M. Smucker. Proc. of SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR (SimInt 2010). ACM Press, 2010.

[3] N. Bevan. Measuring usability as quality of use. *Software Quality Journal*, 4(2):115–130, 1995.

[4] C. Görg and J. Stasko. Jigsaw: investigative analysis on text document collections through visualization. In *DESI II Works.*, 2008.

[5] S. Haun, A. Nürnberger, T. Kötter, K. Thiel, and M. Berthold. CET: a tool for creative exploration of graphs. In *Proc. ECML/PKDD*, pages 587–590, 2010.

[6] M. Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[7] J. Heer. Exploring Enron: Visualizing ANLP results. 2004.

[8] Y. Kang, C. Goerg, and J. Stasko. How can visual analytics assist investigative analysis? Design implications from an evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[9] C. Plaisant. The challenge of information visualization evaluation. In *Proc. of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

[10] J. Redish. Expanding usability testing to evaluate complex systems. *Journal of Usability Studies*, 2(3):102–111, 2007.

[11] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proce. of AVI works. on BEyond time and errors: novel evaluation methods for inf. vis.*, pages 1–7. ACM, 2006.

[12] S. Stober and A. Nürnberger. Automatic evaluation of user adaptive interfaces for information organization and exploration. In *SIGIR Works. on SimInt'10*, pages 33–34, Jul 2010.

Author Index

ISBN 978-90-814485-0-5