SIGIR 2007 Workshop

Web Information Seeking and Interaction

held in conjunction with the 30th Annual International ACM SIGIR Conference

27 July 2007, Amsterdam

Organizers

Kerry Rodden
Google
Ian Ruthven
University of Strathclyde
Ryen W. White
Microsoft

Preface

The World Wide Web has provided access to a diverse range of information sources and systems. People engaging with this rich network of information may need to interact with different technologies, interfaces, and information providers in the course of a single search task. These systems may offer different interaction affordances and require users to adapt their information-seeking strategies. Not only is this challenging for users, but it also presents challenges for the designers of interactive systems, who need to make their own system useful and usable to broad user groups. The popularity of Web browsing and Web search engines has given rise to distinct forms of information-seeking behaviour, and new interaction styles, but we do not yet fully understand these or their implications for the development of new systems.

Web information seeking and interaction (i.e., the interaction of users with Web-based content and applications during information-seeking activities) is a topic that unites many strands of academic and commercial research, from studies of information-seeking behaviour to the design and construction of large-scale interactive systems. Designing components to support this interaction (and evaluating these components) is particularly challenging given the scale of the Web, the diversity of the user population, the diversity in tasks being undertaken, and the dynamic nature of the information.

This workshop is intended to act as a focal point for researchers and practitioners whose work is related to web information seeking and interaction, to enable them to share experiences and collaborate.

The papers selected for this workshop are a mixture of research, discussion and position papers. We have deliberately selected a broad range of papers for this workshop to reflect the diverse research areas that contribute to the discipline of Web Information Seeking and Interaction.

We would like to thank our panellists for providing a stimulating start to our workshop and the programme committee for generously providing comments and guidance to the submitting authors.

Website http://research.microsoft.com/~ryenw/wisi/

Programme

09:00 09:30	**Introduction and Ice-breaker:** Interactive (out-of-your-chair) activity to familiarize participants with the goals of the workshop and each other.
09:30 10:30	**Panel:** "Challenges and Opportunities in Supporting Web Search Interaction"
10:30 11:00	**Coffee break**
11:00 12:30	**Break-out sessions and discussion:** Participants break out into small groups, discuss solutions to problems posed by the workshop organizers and other participants, and report back.
12:30 14:00	**Lunch**
14:00 15:30	**Paper presentations:** *Adaptive Personalization of Web Search* Shady Elbassuoni, Julia Luxenburger, Gerhard Weikum (Max-Planck-Institute of Informatics) *Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages* Kerry Rodden (Google Inc.), Xin Fu (University of North Carolina) *Evaluating Engagement in Interactive Search* Heather O'Brien, Elaine Toms (Dalhousie University) *Clickthrough-based Measures of Search Engine Performance* Erik Graf, Craig Macdonald, Iadh Ounis (University of Glasgow)
15:30 16:00	**Coffee break**
16:00 16:45	**Workshop-wide discussion**
16:45 17:00	**Closing remarks and future directions**

Table of contents

Adaptive Personalization of Web Search

Shady Elbassuoni
Max-Planck Institute of Informatics
Saarbrücken, Germany
elbass@mpi-inf.mpg.de

Julia Luxenburger
Max-Planck Institute of Informatics
Saarbrücken, Germany
julialux@mpi-inf.mpg.de

Gerhard Weikum
Max-Planck Institute of Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

In this paper we present a client-side approach towards personalization of web search which adapts the means of personalization to the user need in place. We differentiate three different search goals: re-finding known information, finding out about topics of user interest, and satisfying an ad-hoc information need. Our approach carefully balances these search modes, which is endorsed by preliminary results of a small-scale user study.

1. INTRODUCTION

An often stated problem in state-of-the-art web search is its lack of user adaptation, as all users are presented with the same search results for a given query string. A user submitting an ambiguous query such as "java" with a strong interest in traveling might appreciate finding pages related to the Indonesian island Java. However, if the same user searched for programming tutorials a few minutes ago, the situation would be completely different, and call for programming-related results. Furthermore suppose our sample user searches for "java hashmap". Again imposing her interest into traveling might this time have the contrary effect and even harm the result quality. Thus the effectiveness of a personalization of web search shows high variance in performance depending on the query, the user and the search context. This coincides with the findings in [4] from a large-scale study on MSN query logs. To this end, carefully choosing the right personalization strategy in a context-sensitive manner is critical for an improvement of search results. In this paper, we present a general framework that dynamically adapts the query-result ranking to the different information needs in order to improve the search experience for the individual user. We distinguish three different search goals, namely whether the user re-searches known information, delves deeper into a topic she is generally interested in, or satisfies an ad-hoc information need. We take a relevance feedback approach in the spirit of Rocchio [11]; however, we vary what constitutes the examples of relevant and irrelevant information according to the user's search mode. This strategy yields an adaptive personalization that exploits context, yet avoids pitfalls of earlier approaches.

The remainder of the paper is structured as follows. Section 2 reviews related work, Section 3 introduces our personalization approach, and preliminary experimental results are shown in Section 4.

1st Workshop on Web Information-Seeking and Interaction '07 Amsterdam, The Netherlands

2. RELATED WORK

There are a number of attempts on personalizing Web search. Due to space limitations we give a high-level categorization of what has been done along with some exemplary references which are, however, not meant to be exhaustive. One way of personalizing search is by means of implicit user relevance feedback. Approaches along these lines include [17, 13] which inspired the work presented in this paper. They achieve personalization by a client-side re-ranking of Web search results based on the previous user search and browse behavior. However, each one tackles a single facet of personalization, either biasing search results to general user-interests [17] or respecting the current search session's context [13], while we unify both aspects and dynamically switch between these two search modes. Our approach towards handling the current session context builds upon ideas in [13], and extends them to the whole user clickstream during a search session.

Another path of addressing personalization is by the categorization of both user interests and search results and a biasing of search results according to some similarity measure on these categories. Approaches along these lines include [9, 3, 18]. Personal biases inside the state-of-the-art link analysis algorithms such as PageRank [2] and HITS [8] provide a further means to shift search results according to user interests. E.g., [6] has been the first to propose biasing the random jumps inside the PageRank algorithm towards pages of user interests. In [10] this idea is further extended by automatically learning topic preferences from the user search behavior. Some personalization techniques not only consider a single user, but also take the actions of a surrounding group of users into account, e.g., [14] follows a collaborative filtering approach.

3. OUR PERSONALIZATION APPROACH

We aim at a holistic approach towards search personalization that supports and balances the different search goals a user might pursue. To this end, we differentiate three major search modes as follows.

Re-finding known information. As motivated in [16], returning to information once successfully found is an important user need. Despite the existence of bookmarking tools that would allow the user to achieve this goal in a direct manner, users quite often prefer to re-search for information by re-submitting a previously issued query [16].

Finding out about topics of user interest. By considering the long-term search and browse history of a user, the main topics of user interest emerge. Whenever a user query is ambiguous or broad in nature, superposing the learnt user interests might serve the user search experience. However, as already found in [15] the benefit of such a approach might differ for *recurring* as compared to *fresh* queries which motivates a differentiated usage of long-term user information.

Serving an ad-hoc information need. Yet even though a user might have strong focus on several topics of interest, she still might switch interests or develop some short-term information needs outside the scope of her interests.

Before we dig into the details of the approaches to each of these personalization facets in Section 3.3, our system architecture, the general framework and underlying retrieval model are introduced.

3.1 Personalized search architecture

As especially the browsing activities beyond search are outside the reach of a search engine, client-side solutions are favorable. Moreover, as all user data is kept locally, user privacy is not violated. We therefore set up a client-side search personalization with the use of a proxy which is running locally. It intercepts all HTTP traffic, extracts queries, *query chains*, i.e., subsequently posed queries, result sets, clicked result pages, as well as the whole *clickstream* of subsequently visited web pages, and stores this information to a local database file which we refer to as the *local index* in the following. Accordingly, searches with Google (the same approach can be easily applied to any other search engine as well) are intercepted and search results are re-ranked according to personal preferences. We preferred a proxy over implementing a plugin for browser-independence. Moreover, the proxy is broader applicable as it may bundle several users and thus achieve biasing of search towards community interests, and at the same time when run locally serve as a pure personalization tool. The proxy we are using relies on the UsaProxy implementation [1] that enhances all html files passed through by some Javascript code that sends logging information on events such as the load of page, mouse movements, etc back to the proxy.

For the following discussion we define our notion of a *search session* which is based on heuristics about the user's timing as well as the relatedness of subsequent users' actions. User actions are (1) queries, (2) result clicks, and (3) other page visits. All successive actions are considered to be within the same search session as long as they are no more than 15 minutes apart from each other or their similarity exceeds a certain predefined threshold. When computing the similarity of subsequent actions, a query is represented by the centroid of the top-50 result snippets.

3.2 Retrieval model

The standard vector space model [12] serves as our retrieval model which represents both queries and documents as a vector of features $\vec{X} = (x_1, x_2, \ldots ...., x_n)$, where n is the number of unique terms in the corpus and x_i is the score of feature i. Terms are weighted according to tf-idf [12]. To overcome the lack of web-corpus statistics that is usually prominent in client-side approaches to personalization, we approximate the global document frequency (df) statistics needed from the documents viewed and queries submitted during the search session. That is, not only each page visited or result viewed will contribute to the statistics, but also each query string, as well as each result item (snippet and title), is considered as a document in the corpus. That way it is not only ensured that each term present in a result item has a non-zero document frequency (as the term might not be present in the local index yet), but also session-biased df statistics are created. These are better suited for measuring the discriminative power of terms in the session context than index-wide statistics would be. Similarly, the features and the document lengths of query results are derived from their snippets and titles, as retrieving their full text would be too time-consuming.

To facilitate personalization of search results we utilize the relevance feedback framework introduced by Rocchio [11]. Thus, we associate with each query a query vector which is initially constructed from the query terms. This query is later augmented with terms that best differentiate relevant documents from non relevant ones. That is,

$$\vec{q_1} = \alpha \vec{q_0} + \frac{\beta}{n_1} \sum_{i=1}^{n_1} \vec{R_i} - \frac{\gamma}{n_2} \sum_{i=1}^{n_2} \vec{S_i}$$

where $\vec{q_0}$ is the original query vector, $\vec{q_1}$ is the refined query vector, $\vec{R_i}$ is the i^{th} relevant document vector, $\vec{S_i}$ is the i^{th} non-relevant document vector, n_1 is the number of relevant documents in the corpus, and n_2 is the number of non-relevant documents in the corpus. The parameters α, β, and γ control the influence of relevant, and non-relevant documents on the refined query vector.

Once we have made the choice of using this relevance feedback model to improve the query representation, the problem dwells down to infering relevant and non-relevant documents with respect to the user need currently in place.

Result re-ranking. Our search agent retrieves more results than the typical user is likely to view (50 results). Whenever a user action allows to update the query representation, unseen results are re-ranked. E.g., this is the case when the user submits a query or when she presses the "Next" link to view more results. Yet we refrain from re-ranking when the user returns to a seen result list using the "Back" button, as we perceived this more as irritating than as advantageous.

Merging of personalized and original results. In order to incorporate the query-independent web page importance, personalized result ranks and original web ranks (as an approximation for the real page rank) are aggregated to form the final result ranking. Inspired by rank aggregation methods for the web presented in [5], we use Borda's method to combine the two result rankings. Thereby each result item is assigned a score corresponding to the number of results ranked below it. Then the total score of a result is a weighted sum of its scores with respect to each ranking, such that the combination weight w serves as a personalization

control parameter. In our search agent, we provide the user with a sliding bar, with which she may control the value of w, thus enabling the user to cancel the personalization at any point of time.

3.3 Personalization Strategy

In the following we present for each search mode in detail how it affects the ranking of search results. The decision which search goal a user pursues, and thus which kind of personalization method applies, is currently based on heuristics. However, we plan to study more sophisticated adaptation factors next.

Re-finding known information. Whenever the first query in a session, has occurred in some previous session, we assume the user wants to re-find some information already searched before. We apply three strategies to satisfy this user need. First, we consult the local index for a suggestion how to re-write the query sent to the underlying search engine (in our case Google). This is the only case in which our search agent changes the query sent to Google. We thus implement a more conservative query expansion mode than [13]. Furthermore we give the user full control over this feature so that she may choose to cancel the automatic query rewriting at any time. Considering how the user modified her query in the previous session, gives valuable hints on an improved query formulation. We choose the last query in the previous query chain as the user most likely stopped refining her query when she was satisfied with the results. To ensure the robustness of our method, we additionally require the reformulated query to share at least one term with the original one.

Second, the query representation is updated interpreting all previously clicked result pages as relevant documents, whereas intentionally non-clicked documents ranked above clicked ones are treated as irrelevant to the query. However, non-clicked results that are ranked higher than a clicked one could be interpreted in two different ways: either the user has examined the result title and snippet and was not satisfied with the result, or the user has already seen or knows the result from a previous interaction. Thus in case the local index contains the result, we assume it is known to the user, and do not consider it as an irrelevant document, but ignore it during query refinement.

Third, documents visited in the previous session starting from result pages are returned as additional clickstream results associated with the result item from which they have been reached. We believe this to be useful in cases where the result page is a directory or a summary page with many links to more specific documents.

Finding out about topics of user interest. Whenever there were no interactions recorded for a recurring query or the first query in the session did not occur before, what the user is generally interested in, might be a good guess of her current interest. Thus, we perform personalized pseudo-relevance feedback by assuming that the top-10 documents retrieved from the user's local index are relevant. The terms used to construct the query vector are selected from the titles or summaries of the top-10 documents.

In addition, the returned result set is extended by the top-10 documents from the local index. By doing so, we enable the user to search her own history of viewed Web pages.

Serving an ad-hoc information need. For every query except the first in a session, we refrain to the context provided by the current search session for personalization. The query representation is updated whenever a result click, a page visit or a query refinement occurs. (1) In case of a result click, the user profile of the query to which the result belongs is updated to include terms that best differentiate the clicked and intentionally non-clicked results. For all other similar queries within the session, the query vectors are updated to incorporate terms from the clicked result. (2) In case a page visit has occurred, the query vectors of all queries that are similar to the visited page are updated to incorporate terms from the visited page. (3) Finally, in case the user has refined her query, the new query is augmented with terms from previous similar queries within the current session. Again, for computing query similarities queries are represented by the result sets' centroids. Updating the representation of earlier queries in the current session is typically useful in cases where the user returns back to a query and investigates its unseen results, or in the face of parallel query submissions through tabbed browsing.

4. EVALUATION RESULTS

4.1 Experimental Setup

In order to evaluate the effectiveness of our proposed approach, we asked 9 volunteers to evaluate 10 self-chosen queries. Before the evaluation took place, the participants used our proxy on their local machines to log their browsing activities for a period of 2 weeks. For each participant, 5 of the evaluated queries were about topics the participant had been inquiring during the logging period and were used to assess the effectiveness of our personalization approach in case of re-finding known information, and finding out about topics of user interest. For each query, the participant was presented with the top-50 Google results, respectively additional top-50 results for the re-written query, placed in random order in order to avoid result's position bias. Then the participant was asked to mark each result as highly relevant, relevant or completely irrelevant. The rest of the evaluated queries were used to assess the quality of result re-ranking based on the search session context. Thereby participants performed a normal search with our personalization in place. After finishing their search, they were asked to evaluate the top-50 Google results of the last posed query in that session.

To measure the ranking quality, we use the Discounted Cumulative gain (DCG) [7], which is a measure that takes into consideration the rank of relevant documents and allows the incorporation of different relevance levels. DCG is defined as follows

$$DCG(i) = \begin{cases} G(1) & \text{if i = 1} \\ DCG(i-1) + G(i)/log(i) & \text{otherwise} \end{cases}$$

where i is the rank of the result within the result set, and $G(i)$ is the relevance level of the result. We used $G(i) = 2$ for highly relevant documents, $G(i) = 1$ for relevant ones, and $G(i) = 0$ for non-relevant ones.

4.2 Experimental Results

As shown in Table 1, automatical re-writing of recurring queries clearly improves search result quality. E.g., the query "eccentricity" is reformulated as "eccentricity graph

Result Set	NDCG	Standard deviation
Original Google	0.469	0.178
Automatically re-written	0.803	0.128

Table 1: Average NDCG for recurring queries.

"eccentricity" Google's NDCG: 0.138	"eccentricity graph theory" Personalized NDCG: 0.932
Eccentricity - Wikipedia, ...	Glossary of graph theory - Wikipedia, ...
Orbital eccentricity - Wikipedia, ...	Glossary of graph theory - Information from Answers.com
Eccentricity - from Wolfram MathWorld	Graph Clustering for Very Large Topic Maps
Eccentricity ONLINE	Graph Theory - from Wolfram MathWorld
Scropio	LINK: a combinatorics and graph theory work bench ...

Table 2: Top-5 results (query re-writing).

theory" by our system resulting in a more than 6 times better NDCG (see Table 2 for the top-5 results).

Ranking Method	NDCG	Standard deviation
Original Google	0.806	0.182
Local-index PRF	0.794	0.182
Local-index PRF($w = 0.5$)	0.824	0.182
Textual Similarity	0.681	0.203

Table 3: Average NDCG for pseudo-relevance feedback from the local index.

Table 3 gives NDCG values for pseudo-relevance feedback (PRF) based on the local index. The pure personalized results slightly overdue personalization, however, when combined with the original web ranks choosing $w = 0.5$, the original Google results are outperformed. As an additional competitor we consider the performance of re-ranking the top-50 Google results based on pure textual similarity to the original query, which is consistently outperformed by the personalized results. Results for the sample query "vilnius, lithuania" are presented in Table 4. We see the top-5 results from the user's local index being all about hotels in Vilnius, thus biasing the personalized results more towards travel guides, hotels and outings in Vilnius.

When investigating the effectiveness of the session context for personalization, we find slight but consistent improvements over the original Google results (see Table 5). Again, combining personalized results and original web ranks further improves the ranking. The ranking quality obtained by re-ranking the results based on the local index indicates the need for our approach of different personalization strategies based on the information need.

Top-5 local index results	Google NCDG: 0.731	Personalization NCDG: 0.841
Radisson Sas Astorija hotel Vilnius	Vilnius City Municipality	Lithuania travel guide
Ramada Vilnius hotel	The Web's No. 1 Lithuanian Tourist Guide	Vilnius, Lithuania Restaurants
Crowne plaza Vilnius hotel	U.S. Mission to Lithuania	Lithuania Hotels Booking ...
Novotel Vilnius hotel	Vilnius	Lithuania in your pocket city guide ...
Vilnius forum: Kaunas to Vilnius - trip advisor	Lithuania in your pocket city guide ...	Vilnius Lithuania (Google maps)

Table 4: Top-5 results for query "vilnius, lithuania" (pseudo-relevance feedback from the local index).

Ranking Method	NDCG	Standard deviation
Original Google	0.766	0.18
Session context	0.783	0.154
Session context ($w = 0.5$)	0.784	0.185
Local-index PRF	0.751	0.18
Local-index PRF ($w = 0.5$)	0.729	0.192
Textual Similarity	0.666	0.201

Table 5: Average NDCG for session-context personalization.

5. REFERENCES

[1] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move - user activity tracking for website usability evaluation and implicit interaction. In *WWW*, 2006.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, 1998.

[3] P.-A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschuetter. Using odp metadata to personalize search. In *SIGIR*, 2005.

[4] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW*, 2007.

[5] C. Dwork, S. R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, 2001.

[6] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, 2002.

[7] K. Jrvelin and J. Keklinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, 2000.

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM symposium on Discrete Algorithms*, 1997.

[9] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. In *IEEE Trans. on Knowledge and Data Eng.*, 2004.

[10] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *WWW*, 2006.

[11] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, 1971.

[12] G. Salton and M. J. McGill. Introduction to modern information retrieval. In *McGraw-Hill*, 1983.

[13] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *CIKM*, 2005.

[14] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *WWW*, 2004.

[15] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *KDD*, 2006.

[16] J. Teevan. The re:search engine - helping people return to information on the web. In *UIST*, 2005.

[17] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR*, 2005.

[18] Y. Xu, B. Zhang, Z. Chen, and K. Wang. Privacy-enhancing personalized web search. In *WWW*, 2007.

‘I’ll just Google it!’: Should lawyers’ perceptions of Google inform the design of electronic legal resources?

Stephann Makri
UCL Interaction Centre, 31-32 Alfred Place, London WC1E 7DP, UK.
+44 (0)20 7679 5242
S.Makri@ucl.ac.uk

Ann Blandford
UCL Interaction Centre, 31-32 Alfred Place, London WC1E 7DP, UK.
+44 (0)20 7679 5288
A.Blandford@ucl.ac.uk

Anna L. Cox
UCL Interaction Centre, 31-32 Alfred Place, London WC1E 7DP, UK.
+44 (0)20 7679 5295
Anna.Cox@ucl.ac.uk

ABSTRACT
Lawyers, like many user groups, regularly use Google to find information for their work. We present results of a series of interviews with academic and practicing lawyers, where they discuss in what situations they use various electronic resources and why. We find lawyers use Google due to a variety of factors, many of which are related to the need to find information quickly. Lawyers also talk about Google with a certain affection not demonstrated when discussing other resources. Although we can design legal resources to emulate Google or design them based on factors perceived to make Google successful, we suggest this is unlikely to better support legal information-seeking. Instead, we suggest the importance of taking a number of inter-related tradeoffs, related to the factors identified in our study, into account when designing electronic legal resources to help ensure they are useful, usable and used.

Categories and Subject Descriptors
H.1.2 [**Human Factors**]: Human information processing.

General Terms
Human Factors

Keywords
Information-seeking, Google, law, legal, digital libraries, Grounded Theory, user studies.

1. INTRODUCTION AND RELATED WORK
Google is arguably one of the greatest Internet success stories of our era. In a study by Aula et al. [1] of 236 experienced web users, Google was used as a primary search engine by 95.3% of them. Indeed, in 2006 the word ‘Google’ became a verb in the Oxford English Dictionary. In this short paper, we examine what we can learn from Google’s success when designing electronic legal resources. We discuss, by referring to a series of interviews with lawyers and law librarians, the perceived factors that make Google successful. We suggest that rather than design electronic legal resources to be ‘more like Google,’ we should learn from users’ affectionate comments about Google and design systems with an awareness of the factors perceived to make Google useful, along with an awareness of the associated design trade-offs.

Most related to our work is a study by Fast and Campbell [2], who observed and compared Librarianship and Information Science students searching the Web using Google and searching a web-based library catalogue (OPAC). As well as video and audio recordings, they collected retrospective verbal reports from the students and asked them questions about their perceptions of Google and OPACs. They presented their results in the form of five paired categories: organisation and clutter, trust and evaluation, expectations and confidence, time and effort and freedom and control.

The study revealed two paradoxes. Firstly participants praised the way OPACs were organised, but preferred to use the Web even though they noted it to be disorganized. Secondly, they displayed trust for documents in the library catalogue, but remained confident that they could evaluate the trustworthiness of documents on the web, even though they noted these documents could sometimes be untrustworthy. Fast and Campbell suggest the students’ preference for Google might be due to the confidence that systems like it, which have a low skills threshold, provide, along with design and interface factors. Arguably part of the preference for Google may also be because, unlike library catalogues, it provides access to many of the documents it indexes.

Our study also examines perceptions of Google, but using lawyers as opposed to Librarianship and Information Science students. As with other busy professionals, legal information-seeking is often characterised by heavy time pressure. For lawyers, this means pressure to gain a complete, correct and current picture of aspects of the law, often in a limited amount of time. Legal information-seeking has caused interest in the fields of Information Behaviour Research and HCI research alike, with a number of recent user-centred studies such as [3] and [4].

2. DATA COLLECTION AND ANALYSIS

Our study involved a series of semi-structured interviews with twenty-eight academic lawyers, five law librarians and fifteen practicing lawyers. Most of the academic lawyers were based at a large London university, whilst two were from a nearby vocational law college. Academic lawyers included taught students from first year LLB (Batchelor of Law) undergraduates to LLM (Masters of Law) level and also included research students and staff, lecturers and a Professor of Law. The law librarians worked for libraries that belonged to or were affiliated with the two academic institutions. The practicing lawyers worked in the Dispute Resolution department of the London branch of a multinational law firm and ranged from Trainee to Associate level.

During the interviews, the lawyers were asked questions related to their background and the extent to which they perform electronic information-seeking as part of their work. They were also asked what electronic resources they regularly use, in what situations they use them and why they use them. These interview questions formed part of a broader Contextual Inquiry into academic and practicing lawyers' information-seeking behaviour. The academic strand of the study is discussed in [5]. During the study, as well as using Google and Google Scholar, lawyers chose to use a variety of dedicated electronic legal resources, including resources produced by LexisNexis Butterworths and Westlaw, two major publishers of electronic legal resources.

The interviews were transcribed and analysed according to the open and axial coding stages of Strauss and Corbin's Grounded Theory [6] and excerpts from the transcripts are presented below. In the excerpts, academic lawyers are denoted by an 'A,' law librarians by an 'L' and practicing lawyers by a 'P.' '[…]' denotes omitted text.

3. FINDINGS

We found that lawyers select electronic resources, including the Google search engine, due to a variety of factors, which include the perceived *quality of results*, *degree of flexibility and control offered*, *simplicity and approachability*, *familiarity* and *speed/time-saving benefits*. These factors are highly subjective and inter-related. We argue that many of these factors are linked to the important need for lawyers to find information quickly, which many academic and practicing lawyers pointed out was extremely important when working on client advice, preparing for court or preparing lectures.

Most lawyers were aware that Google provides legal information that is more general than that provided by dedicated legal resources and therefore should be used for different search purposes. However, bearing in mind these different purposes of use, Google is perceived by lawyers to provide **quality results** or as one undergraduate student phrases it, it 'tends to pull up exactly what I need.' This is not the case with other electronic legal resources. Related to the quality of results, some lawyers commented on the wide document coverage provided by Google. Lawyers were aware of the need to be cautious with the regard to the authority of documents and recognised that Google was useful for, as one vocational student phrased it, 'gaining a layman's perspective' on legal issues as opposed to a legal perspective.

Another pair of factors identified was that of the **degree of flexibility and control** that lawyers perceive Google to offer (related to Fast and Campbell's 'freedom and control' category). These factors are illustrated by the Lecturer and student below respectively. Not only does Lecturer A6 highlight tolerance of 'vague' search terms in Google, but also speaks of getting 'the result *quicker'* (referring to obtaining a particular case), relating Google's search input flexibility to potential time savings:

"The difference is probably that in the British and Irish Legal Information Institute and in Westlaw, the search engines need a greater degree of precision. You know, full case names, citations, something like that. […] With Google I find that a vague approximation of the relevant terms actually gets the result quicker." - A6 (Lecturer)

"With Google, you just define everything. You're in control with Google. Well that's what you think anyway, and I like that. […] You can define everything, you can choose everything." – A11 (Undergraduate student)

Another factor that influences Google's use is its perceived **simplicity and approachability**, as explained by the undergraduate student and law librarian below:

"I used to hate computers. So Google is something simple and looks approachable to me. […] Google made me like computers!" – A11 (Undergraduate student)

"I think law students are the same as all other students, are the same as all other people that are not involved in the information profession. They just think Google is a gift from heaven and it's fabulous. ***R: What exactly is it about Google?*** *Ease. Ease of use. Solely and specifically ease of use. One box, search terms in, voomph! Twenty seconds later, results back." – L1 (Law Librarian)*

Again, Law Librarian L1 illustrates the link between the simplicity and approachability of a resource and speed/time savings. In addition, the above factors can all contribute to lawyers' repeated use of Google which, in turn, can lead to **familiarity** with the system (and speed and time savings when using Google over other resources):

"I don't think I know how to look into the online legal journal databases very well, but I know how to use Google very well because I do it all the time when I use the Internet." – A8 (Undergraduate)

Indeed, **speed/time savings** are important factors in their own right for explaining Google's popularity, particularly amongst lawyers who have indicated that legal work is often particularly time-sensitive:

"Sometimes I've had a very pressured time just to get an answer to something. Google. It's brilliant! The other

search tools, you really have to have marshalled your thought process a bit more, I think. And while they're more effective, if you've got limited time, I'm a great believer in Google - I think it's great!" – P5 (Associate)

Like the Associate above, many lawyers spoke about Google using affectionate terms such as 'it's brilliant.' One student claimed he was 'very grateful for Google.' This affection for Google also extended to practicing lawyers. Indeed, members of all groups of lawyers in our study spoke of Google in a positive light (and none spoke of it in a negative light). In addition, although the lawyers in our study displayed varied search sophistication (in general taught students were less sophisticated than other groups of lawyers at information-seeking), the factors we have highlighted were identified by lawyers across the board, not just by those in certain groups.

4. IMPLICATIONS FOR DESIGN

When referring to dedicated electronic legal resources, lawyers were just as negative as they were positive. Many lawyers, particularly taught students, spoke of frustration concerning knowing where in the system to go in order to find a particular type of legal document. Lawyers also mentioned (or demonstrated in the Contextual Inquiry part of the study) that they sometimes found it difficult to know where within a document or meta-data a search that is restricted to a particular segmented field might match their search terms to in order to bring back the results.

Figure 1 illustrates a mock-up of part of a (fictitious) electronic legal resource that allows users to search for legal case reports. These reports can be searched in the traditional way by entering search terms (perhaps connected by Boolean syntax) or by entering text into a number of segmented fields (such as a 'case name' field which might search for the text entered in the field in the title of the case or a 'judges' field which might search for cases that have been heard by a particular judge or judges).

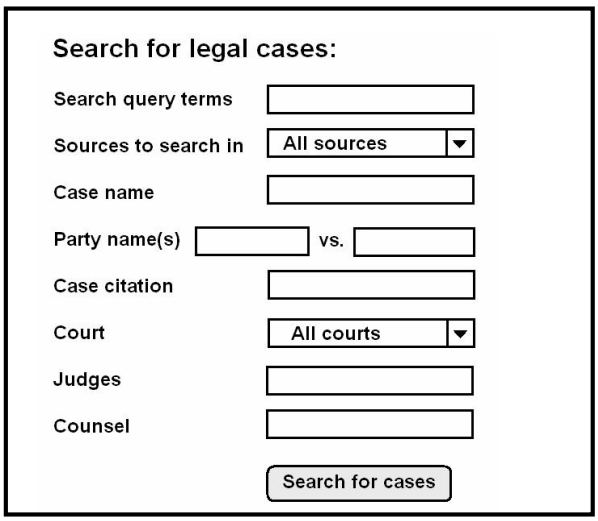

Figure 1. Mock-up of segmented search fields commonly used in electronic legal resources to facilitate searching for legal cases.

Although we can design legal resources to be 'more like Google' by reducing complicated system features such as the segmented search fields above and by providing a simple open search field, we suggest this is unlikely to be useful for supporting legal information-seeking. This view is supported by this law librarian, who explains that the kind of 'woolly' searching that Google provides may not yield suitable results if implemented in an electronic legal resource:

"I think there are advantages in making systems more Google-friendly or Google-like but not to give in to that whole system and assume that that kind of woolly type of searching is going to produce the results that you really want." – L3 (Law Librarian)

It is also tempting to design legal resources based on the factors that we have highlighted that make Google successful. However, this approach (along with the approach of designing to emulate Google is potentially risky. This is because both of these approaches do not take the *information being sought* (and therefore the information-seeking tasks that the electronic resource should be designed to facilitate) into account. For example greater control, as provided through the segmented field searches in figure 1, may be preferable to lawyers when looking for a particular case or citation of a case, but not when trying to gain an overview of a legal area by examining various cases that have dealt with a particular legal subject.

Rather than prescribing how to design 'optimum' electronic legal resources based directly on the Google search engine or around the factors which are perceived to make Google successful, we suggest the need to make considered design decisions by making careful tradeoffs between the factors we have discussed. These tradeoffs must be made based on the *context* in which the resource will be used. For example, providing a single open search field (as opposed to several segmented fields) to allow users to search highly organised repositories of legal cases is unlikely to produce *quality results* in situations where lawyers have particular details about the legal material being sought and simply want to find it.

However, introducing too many segmented search fields to facilitate more powerful searching might improve the quality of results, but impact negatively on the *simplicity and approachability* of an electronic legal resource, making it more difficult to increase *familiarity* with the resource. It might, however, impact positively on the *degree of contro*l offered and, if it improves the quality of results, might also provide *speed/time-saving benefits*.

It is important to highlight, however, that the process of considering these tradeoffs is likely to differ for each electronic resource being designed. Therefore we do not believe that it is useful to design a resource based on factors perceived to make Google or any other resource successful. Instead, we suggest the importance of taking a number of inter-related tradeoffs, related to the factors identified in our study, into account when designing electronic legal resources. We also suggest that the balance of these tradeoffs that is likely to make an electronic legal resource successful will be

highly dependent on the types of information that the resource is designed to facilitate searching. We argue that only by considering the balance of these tradeoffs for each electronic resource that we design can we help ensure that our resources will be useful, usable and used.

We suggest that future research in this area might focus on examining a particular electronic resource or resources in light of the tradeoffs that we have discussed with the aim of suggesting ways of improving the design. For example, it may be possible to prototype a new electronic legal resource and ask lawyers to explore using the prototype and discuss their experience making reference to the factors and associated tradeoffs that we have identified.

5. ACKNOWLEDGMENTS

This work is partly supported by an EPSRC DTA studentship.

6. REFERENCES

[1] Aula, A., Jhaveri, N. and Käki, M. (2005). Information Search and Re-access Strategies of Experienced Web Users. In Proceedings of the 14th International Conference on the World Wide Web, pp. 583-592.

[2] Fast, K. and Campbell, D. (2004). 'I still like Google': University Student Perceptions of Searching OPACs and the Web. In Proceedings of the 67th ASIS&T Annual Meeting, pp. 138-146.

[3] Jones, Y.P. (2006). "Just the Facts Ma'am?" A Contextual Approach to the Legal Information Use Environment. In Proceedings of the 6th ACM Conference on Designing Interactive Systems, pp. 357-359. University Park, PA, USA.

[4] Komlodi, A. & Soergel, D. (2002). Attorneys Interacting with Legal Information Systems: Tools for Mental Model Building and Task Integration. In Proceedings of the 65th Annual Meeting of American Society for Information Science and Technology, Philadelphia, USA, pp. 152-163. ACM Press.

[5] Makri, S., Blandford, A. & Cox, A.L. (In Press). Investigating the Information-seeking Behaviour of Academic Lawyers: From Ellis's Model to Design. To appear in the Information Processing and Management special issue on Digital Libraries in Context of Users' Broader Activities.

[6] Strauss A. & Corbin J. (1998). Basics of Qualitative Research. Sage Publications, UK.

Increasing the speed of Information Access on the Mobile web using HTML feature extraction

Andreas Komninos
Glasgow Caledonian University
Cowcaddens Road
Glasgow G4 0BA, UK
+44 141 3313095
andreas.komninos@gcal.ac.uk

Chris Milligan
Glasgow Caledonian University
Cowcaddens Road
Glasgow G4 0BA, UK
+44 141 3313095
cmilli10@caledonian.ac.uk

ABSTRACT
Motivated by the cumbersome process of extracting information from webpages as rendered on mobile device web browsers, this paper focuses on describing an alternative and promising approach to facilitating the process for users. We present early work-in-progress on a system that attempts to extract information sections of a webpage and presents the extracted sections first, with the remaining page following. Early trials of a rudimentary prototype show promising results and we discuss further work to be carried out for the improvement of the system.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – *Information Filtering, Retrieval Models*

H.5.2 [**Information Systems**]: Information Interfaces and Presentation - *User Interfaces*

H.5.4 [**Information Systems**]: Hypertext/Hypermedia - *Navigation*

General Terms
Algorithms, Design, Experimentation, Human Factors

Keywords
Mobile Information Access

1. INTRODUCTION
Browsing the web on mobile devices is well known as a problematic process, mostly from a usability point of view. Because of the general layout of normal web pages, which are designed for viewing on desktop computers, the inevitable vertical and horizontal scrolling required to access information contained therein, when using a mobile device browser, poses a serious impediment to the process of mobile information access. Due to the nature of mobile devices and their mode of use, which is drastically different to that of desktop computers, it is clear that alternative approaches are required to rendering webpage information on mobile device screens. The following section critically discusses work already undertaken in the field in a brief manner. The description of related work is non-exhaustive but indicative of the state of the art in this area. A description of our own approach is discussed thereafter, followed by recommendations on future work that we are planning on our early prototype.

2. RELATED WORK
We mentioned earlier the difference in the mode of use of mobile devices, compared to desktop computers. Typical usage patterns for mobile devices show that these are used for very short ("burst") periods of time during the day, when the user requires immediate access to some information or function of the device, followed by large periods of inactivity. The "always-on-standby" model for managing power on mobile devices such as PDAs or smartphones is a good example of functionality derived from the requirement arising from these usage patterns. Fujmoto [1] uses the term "nagara mobilism" (nagara = "while doing something else") to explain that this pattern of usage is central to the behaviour and adoption of devices by young users. This mode of use highlights the need for quick access to information that is relevant to the user's tasks, something that is currently not well supported in mobile web browsing.

To solve the problem of webpage rendering on mobile devices, two approach categories can be identified within existing literature and commercial systems: Client-based and Server-based processing of HTML documents. The first approach delegates the task of processing and rendering HTML documents in a more appropriate form to the mobile device itself. Documents are processed after having been downloaded in a variety of methods, most often in an attempt to eradicate horizontal scrolling, which imposes the largest interaction cost and impediment to mobile information access. Yin et al [2] devised the method of taking a normal webpage and rearranging the HTML in a way that eliminates horizontal scrolling. Their system examines the semantic relationships between HTML elements and sections to intelligently decide the order in which they will be "stacked" on top of each other. Other papers have also incorporated this method into their projects such as Liu et al[3], Dontcheva et al [4] etc. SmartView [5] divides the webpage into logical sections which can then be viewed independently of the rest of the document, although this requires explicit user instruction. This system benefits the users of PDAs whose touch-sensitive screens are easy to navigate, but usability problems would probably arise during use on a mobile device where the only navigation mechanism is the joypad. The Access NetFront browser [6] is a browser which implements this style of display with no horizontal scrolling, using a technology called SmartFit. SmartFit uses a process of restructuring the node arrangement so that there are no

nodes side by side. This is used to position nodes with a single breadth, one on top of the other. Another option for this browser is a process named JustFit, which allows for the page to be "squeezed" so that layout is very narrow, but all the sections are viewable without horizontal scrolling. As the nodes are squeezed, they appear long and narrow and are generally hard to read. Other commercial browsers like Opera [7] and Thunderhawk [8] address the same problem by "stacking" and providing a zoomable overview of sections respectively. A combination of both technologies is appearing in Microsoft's latest DeepFish browser [9].

With mobile device processing power increasing, the requirement on resources for client-side adaptation is not as taxing. Another advantage of client-side adaptation is that the client knows its own properties, thus being able to guide the adaptation process more effectively than a server-side system which has to rely on standardized profiles. A client-side adaptation system can also more easily forward decision-making to the user when the right decision is not obvious. For example, it can ask the user whether he wants a shortened, high-usability or a full, low-usability version of the content. This information brings to light the fact that device properties may have an influence on the behavior of the adaptation system.

On the other hand, server-side adaptation, where HTML documents are processed by a proxy before being served to the device, has not been studied to the same extent. The process occurs entirely on the server end, which means that the server has to estimate the characteristics of each mobile device in order to give an accurate transformation of web content. Of course this is not always possible so the content is adapted for a stereotypical mobile device. An advantage of server-side adaptation however, is that (especially with the removal of irrelevant content) the amount of data that a device would have to download and store would be slightly less, thus reducing download times and system overhead. Finally, server-side conversions may result in wasted (expensive) bandwidth if the adaptation is not desirable or restructures a page in such as manner that it is rendered unreadable by the user. Thusfar, server-side adaptation is provided by a very small number of proxies. Google have their own technology for adapting pages for mobile devices by segmenting and presenting a single page as multiple pages. Their system will also attempt to take the user to the "section" sub page which is most likely to contain text relevant to the query. A server-side webpage conversion service is offered by Skweezer.net [10] using the Ask.com web search engine.

All the approaches mentioned above try to address the problem of rendering standard web sites on small screen devices. While the approaches are more or less successful, they contribute little to the problem of facilitating access to relevant pieces of information. In eradicating the problem of horizontal scrolling, they aggravate the amount of vertical scrolling that is required to navigate the page. The Google approach seems to be the only one trying to intelligently aid the user, however, because of its default behaviour that strips each page of all non-textual elements and the non-existent control that the user has over the process, it has been heavily criticized with some users and content developers treating it as a form of "censorship".

3. METHODOLOGY

Our system is based on the simple assumption that when looking for relevant information on a webpage, a user will most likely prefer to have immediate access to those sections of the webpage that are likely to contain the information required. We therefore hypothesized that if, through query analysis, we could display those sections before the entire webpage, users would be able to obtain the information required much more quickly and thus reduce the need for horizontal scrolling.

Work by Kamvar and Baluja [11] on searches conducted using Google through mobile devices highlights the type of query most likely to be sent as predominantly belonging to one of the following categories: Local services, Travel & Recreation, Technology and Entertainment. The existence of "technology" searches can probably be attributed to the early adopters of mobile web technology being generally interested in technology. However the other categories, especially local services, hint towards the type and, importantly, size of information required: Addresses, types of business, short reviews or directions. Such information is typically collated with several unwanted elements (e.g. websites will list all Italian restaurants in a particular city). It is clear that users require "snippets" of information that would allow them to carry out a very specific task, not the entirety of the website information.

Based on these observations and our assumption as stated above, we began building a prototype system that operates on the principles of

a) Identifying the logical sections of the webpage that contain information relevant to the query, as input by the user;

b) Reconstructing the webpage presented to the user in such a manner that these sections are presented first, stacked on top of each other, followed by the remaining web-page which is otherwise not manipulated.

To accomplish the above, our system uses currently a naïve approach that assumes logical sections on a webpage correspond to physical sections marked by , <td> and <p> HTML tags. The sections are again very naively weighted for relevance to the original query by determining the Term Frequency of each query keyword in those relevant sections. We limited the system to display only the top 5 sections in terms of their calculated weight, in an attempt to limit the (potentially) lengthy additions to the top of the resulting viewable document.

Futhermore, each section text is "wrapped" around some custom HTML to ensure that it is clearly presented as an extract of the original website and not part of its original structure. For this purpose the extracted sections are presented as part of a table with double border and gray background.

It is important to note here that we chose not to strip other HTML elements from the sections marked by our designated tags. For example, text following a <p> tag, which can contain other tags, such as images, formatting or link tags, is kept "as-is". The reason for that is that we do not want to extract from the context of the retrieved content (for example, an image might be part of the required information of a link might be of benefit to the user for following more information). Sample screenshots of our system can be seen in Figure 1 below.

Figure 1. The left screenshot shows the query interface (up to 3 keywords are allowed). On the right, the original page with sections added on top (gray boxes)

We proceeded in carrying out preliminary tests of this rudimentary prototype, although it is clear that it is quite far from being perfect. We will discuss our planned improvements in the following sections; for now we will focus on the findings of our initial investigation, which aimed to assess the users' perceptions of the utility of such a system and whether it would be possible, even with a naïve approach, to obtain results that might be encouraging towards further development.

4. INITIAL EXPERIMENTATION

We asked 4 subjects to try out our early prototype by asking them to find out certain pieces of information by browsing the mobile web. Three of the subjects were computing students, with one being a computing novice. One of the volunteers owns a PDA and has used it to browse the web, and another volunteer uses PDAs quite frequently (although they don't own one).

4.1 Search Comparison Test

The first test to be carried out was in order to assess the ease of information retrieval from the prototype. The test setup involved an initial browse of a web page on a desktop computer to pick a random piece of information. The testing participant was then told to find this piece of information using three keywords relating to the information, and to signal once found. We asked for 3 keyword queries to replicate the fact that the average query length for the top 5 search categories, as described in [11]. Timing started from when the user clicked on the Google search results link for the particular page. When the piece of text had been found, the timer was stopped and recorded. This process was achieved using Pocket Internet Explorer with Google.com, and then using the prototype browser. Each participant repeated the task three times with different items of text from different web pages to search. The following figure (figure 2) shows the average access time, which appears to be roughly 10% less when using our prototype.

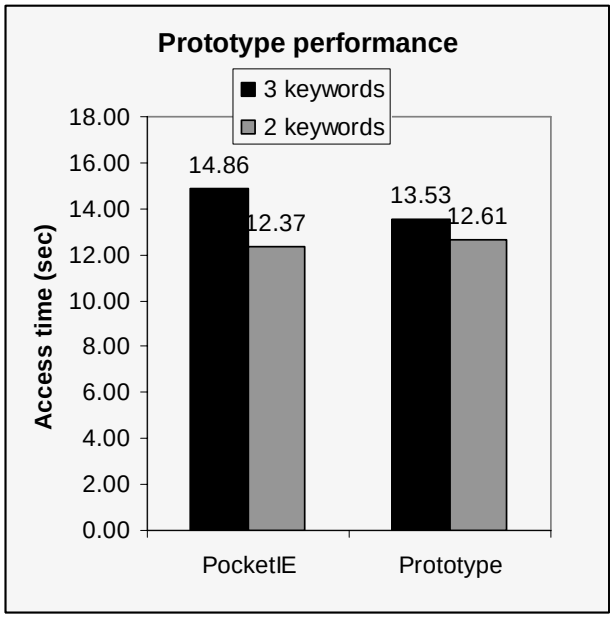

Figure 2. Performance of prototype vs. PocketIE

We repeated the experiment instructing the users to formulate 2 queries this time (again from [11] this figure is the average query length for all categories), using the same procedure and similar tasks. In this instance we found the performance to be comparable with PocketIE being marginally (~0.3 sec) better than our prototype.

Finally we asked the users to complete a post-experiment questionnaire, in which we asked them to rate their perceptions of the system on a scale of 10 (a larger score is more positive). Overall the prototype received average feedback on the look of the reconstructed page (5.4/10) and an average perceived effectiveness score of 5.0/10. While these results are perhaps not exactly splendid, we must consider them as a sign of indifference to the effect of the overall presentation of the viewed document, which in itself, is a positive finding as the users don't seem to mind the addition of extra elements on the document. The perceived effectiveness score is also rather average, but again we consider this to be an encouraging result given the naivety of the adopted approach for this early prototype.

Another section of the post-experiment questionnaire allowed subjects to leave general comments of their opinion of the prototype. This was probably the most encouraging section with all users commenting positively on the intuitiveness of the browser and its speed of use. Three users also commented positively on the simplicity of the design. Finally two users commented on the effectiveness of the section extraction methods as an area they would want to see improved.

5. FUTURE WORK

When testing the prototype we anticipated that the initial reaction to it would not have been enthusiastic, given the simplicity of our approach. However, based on the test subject comments, we were encouraged by the fact that it was immediately obvious to them that our approach would be a helpful aid in browsing the mobile web, if it could be perfected further. The users readily identified

the problem of horizontal scrolling and the lack of speed in finding information on the mobile web. Furthermore, we were encouraged by the findings that seem to indicate that our system, even in its simplest form, is not only non-disruptive to the users, but can also outperform the standard browser for a large proportion of mobile devices.

Although our initial early trial is based on a small sample and cannot be presented as conclusive in any manner, based on its findings, we are ready to conduct much further work on the prototype. It is essential for us to improve the performance of the section identification algorithm, something that we anticipate to be a great challenge. Further from overcoming the problem of parsing loosely structured HTML documents that do not necessarily conform to development standards, the identification of what constitutes a "logical" section within a document will be a hard challenge. Context cues such as colours, blank space dividers, fonts or images, allow the human brain to immediately identify and separate content into logical sections such as the webpage designers meant it to be seen. While difficult, this process is not impossible as past research shows and semantic analysis in conjunction with DOM analysis of the webpage, filtered through layout heuristics, could greatly help.

More importantly, it is important for us to employ intelligent selection mechanisms for choosing the sections that will be displayed on top of the page. TF/IDF instead of simple TF is an obvious candidate for improvement, however we feel that we should be looking at a combination of various weighting heuristics to obtain a more accurate results. More specifically, a Bayesian probability model could possibly increase the selection process accuracy while maintaining the number of sections displayed at a minimum. Further to this, it would be extremely interesting to apply a Markov chain model trained on implicit relevance feedback indicators and user behaviour modeling, to try and accurately assess the order that sections should be presented to the user.

Having mentioned user behaviour modeling, we should also mention here our intention to augment the system by exploiting user models built and trained over time to perform tasks such as query disambiguation and expansion. We have previously explored such methods with good success in the past [12] and we feel that they would be highly appropriate here, given the low probability of long queries sent through a mobile browser. The process of query expansion and disambiguation can help locate sections in a given document that would have otherwise been given a low weighting if, for example, synonyms or closely related terms omitted from the query can be found in a text section.

Once a more advanced prototype has been completed, we would be interested in comparing its performance with a variety of existing systems and with a range of different algorithmic functionality options to determine conclusively what method works best.

6. CONCLUSIONS

In [5], the authors describe how a "section extraction" system could work in conjunction with a search engine by marking each section with a number of small boxes to indicate its potential relevance to a query. This might work quite well for a PDA but considering the navigation mechanisms of a typical phone (joypad), we believe that this will not necessarily result in any significant improvement as the interaction cost of navigating to the relevant section will probably remain just as high.

We are confident that our idea of intelligently "stacking" extracted sections can aid users to navigate the mobile web, thus helping towards the solution to a usability problem that has hindered the use of data services on mobile devices so far.

7. REFERENCES

[1] Fujimoto, K. "The Anti-Ubiquitous "Territory Machine"-- The Third Period Paradigm: From "Girls' Pager Revolution" to "Mobile Aesthetics", in Personal, Portable, Pedestrian: Mobile Phones in Japanese Life, edited by M. Ito, D.Okabe, and M. Matsuda. Cambridge: MIT Press, 2005.

[2] Yin, X. & Lee, W.S. "Using link analysis to improve layout on mobile devices", Proceedings of the ACM 13th international conference on the World Wide Web, New York, 2004

[3] Liu, Z., Ng, W.K., Lim, E.P., & Li, F., "Towards building logical views of websites". Data & Knowledge Engineering Volume 49, Issue 2, 2004

[4] Dontcheva, M., Drucker, S. M., Wade, G., Salesin, D., Cohen, M. F., "Summarizing Personal Web Browsing Sessions", ACM 19th annual Symposium on User Interface Software Technology (UIST), Montreux, 2006

[5] Milic-Frayling, N., Sommerer, R., "SmartView: Flexible Viewing of Web Page Contents". Proceedings of the ACM 11th World Wide Web Conference, Hawaii, 2002.

[6] Access NetFront: http://www.access-netfront.com

[7] Opera mobile browser: http://www.opera.com

[8] Thunderhawk browser: http://www.bitstream.com/wireless

[9] Microsoft DeepFish browser: http://labs.live.com/deepfish

[10] Skweezer mobile webside adaptation: http://www.skweezer.net

[11] Kamvar, M., Baluja, S., "A large scale scale study of wireless search behavior: Google Mobile Search", in Proceedings of the 24th ACM Conference on Human Factors in Computer Systems (CHI2006), Montreal, 2006.

[12] Komninos, A., Dunlop, M.D, "A calendar based Internet content pre-caching agent for small computing devices", Journal of Personal and Ubiquitous Computing (online First), Springer, 2007

Separating Human and Non-Human Web Queries

Yuye Zhang

NICTA Victoria Research Laboratory,
Department of Computer Science and Software Engineering
The University of Melbourne, Australia

zhangy@csse.unimelb.edu.au

Alistair Moffat

NICTA Victoria Research Laboratory,
Department of Computer Science and Software Engineering
The University of Melbourne, Australia

alistair@csse.unimelb.edu.au

ABSTRACT

We describe the evaluation of a set of web search queries that were issued to the MSN search service in May 2006. We compare two mechanisms for filtering the query stream so as to remove queries from automated sources, and contrast various attributes of the respective output streams. Our findings show that the previous implementation of query-based filtering may have removed large segments of queries from within sessions, consequently impacting on query-to-query analysis. Conversely, session-based filtering provides more realistic output, but with the risk of additional machine-generated queries being included in the analysis. We have also discovered some inconsistencies involving persistent URL rankings within the query log which may affect future research utilizing this dataset.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, search process.

General Terms

Measurement, performance, experimentation.

Keywords

Query log analysis, web search.

1. INTRODUCTION

The statistical analysis of data collected by large-scale search engines has served as a basis for many activities in Information Retrieval. By providing insight into various aspects of user behavior in a non-intrusive manner, query log analysis [Silverstein et al., 1999, Spink et al., 2001] allows researchers to form models of user activity, which can then be used to drive enhancements of the retrieval system and thereby improve search quality. However, in order to accurately represent underlying "human" user behavior, it is important that the initial source of data (the query log) should be largely free of activity from automated sources. Performing the necessary filtering can sometimes prove difficult, as publicly-available query logs tend to have been quite extensively pre-processed to remove all information that might allow individual users to be identified.

In this paper we build on our previous analysis of a MSNSearch "query and clickthrough" resource [Zhang and Moffat, 2006]. This query log contains approximately fifteen million queries, sessions and clickthroughs from users in the United States captured during May 2006. Following previously published work, we propose a variation on the filtering method used to remove non-human search interactions, and provide a comparison of key output statistics across the two alternative filtering regimes. We also report on some idiosyncrasies (or possible inconsistencies) in the data set – problems that may impact on the analysis undertaken by others making use of the same resource.

SIGIR Workshop on Web Information-Seeking and Interaction, July 27, 2007, Amsterdam, The Netherlands.

2. LOG FILTERING METHODS

As was noted in our previous work [Zhang and Moffat, 2006], the MSNSearch dataset contains search requests made to `http://search.msn.com` by US users, with each request being timestamped and anonymized to remove any personally identifying information. The requests were grouped into sessions before being made available (presumably using an IP address based segmentation metric), and were distributed with sets of associated clickthroughs. The logs do not represent all traffic to the MSNSearch engine during that month, and were made available simply as a representative sample extracted by the Microsoft engineers that prepared the dataset.

During our initial analysis of the query set, we found that a large number of queries were within sessions that we suspected had been generated by automated processes. We thus elected to remove all queries with zero associated clickthroughs, taking post-query activity as being evidence that the query had originated with a human user. By removing the possibly-automated queries, the log analysis would, we believed, be more informative of human-search interactions.

However, this *query-based filtering* heuristic had the negative impact of removing all legitimate queries where the user chose not click on any results, whether because the results presented were unsatisfactory, because the query was satisfied purely by the on-screen snippet information, or because of some other effect. In particular, the "no click, so remove the query" filtering mechanism removed queries that were embedded in search interaction sessions in which there was in fact evidence of user clickthrough via other queries. These removals disrupted the sequential representation of user interactions in our session-based analysis, and put us at risk of drawing erroneous conclusions.

To explore the extent of this risk, we replicated the affected experiments by working at a higher level, and filtering out all interaction *sessions* for which there were no clickthroughs. That is, queries that did not have an associated clickthrough of their own, but were within a session where another query did have a clickthrough, were retained in the log instead of being discarded. Although this method has the potential to return more false positives in relation to machine generated material, it might be a better com-

Attribute	Original log	Query filtered	Session filtered
Number of queries	14,923,285	8,831,275	12,231,093
Number of unique queries	7,095,622	3,875,436	5,589,822
Number of terms	35,824,851	20,641,810	29,774,415
Number of unique terms	2,605,699	1,151,998	1,785,229
Number of sessions	7,470,913	5,684,599	5,684,599
Average query length (terms)	2.401	2.337	2.434
Average session length (queries)	1.997	1.554	2.152

Table 1: Attributes of the MSNSearch query log, in original form, and with two different filtering regimes applied.

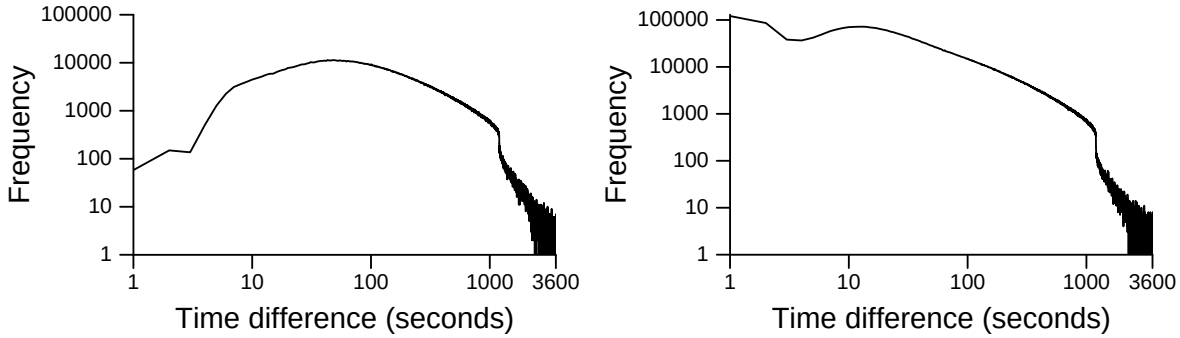

Figure 1: Time interval between queries in multi-query sessions for query-based filtering (left) and session-based filtering (right). Query-based filtering removes many requests which are submitted in quick succession, conversely retained in session-based filtering. These requests are likely to originate from automated sources, although users are capable of issuing them manually by utilizing suggested query terms from the search engine.

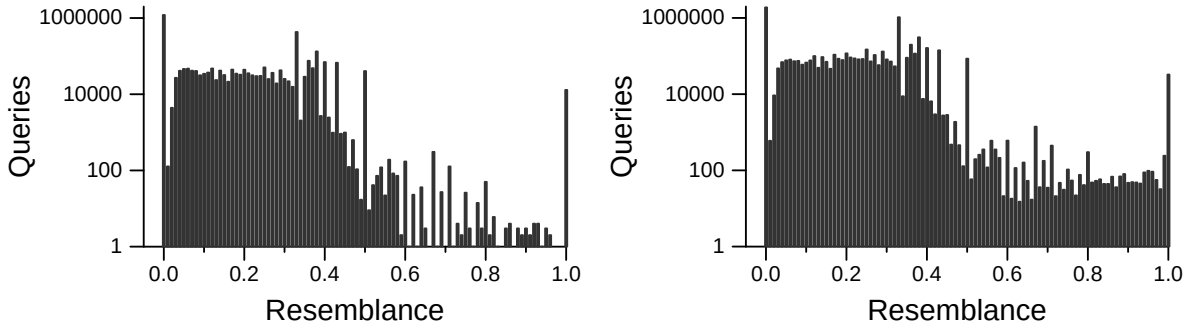

Figure 2: Trigram resemblance between queries in multi-query sessions for query-based filtering (left) and session-based filtering (right). There is an overall increase across all resemblance values in session-based filtering, with significant gains where trigram resemblance is greater than 60%. This change shows that query-based filtering had removed a large portion of queries with overlapping terms from within sessions, causing disruptions to query continuity.

promise in allowing us to model user interaction more accurately. We report our findings for this *session-based filtering* approach as follows.

3. QUERY VERSUS SESSION FILTERING

Table 1 shows the effects of the two different filtering methods on the dataset. Note that the number of sessions is the same using the two filtering methods, since a session is only removed in either approach if all queries belonging to that session have no click-throughs. General volumes for session-based filtering moved to an approximate midpoint between the original query log and query-based filtering, with some differences in the number of queries/terms. This is attributable to the growth in repeated entries as the resultant dataset is enlarged. From the session length statistics, we can see that the number of queries within a session drops significantly in query-based filtering due to more queries being removed. On the other hand, because session-based filtering removes the same number of sessions but retains the continuity of queries within the retained sessions, the resultant sessions are much longer. Average query length remains fairly unaltered in the different filtering approaches.

Figure 1 shows the time difference in seconds between consecutive queries for multi-query sessions. Comparing the two graphs, it is evident that session-based filtering results in many more queries received within just a few seconds of each other, which we had originally believed to be indicative of possible machine issued queries

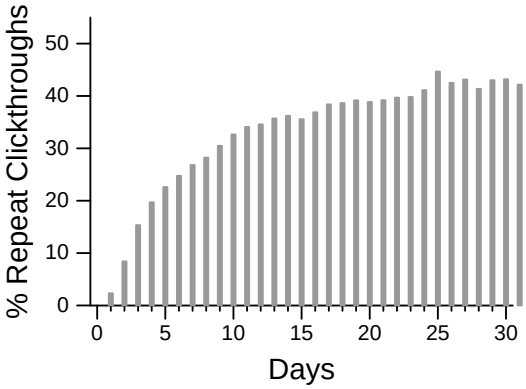

Figure 3: Days between consecutive clickthroughs on the same query/URL combination where URL ranking was changed on the second clickthrough. As time progresses, the probability that there is a change in a URL's ranking increases, even in response to the same query.

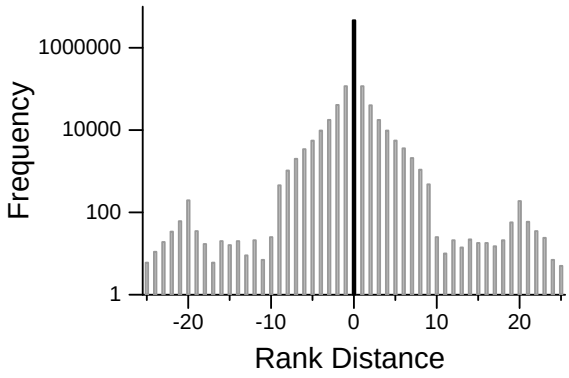

Figure 4: Difference in rank between consecutive clickthroughs on the same query/URL combination. In the majority of cases, URL position for a given query is relatively stable, but there are surprising breaks in that pattern at $+20$ and -20 that may indicate errors in the data.

arriving at very short intervals. In contrast, the query-based filtering output is visibly biased towards longer intervals. If we accept that a session-based approach is the more useful test for the queries being of human provenance, then one possible explanation for the short inter-query times is the detection at the search service of potential spelling mistakes, and the resultant immediate reissue of a corrected query via a "did you really mean?" followup prompt from the system.

The second peak of query activity (at around ten seconds per query) in the session-based filtering method is still a fair representation of user activity, and corresponds to users looking quickly at a first page of suggested answer snippets, and then reformulating their query. The frequency of occurrence of longer query intervals – those greater than one minute – remains unaltered in the session-based filtering method.

To calculate the syntactic similarity between consecutive queries in the same session we used the standard n-gram resemblance computation, defined for queries A and B as:

$$R(A,B) = \frac{|S(A,n) \cap S(B,n)|}{|S(A,n) \cup S(B,n)|},$$

where $S(D,n)$ is the multiset of substrings of length n characters in the string D, omitting any substrings containing whitespace. Figure 2 compares the distribution of syntactic resemblance for multi-query sessions for the two filtering methods. In comparison to query-based filtering, session-based filtering results in a significantly higher portion of query pairs with similarity scores in excess of 60%. In real terms this translates into, approximately, a two-words-in-three overlap.

Based on these findings, we conjecture that the query-based filtering method used in our first analysis had indeed removed reasonable queries from within multi-query sessions arising from the actions of real users. When the integrity of individual sessions remain unaltered as per the session-based filtering method, we can see that in actuality consecutive queries within a given session actually can share a fair degree of syntactic similarity, although this is still outnumbered by those queries with little to no resemblance.

4. CHANGES IN RESULT RANKING

An important component of query log analysis is to validate that any processing undertaken on the data to reduce its volume is statistically sound, and that findings on subsets of the data reflect the data as a whole. While some types of anomaly are easily detected and rectified, others can often go unnoticed. This section comments on some aspects of the Microsoft data that we found to be puzzling.

The clickthrough log supplied by Microsoft includes information about the absolute rank of the URLs clicked in response to each query. When coupled with the timestamp information supplied in both the query and clickthrough logs, the ranking of some URLs for some common queries can be traced through the month. While, for a given query, a given URL can be expected to have a roughly consistent answer rank, there are several scenarios under which the ranking might change, including if the page in question is re-crawled and its content changes, or if the collection as a whole is reindexed and the content of other pages has changed. Nevertheless, we would expect such change to be evolutionary in nature, rather than revolutionary.

Figure 3 shows the percentage of URLs (gathered from the clickthrough data) which exhibit a change in their rank position in the answer list for some query, plotted as a function of the interval between the two queries in the repeat query pair. As the time between clickthroughs on that URL/query pairing increases, the likelihood of a change increases, as we had expected. (Note that because we are only able to determine answer rank via the clickthrough data, Figure 3 is not a perfect depiction of changes in result ranking.)

We also examined the size of jumps in ranking of a URL/query pairing between subsequent clickthroughs, with results presented in Figure 4. As expected, the vast majority of repeat queries and repeat URLs experienced no change in ranking, as shown by the strong peak in the middle of the graph.

Assuming that the popularity of a URL plays a role in determining changes to its rank position in response to a query, then we can assume that the jumps will be inhibited somewhat by page boundaries and that jumps of magnitude greater than nine positions will be much less commonplace due to results beyond the top ten being less likely to be viewed (and therefore clicked). This effect may account for the drop off in changes around the boundaries around the $+/-9$ segments. However, the anomalous number of magnitude twenty jumps also caught our attention. Investigation showed

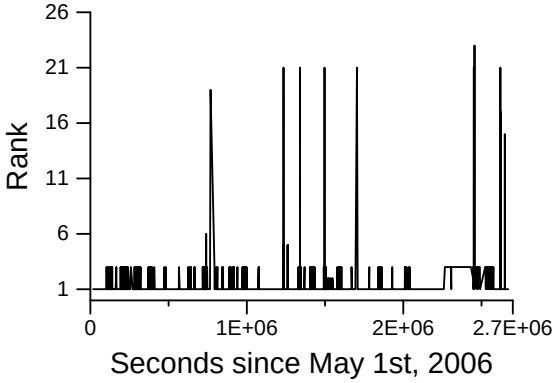

Figure 5: Changes in the rank of the URL `http://www.yellowpages.com` for the query "yellow pages" over the month of May, as captured by clickthrough instances. The ranking should stay relatively constant, and the surprising jumps in the rank of that answer suggest a sporadic error of some sort within the dataset.

that there were some quite surprising shifts in answer rank taking place, even for identical queries.

As an illustration of this discrepancy, Figure 5 shows the changes in the ranking of the URL `http://www.yellowpages.com` for the single query "yellow pages" over the month of May, sampled over the 3,551 dataset instances of the query "yellow pages" that were linked with a subsequent clickthrough to that page. Quite remarkably, there are several instances captured in which the rank of the (presumably correct) page `http://www.yellowpages.com` suddenly jumps to ranks in the twenties. We suspect the anomalous peaks for this and a variety of other queries may be caused by an error within the Microsoft extraction processes used to generated the data. These rank shifts are relatively infrequent across the dataset as a whole; nevertheless, they do raise the possibility of other similar errors, a possibility that needs to be taken into consideration by other researchers working with this data.

5. RELATED WORK

Jansen and Spink [2006] provide an overview of published query log analysis work prior to 2002, including comparison of the reported statistics for datasets and outcomes where applicable. A study conducted by Silverstein et al. [1999] remains one of the largest of its kind to date, utilizing over a billion queries taken from AltaVista, the most prominent search engine in its day. We believe our work offers insight into themes not normally discussed in traditional query log analysis and instead looks into novel methods of capturing interesting behaviors, of both the users and the search engine itself.

Analysis conducted for this investigation can be furthered using more specialized methods. Spam removal algorithms based on machine learning techniques may be utilized to supplement existing filtering techniques, including integration of features based on natural language, and stopwords to remove undesirable queries. Similarly, such measures can used to model inter-query similarities at a semantic level. Furthermore, anomaly detection may be expanded to verify the consistency of session sizes and associated users, although this may be a much harder task due to the anonymized nature of most query log data.

6. CONCLUSION

Our experiments on the MSNSearch dataset have shown that our previous implementation of query-based filtering may have removed large segments of queries from within sessions, and affected our session-based analyses in which we examined continuity characteristics between queries in multi-query sessions. By comparison, session-based filtering implemented in a similar manner resulted in more realistic output, but with the potential risk of additional non-human-origin queries being integrated into the analysis.

We also looked into possible inconsistencies within the clickthrough data, by utilizing individual clickthroughs as snapshots to gauge modifications in query/URL pairings over the collection period. Although we discovered there are in fact inconsistencies within the data manifesting as anomalous ranking changes (possibly resulting from the use of faulty scripts at the time the data was generated), such inconsistencies were small in number and would not greatly affect any conclusions drawn from the data. Regardless, care should be exercised when using the affected data in its current form for future activities.

Acknowledgements

Microsoft Research provided query logs described in this paper, as well as associated funding, via their "Accelerating Search" Live Labs research incentive.

References

Bernard J. Jansen and Amanda Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42(1):248–263, 2006. ISSN 0306-4573.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. ISSN 0163-5840.

Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001. ISSN 1532-2882.

Yuye Zhang and Alistair Moffat. Some observations on user search behaviour. In *Proceedings of the Eleventh Australian Document Computing Symposium 2006*, pages 1–8. Queensland University of Technology, Australia, 2006.

Interaction Pool: Towards a user-centred test collection

Hideo Joho
Department of Computing Science, University of Glasgow, UK.
hideo@dcs.gla.ac.uk

Robert Villa
Department of Computing Science, University of Glasgow, UK.
villar@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science, University of Glasgow, UK.
jj@dcs.gla.ac.uk

ABSTRACT

The advance of evaluation methodology is essential for the development of interactive systems that are based on the understanding of information seeking behaviour. This position paper presents a (rough) design of a community-based approach called the *interaction pool*, a repository of annotated interaction data that can be harnessed and shared by a research community interested in information seeking behaviour, interaction design, interface engineering, and realistic system evaluation. The design of such a repository was motivated by the need to develop a user-centred test collection which inherited the advantages of existing system-centred test collections while considering the characteristics of user-centred research and development.

1. INTRODUCTION

Evaluation of interactive systems and measuring their effects on information seeking behaviour are challenging. The comparison of different interface designs and interactive support systems are even more challenging. In Information Retrieval (IR), common test beds, called *test collections*, have been created and shared by the IR community, being used for extensive testing and comparison of retrieval algorithms over some decades.

While existing test collections have been an important asset for IR research, they are mainly designed for algorithmic evaluation, thus, user interactions and contexts of search are often simplified. Such a test collection is referred to as a *system-centred* test collection in this paper. This position paper is concerned with the design of test collections such that user interactions and search contexts are captured as part of the resource and shared by a community. We will refer to this as a *user-centred* test collection. We believe that such a test collection can facilitate the comparative evaluation of interactive systems and information seeking research while inheriting the advantages of existing approaches.

To maximise the benefits of a user-centred test collection, it is important to obtain feedback from the researchers in the relevant areas. For example, during the design of early test collections, Sparck-Jones and Van Rijsbergen [5] carried out a study to elicit the properties of test collections. While the specification of a test collection is not the main focus of this paper, we hope that this paper will set a tentative ground to discuss the properties of a user-centred test collection.

The rest of the paper is structured as follows. Section 2 summarises how existing test collections work and highlights their advantages and limitations. Section 3 presents a design of *interaction pool* which constitutes a central part of a user-centred test collection. Section 4 illustrates how the interaction pool can potentially facilitate the research on interactive systems and information seeking behaviour. Section 5 discusses several issues that are open for discussion in the context of a user-centred test collection. Finally, Section 6 concludes this paper.

2. SYSTEM-CENTRED TEST COLLECTION

A test collection usually consists of a document corpus, a set of topics, and a list of documents that are relevant to each of the topics (called *qrels*). The document corpus tends to be a static collection so that the performance measures are not violated by content changes. A topic is a description of a searcher's information need. A participant of a system-centred test collection then indexes the document corpus, performs a retrieval using the topic descriptions, and finally, submits the top N[1] ranked documents to the organiser. A *document pool* is then formed by using the top M[2] documents submitted by each of the participated systems (See Figure 1). The assessor of a topic judges the relevance of documents in the document pool, which become qrels.

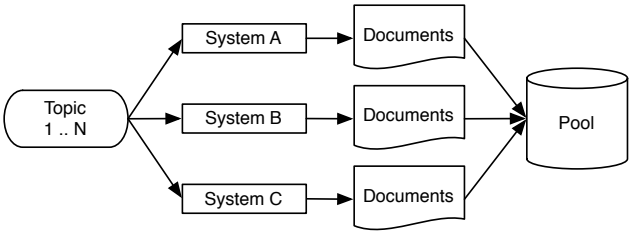

Figure 1: Pooling of documents.

The advantages of this approach is that participating systems use a common set of documents, topics, and qrels,

[1]e.g., 1000 docs
[2]e.g., 100 docs

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

which makes the comparison among the systems fair and more reliable than tests performed in different conditions. By pooling the documents retrieved by different ranking algorithms, bias towards a particular system is minimised in the evaluation. This also makes it possible to assess a future system using the existing resource. Therefore, the use of a common data set and a pooling method is inherited and assumed in our design of a user-centred test collection.

From the interaction point of view, however, the data stored in a system-centred test collection is the minimum set of interactions where a user submits a query and a system returns a set of (ranked) documents in response to that. The document pool, therefore, stores and evaluates the *outcome* of single search iterations harnessed by participants. However, in a study of interactive systems and information seeking behaviour, the *process* and *context* of search are of great interest. For example, search is often an iterative process which uses multiple queries and browsing of documents. Furthermore, context influences how a search session is developed and how document relevancy is perceived by searchers [3]. A system-centred test collection is not designed to store such data, although effort has been made to elicit some of the contexts inherent in test collections [1].

Another significant property of a system-centred test collection is that document relevancy is determined by a single assessor (who is often a topic creator). This is related to the lack of interaction in the design of system-centred test collections. As discussed above, the relevance of documents can vary over searchers and search contexts. In a user-centred test collection, therefore, the data should contain the document relevancy perceived by different searchers and different contexts. The interaction pool discussed in the next section is designed to address these issues of existing test collections.

3. DESIGN OF INTERACTION POOL

An interaction pool (See Figure 2) is an extension of the document pool where multiple iterations of search are stored. The interaction data such as the queries submitted by users, retrieved documents, click-through documents and their rank positions, next / previous result page viewing actions, are populated along with a timestamp in the interaction pool. Similar to a document pool, the interaction pool contains a range of interaction paths that would be recorded in participated studies which might use different search engines, interfaces, and support systems. This enables researchers to study the process of search harnessed by participants. As such, an interaction pool constitutes a central part of a user-centred test collection.

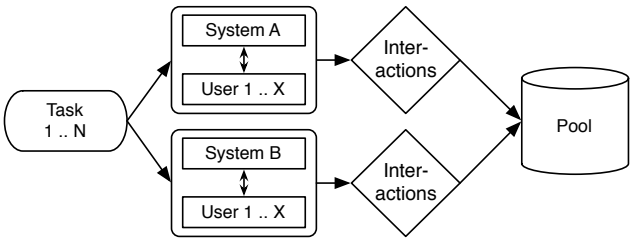

Figure 2: Pooling of interactions.

Participation in a user-centred test collection might occur as follows. First, a set of work/search tasks are defined. Participants carry out a user study using their own choice of systems and the given tasks. The interaction logs are recorded during the study and the data is submitted to the organiser to populate the interaction pool.

Another component considered in the design of an interaction pool is the search metadata. The metadata can consist of a work/search task description (providing a context of search as opposed to a description of what is relevant or not), a user's background, search contexts, system/interface descriptions, subjective assessments, and other information that allows us, for instance, to cluster the interaction data for a granular analysis of people's searching process.

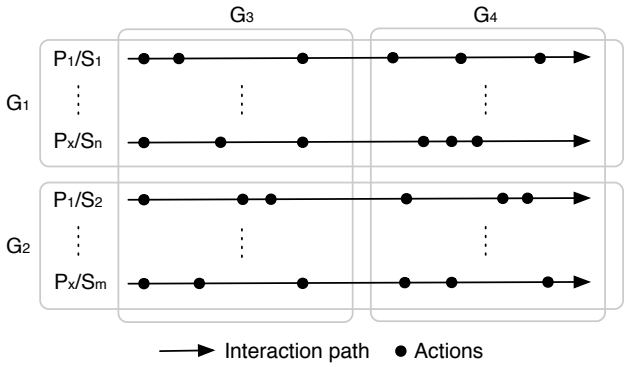

Figure 3: Grouping of interaction data.

Figure 3 illustrates the examples of grouping the interaction data. P denotes a participant ID and S denotes a search session ID. P_1/S_1 to P_x/S_m are a set of search sessions based on a task populated by participants. The horizontal arrows represent an interaction path where the dots indicates user actions occurred in the path. The first case (G_1 and G_2) groups the search sessions into two categories based on a facet or context of search environments. The facet/context can be anything as long as it can be extracted from the metadata (e.g., a user's role in an organisation, level of familiarity/interest with a search topic, search device used). The second case (G_3 and G_4) divides the interaction paths into two different stages of search sessions. In this way, one can analyse the search behaviour at the beginning to middle and middle to the end of search. These are just two examples and other usages of the interaction pool entirely depends on research interests.

Table 1: Aggregate relevance judgements

	Single	Interaction pool			
Doc	assessor	P_1/S_1	P_1/S_2	...	P_x/S_m
D_1	Rel	Rel	Rel	...	Rel
D_2	Non-rel	Non-Rel	Rel	...	Non-Rel
⋮	⋮	⋮	⋮	⋱	⋮
D_n	Rel	Non-Rel	Non-Rel	...	Rel

The interaction pool can also be used to store multiple relevance judgements of retrieved (or click-through) documents. In system-centred test collections, relevance assessments are usually carried out by a single assessor for each topic (See Table 1). In the interaction pool, however, document relevancy is no longer uniform and can vary across

the populated search sessions. Similar to the grouping of interaction data, the aggregated relevance judgements give us an advantage of investigating the effectiveness of interactions and systems from different facets/contexts. Since not all documents are retrieved by every sessions, the annotations such as *shown*, *clicked*, *rel/non-rel* can be associated with retrieved documents.

4. BENEFITS OF INTERACTION POOLS

The previous section described a sketch of an interaction pool that can be harnessed and shared by participants of a user-centred test collection. This section illustrates how such a resource can potentially facilitate the research on information seeking behaviour, user interface/interaction design, and system evaluation.

Information seeking behaviour.

The interaction pool can offers an opportunity to verify existing information seeking models that might have been developed through an ethnographic study. Researchers can access to the pool to analyse whether or not a modelled behaviour can be found in the interaction of various systems, users, and search contexts. While the interaction data in the pool are likely to be based on controlled environment experiments, the search metadata should reduce the level of uncertainty involved in the interpretation of information needs behind the seeking behaviours, compared to, for example, the analysis of search engines' query logs.

In a different scenario, the pool itself might become the rich source of investigation. For instance, one can attempt to mine behavioural patterns from the interaction data. As illustrated above, it will be easy to partition the data based on the annotated metadata, or re-organise the data set to highlight a certain facet/context of search.

Interaction/interface design.

For those who are interested in evaluating the usability or effectiveness of a new search interface, the interaction pool offers realistic user input for benchmarking, whether a user study or simulated study [4] is carried out in the investigation. For example, researchers can extract real queries formulated by the users of the interaction pool and use them as the input to a simulated study of a new interface. Since the users are likely to have a different interpretation of the information need of a given task, their queries are more realistic and diverse than those arbitrary formulated from a task description. The click-through documents in the interaction path can also be exploited as a user feedback trail or path during the task. Overall, the interaction pool can provide extra information for more realistic and controlled simulation of users in the study.

When a user study is conducted independently, the results of the new interface can be compared to the average performance obtained from the interaction pool, or compared to a particular set of search sessions selected by the facets/contexts given in the metadata. For example, one can measure if the new interface allows users to complete a task faster than the average performance in the pool. The subjective assessments can also be compared to other participants' data. When the interaction data in the pool can be used as a baseline performance, then participants can reduce the resources required to carry out a user study (e.g., time and number of subjects). Given that a user study tends to be an expensive process, the interaction pool can reduce evaluation effort. Like a system-centred test collection, we would expect that the experimental resources such as the tasks, document collections, and user's interaction data can be re-used by or support a future interactive IR study.

System evaluation.

The interaction pool offers a new challenge for those who are interested in system evaluation. While a system-centred test collection is designed to determine, for example, if System A is better than System B based on N *topics*, a user-centred test collection is designed to find the difference between the two systems based on N *contexts*. In particular, the notion of uniform document relevancy is no longer compulsory. The interaction pool allows researchers to control how document relevancy is determined by a given facet or context of a search environment.

In the simplest example, the qrels of a task can be generated as many individual search sessions, and the performance of systems can be measured by those individual qrel sets. When a certain facet/context is given, aggregated relevance judgements can be used to measure the system performance for contextual relevance. Since the path of user actions is stored in the pool, one can test the performance of relevance feedback techniques based on a range of interaction patterns.

5. OPEN ISSUES

The requirements and specifications of a user-centred test collection are still under developement. As such, there is a number of open issues. The following are some of the issues that emerged from the preparation of this paper.

Legal/Ethical issue Sharing interaction data imposes an additional element to consider when legal and ethical issues are concerned. This might be as simple as adding a section in a consent form noting that the collected data will be anonymised and shared by the research community. The issue might be more complex for industrial participants. A collective effort needs to be made by the community to share the data since conditions may vary across countries and companies.

Research assets When the interaction data constitutes a fundamental asset in a study, it is conceivable that researchers are not willing to release such data to the community immediately. We need to consider how to achieve a win-win situation for the participants of test collection. Needless to say, participation in an interaction pool means that researchers can access a potentially large quantity and diversity of annotated interaction data which might be infeasible to obtain by a single researcher or research group.

Document collection In existing system-centred test collections, a static document collection (.e.g., web corpus) is often offered to participants. A static collection allows researchers to measure the performance of systems without the effects of content change. On the other hand, participants are responsible for indexing a common document set provided by a test collection. This might be too much effort for, or at least not the

main interests of, some of the participants of a user-centred test collection. An alternative choice is not to have any restrictions on the selection of document collections. Participants can use a search engine's API, for instance, to develop a new interface. In this case, we would need to devise the performance measures that are independent of document collections.

Work/Search tasks It is generally believed that studying people's searching behaviour in the context of tasks (e.g., work task or search task) is beneficial [2], and that a simple description of the search aim by asking users to find as many relevant documents as possible is not realistic. To attract many researchers to participate and contribute to the population of an interaction pool, we need to devise a set of work or search tasks that are realistic and interesting to the research community.

Annotation scheme Participation in a system-centred test collection such as TREC[3] is facilitated by the simple annotation adopted in the data submission. While the data in the interaction pool requires a more complex annotation scheme than a list of document IDs, we should aim to define a standardised scheme which is as simple as possible. A related issue is to formulate a core set of metadata and actions that need to be recorded for the population of an interaction pool.

Infrastructure When the core set of metadata and interaction data is defined, and an annotation scheme is specified by a community, then we would expect to have a repository server which enables participants to access to the pool through some sort of API.

Performance measures In a system-centred test collection, the performance of ranking algorithms is typically measured by precision, recall, and their variants. It is still not clear what performance measure is appropriate for interactive systems and their effects on information seeking behaviour. However, a user-centred test collection has the potential to employ the measures based not only on retrieval effectiveness (e.g., precision/recall) but also on interactions (e.g., number of actions, time to complete a task, etc.) as well as subjective assessments (e.g., "Would you use it if it's available on the web?").

An approach to establish the performance measures for a user-centred test collection can be to analyse the central dependent and independent variables frequently investigated in existing interactive IR studies.

Scale of data There are unknown properties in the current design of interaction pool: how many participants are needed to achieve a meaningful interaction pool; how many tasks should each participant carry out; how many subjects should each participant recruit for populating the pool. While the size and diversity matter in our design, a continuous co-ordination by a community is essential for the development of a successful test collection.

[3] `http://trec.nist.gov`

6. CONCLUSION

This position paper discussed a design of *interaction pool* aiming towards the development of a user-centred test collection. We illustrated how such a resource can support the evaluation of interactive systems. A number of open issues were also discussed. This paper is intended to stir the discussion of evaluation methodology as opposed to presenting a precise specification. We believe this workshop is an ideal forum to discuss such issues.

7. ACKNOWLEDGEMENTS

The authors thank to the anonymous reviewers for their constructive feedback on the paper. This work was supported by EPSRC (Ref: EP/C004108/1), and EU IST FP6 projects (Ref: 033715 (MIAUCE), 027122 (SALERO)). Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] J. Allan. HARD track overview in TREC 2005 high accuracy retrieval from documents. In E. M. Voorhees and L. P. Buckland, editors, *NIST Special Publication: SP 500-266, Proceedings of The Fourteenth Text REtrieval Conference*, 2006.

[2] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[3] L. Freund, E. Toms, and C. Clarke. Modeling task-genre relationships for IR in the workplace. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 441–448, 2005.

[4] H. Keskustalo, K. Järvelin, and A. Pirkola. The effects of relevance feedback quality and quantity in interactive relevance feedback: A simulation based on user modelling. In *Proceedings of the 28th European Conference on Information Retrieval*, pages 191–204, London, UK, 2006. Springer.

[5] K. Sparck-Jones and C. J. van Rijsbergen. Report on the need for and provision of an 'ideal' information retrieval test collection. Technical Report British Library Research and Development Report 5266, University Computer Laboratory, Cambridge, 1975.

Using Subjunctive Interfaces to Show Web Retrievals in Context

[Extended Abstract] *

Aran Lunzer
Meme Media Laboratory
Hokkaido University
Sapporo 060-8628, Japan
aran@bigfoot.com

ABSTRACT

A wide range of applications deliver information from the Web in response to users' requests. Many of these applications can deal with information needs that are only vaguely expressed, compensating for the wealth of potentially relevant results by ranking or filtering them to present a manageable subset. However, in many cases the criteria for this ranking or filtering are not revealed. Thus users are in the unsatisfactory position of having to accept on faith that the presented information is, by some opaque measure, the best available. We propose an approach, based on subjunctive-interface techniques, to extending such applications so that information is delivered along with context that makes it easier to judge the information's quality.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*input devices and strategies, interaction styles*

General Terms

Design, Human Factors

Keywords

Subjunctive interfaces, Web search, Web applications

1. INTRODUCTION

Recommenders, Web search engines, and context-aware systems are examples of applications that deliver information based on criteria actively or passively supplied by a user. In many cases, the volume and variety of information that the application could deliver is so large that, even if the user has specified some filtering criteria, the application must still apply further criteria to constrain what is delivered. A search engine such as Google, for example, may discover a million pages that match a user's keyword query; it would be unhelpful to try to present them all, so the engine applies criteria to rank the pages such that those estimated as being especially useful are presented first.

In terms of enabling users to benefit from the delivered information rather than be overwhelmed by its quantity, such constraints are vital. But the fact that users are routinely ignoring the vast majority of available results, without any means to check the quality of those results relative to the ones that are seen, can be regarded as a problem in at least the following respects:

1. An application's standard criteria for favouring some results over others might not fit the user's current needs.
2. The user may therefore end up accepting results that are substantially inferior to others that are available.
3. Nonetheless, the user may believe that the obtained results are the best.
4. Even if the user realises that alternative results may be better, he or she will feel powerless to go in search of them.

Clues as to how to address this problem are found in some existing retrieval applications. For example, many travel-booking sites now automatically deliver not just results that exactly match the user's query, but also results from nearby queries, such as for the preceding or following date, or for an airport in a neighbouring city. Informally we can say that those additional queries are providing context for the precise-match results; in many cases such context can help a user to judge result quality, such as by revealing that the lowest fare on the originally specified date is high compared to other days.

Our goal is to encourage the provision of this kind of result context in retrieval applications in general. Especially for

*A full version of this paper, entitled *Using Subjunctive Interfaces to Put Delivered Information into Context*, appears in K.P. Jantke, R. Kaschek, N. Spyratos and Y. Tanaka (eds.), *Knowledge Media Science: Preparing the Ground*, to be published in 2007 in the Lecture Notes in Artificial Intelligence series.

applications that normally rely on hidden criteria to produce and rank their results, we believe that context illumination based on a range of alternative criteria could help users to make informed judgements about result quality.

The work reported here moves towards that goal by proposing interface techniques that would allow the display of result context to be generalised. Our approach is based on subjunctive-interface techniques [1, 2], which allow alternative execution scenarios for a given application, based on alternative values for its inputs, to be set up, viewed and manipulated in parallel. We believe that such parallel scenarios offer a simple yet flexible way to reveal context.

2. A SUBJUNCTIVE-INTERFACE APPROACH TO RESULT CONTEXT

In this section we explain how subjunctive-interface techniques could be used to provide context for results, then show an example built using our prototype system.

2.1 Concept

The essence of a subjunctive interface, as outlined above, is that it provides facilities for handling multiple scenarios in parallel. Our intention is to use these facilities to support a simple form of result-context delivery, through what we will call a subjunctive result explorer.

For example, given an application that finds the lowest price of a flight from one city to another on a specified date, a simple form of context can be provided by automatically setting up two additional scenarios to handle enquiries for the day before and the day after. Moreover, the flexibility of a subjunctive interface allows for arbitrary definitions of the contextual information: instead of the day before and day after, the interface could show results for a week earlier and a week later, or for a range of consecutive dates, or even for a range of alternative departure cities or destinations (though automatically generating suitable alternative cities is clearly not as straightforward as generating alternative dates).

Then, through the basic mechanisms of a subjunctive interface, the subjunctive result explorer will produce visual displays that help the user to understand which results arise from which scenarios.

In the following descriptions, the user's initial request to a subjunctive result explorer is referred to as the primary specification. Secondary specifications are then added through the operation of the explorer, driving secondary retrievals that provide context for the primary retrieval's results.

2.2 Basic requirements

In this section we consider various conditions that must be met when developing a subjunctive result explorer for some application.

2.2.1 No external side effects to result generation.

This obvious but vital condition is what makes it possible to offer unhindered exploration of multiple results. For example, scenarios evaluated in exploring result context at an online shopping site must not go so far as to complete a purchase.

2.2.2 Comparisons of like with like.

A constraint arising from the use of subjunctive-interface techniques for the result explorer is that the primary and secondary retrievals should be of the same form: that is to say, all specifications are in terms of values for the same set of parameters, and all deliver results of the same kind. If a user insists on investigating widely different types of retrieval, the prerequisite to tackling this with a subjunctive interface would be to express those retrievals using a common set of inputs and results. For example, a user wanting to compare flight itineraries against package tours, or car journeys, would have to do so in terms of result values shared by all cases, such as the journey's estimated cost and duration.

2.2.3 Proactive offering of context.

The greatest benefits of displaying context are likely to arise when such information is added autonomously by the system. A key challenge in such autonomous support is to generate alternative retrievals whose processing provides context that is meaningful for the user. In general it is unlikely, for example, that random variation of ingredient values would be helpful. What we do expect is that each application will have some standard forms of variation that are known to have a high probability of delivering interesting results; these should be encoded as executable rules for generating secondary retrievals based on the user's primary specification. That said, rather than have the system immediately apply those specifications to generate extra results, in some cases it may be more appropriate just to make suggestions that the user is then free to pursue or ignore. This is related to the following point.

2.2.4 Safeguards against data deluge.

Proactive generation of context brings a risk of creating a flood of secondary retrievals. Such a flood is not only potentially expensive in its use of computing resources, but can lead to overwhelming numbers of results, from which it is hard for a user to gain any insight. One approach to reducing this risk is to give the user control over which secondary retrievals are run. Another is to apply filtering techniques at various stages of the processing (discussed in the full version of this paper).

2.2.5 An interface that makes clear the relationships between retrievals.

The user must be supported in understanding which alternative retrievals have been evaluated, and what results were delivered by each retrieval. We believe that the RecipeSheet [3, 4], our current platform for research on subjunctive-interface techniques, can meet this requirement. A recipe sheet offers facilities analogous to those of a spreadsheet, allowing a user to set up a graph of cells connected by recipes. Recipes replace the standard spreadsheet's use of formulas for deriving cell values, and have the property that the processing for a recipe is itself a dynamically supplied ingredient (i.e., an input). The RecipeSheet can therefore provide uniform handling of variation in both inputs and processing, which we believe will be needed in many applications to allow useful ranges of queries to be processed and displayed.

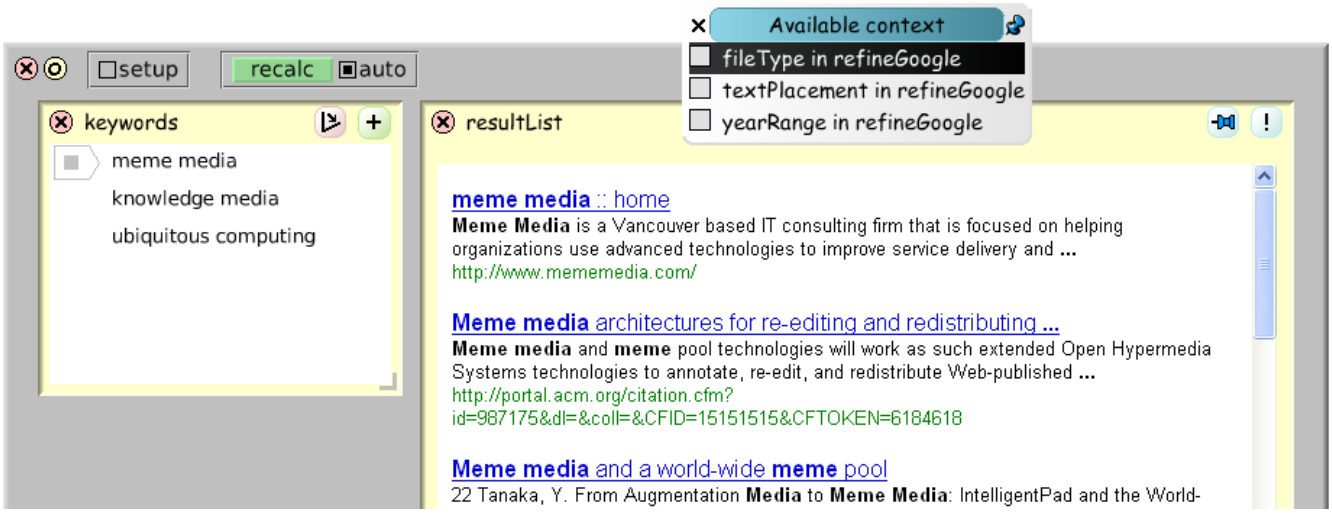

Figure 1: User asking for context-expansion suggestions on the results of a Google search for 'meme media'.

2.3 Application examples

We have started to test out these ideas by building simple application examples using the RecipeSheet.

Figures 1 and 2 demonstrate the addition of context to results from a Google search. Having run a standard search, the user probes the result list for suggestions on how to obtain context information. The system reveals three ingredients that may be worth exploring – relating to file types, the position of keyword terms (e.g., in a page's title, text, or URL), and the inclusion of particular year numbers. Figure 2 shows the situation after the user chooses to explore alternative year ranges[1]. In the setup seen here, a merged display of the top ten matches from each search shows which pages turned up in which searches. One potential benefit of running multiple searches with alternative year ranges is seeing how the usage of a term on the Web has changed over time; another is reducing the risk that the results will be swamped by matches from some narrow time range.

A further application, suggested by our bioinformatics colleagues, relates to the parameter settings commonly used in tools for manipulating genome and protein sequences. Tools such as BLAST or ClustalW include a number of parameters that can be set when submitting a job, but through experience it has become customary for bioinformaticians to use fixed values for these parameters on the basis that they give satisfactory results, and that trying alternative values does not usually reveal particularly interesting variation. However, such habits can lead the scientists to fail to notice those situations where supplying an alternative value would in fact lead to significant changes. In the case of BLAST, such changes are especially likely when performing searches across species. A test in which we performed a homology search for a human protein against a fruitfly database showed that the top twenty matches found using the BLOSUM62 weighting matrix contained only fourteen of the top twenty found using the matrix PAM30. We are now considering what approach to secondary retrievals would be effective in alerting users to such differences.

[1]The example shown here is based on adding a numerical-range specification to the search keywords. There is no guarantee that a search using a range such as '1990..1999' returns only pages relevant to those years, but in practice most Web pages mentioning a number around 2000 are, not surprisingly, referring to a year.

3. DISCUSSION

One way to characterise this work is to say that we are trying to help users obtain visualisations that offer insight into the variety and properties of results in a given domain. Building these visualisations in general requires the cooperation of several processing components, often specialised to the domain or individual application. In that sense, what we will be providing is a framework onto which developers and end users will add components suited to particular applications. Our challenge, therefore, is to ensure that the framework strikes an appropriate balance between powerful flexibility and ease of use.

This framework for building subjunctive result explorers itself relies on the services of the RecipeSheet. As it stands, the RecipeSheet already provides a usable base for setting up, viewing and manipulating scenarios, but it will need to be extended in various ways to provide the general flexibility that is needed. One pressing need that we recognise is to overcome the current limitations on the numbers of scenarios that can be handled, in particular given that in the present model every cell on a sheet is required to participate in every scenario. We are developing a new calculation model and interface facilities to support alternatives, and hence scenarios, that are localised to a single branch of the inter-cell dependency graph.

Once the framework is established, and result browsers are built for various applications, we can start to evaluate their impact on users' result-seeking behaviour. Our hypotheses include a number of potential benefits. The easiest kind of benefit to observe would be cases where automatic generation of secondary results causes users to notice valuable regions of the result domain that other users tend not to ex-

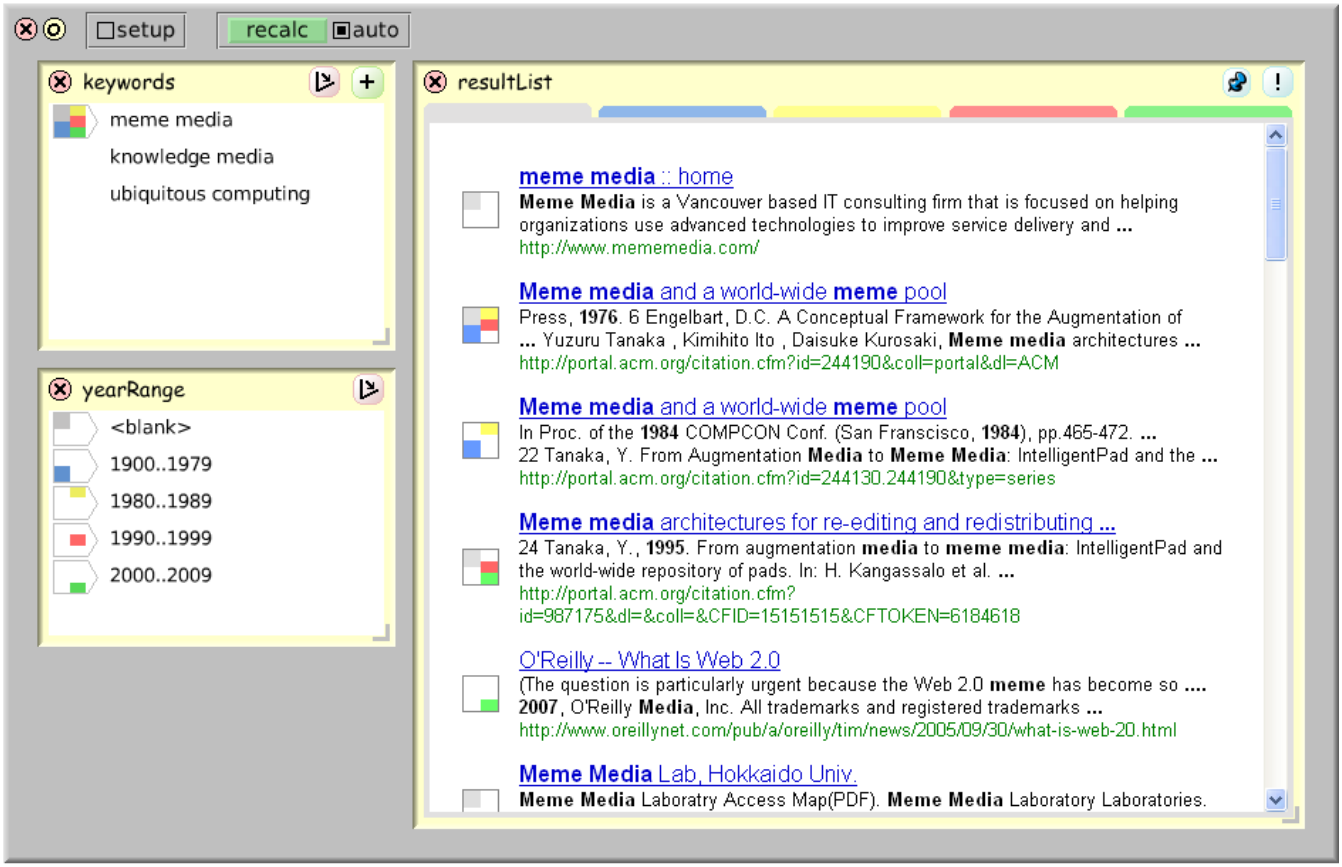

Figure 2: Google search for 'meme media' evaluated five times using alternative year ranges (including the default search, with no range). The five corresponding top-ten result lists have been interleaved, with each result marked to show which lists include it. In this case the third and fifth results, although receiving high rankings in the 1980s and 2000s respectively, do not appear in the default list at all.

plore. A benefit that would be harder to detect is where the system generates a range of results that turn out not to differ significantly from each other (e.g., a range of travel dates turning out to have identical fares); the benefit here lies in reassuring the user that, at least in terms of the displayed properties, no choice is worse than any other.

For reasons such as the above, it will certainly be easier to demonstrate benefits of subjunctive result exploration in some domains than in others. However, if users merely become more conscious of the fact that delivered information tends to sit within a context of nearby alternatives, we will have helped them to understand the power and the limits of the Web retrieval tools they use.

4. REFERENCES

[1] A. Lunzer. Benefits of Subjunctive Interface Support for Exploratory Access to Online Resources. In: G. Grieser, Y. Tanaka (eds.) *Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets. LNAI 3359,* pages 14–32. Springer, 2004.

[2] A. Lunzer and K. Hornbæk. Usability studies on a visualisation for parallel display and control of alternative scenarios. In *Proceedings of AVI 2004*, pages 125–132. ACM Press, 2004.

[3] A. Lunzer and K. Hornbæk. An Enhanced Spreadsheet Supporting Calculation-Structure Variants, and its Application to Web-Based Processing. In K.-P. Jantke, A. Lunzer, N. Spyratos and Y. Tanaka (eds.) *Proceedings of the Dagstuhl Workshop on Federation over the Web,* May 2005 (Lecture Notes in Artificial Intelligence, Vol. 3847), pages 143-158, 2006.

[4] A. Lunzer and K. Hornbæk. RecipeSheet: Creating, Combining and Controlling Information Processors. In *Proceedings of ACM UIST 06,* pages 145-153. ACM Press, 2006.

Naming the Topic or Reversing Query Terms from Result Documents – Successful Strategies in Web Search

Anne Aula
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043 USA
anneaula@google.com

ABSTRACT
Numerous studies show that the strategies of most of the web searchers are very simple: they use short queries without operators and modifiers or make mistakes with them, and they mostly rely on the first page of results returned by the search engines. However, the question of whether the users are successful with these simple strategies has received less attention. This paper describes strategies that web searchers have for query formulation and results evaluation and focuses specifically on factors that affect the success of these strategies. Based on the understanding of the limitations of the users' strategies, the paper presents ideas on how search engines could more effectively support the users in the information search process by engaging the users in a dialogue-like interaction.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – *search process.* H.5.2 [**User Interfaces**]: User-Centered Design. H.5.3 [**Information Systems**]: Online Information Services – *web-based services.*

General Terms
Design, Human Factors.

Keywords
Web Search Engines, User Search Strategies, Search Process.

1. INTRODUCTION
Studies show that the users mostly employ very simple strategies to web search. They formulate short queries with one or two query terms and seldom use Boolean operators or term modifiers, they often only submit one query during the search session, and they re-access information mostly by using browser bookmarks or re-searching information with search engines [5][10][11][12][13]. However, there is no clear understanding on how successful the users are with these simple strategies.

PC magazine recently posted a list of tips on how to become a more successful web searcher. The list stated, among others, that it is beneficial to be use quotation marks around the query, to use a minus sign to exclude terms, and use plain English in queries[1]. Another tip was recently given by teenagers who told that at school, they are taught to add a '+' sign in between query terms to make the queries more precise on Google. Similar (albeit more accurate) tips are also given on the Google's help page[2]. Among other tips, the users are told that it is good to be precise when formulating queries and that there are several different operators and modifiers they can use to improve their success in search. All these tips seem to suggest that the users have problems with their simple approach to web search. What if they started using the strategies the tips suggest – would they be more successful?

Defining success in web search is a challenge that is yet to be overcome. In our user studies with predefined search tasks, we have found it useful to consider both the outcome of the search (did the users find what they were asked to find) and the resources (typically time) needed to complete the search. Along these lines, we have used *task completion speed* (number of tasks completed divided by the time it took to complete the tasks) as an approximation of the users' overall success or level of expertise [8]. Using task completion speed to differentiate between less and more successful searchers requires that the tasks are the same for all the participants. Thus, this metric is mostly useful in controlled user studies.

This paper will discuss the factors that affect the success of the users' strategies based on a number of published and unpublished user studies. The examples illustrate the differences in the users' strategies in two different phases of the search process; query formulation and results evaluation. After presenting the different strategies and discussing their success, the paper proposes that search engines should take a more active role in trying to discover the user's underlying information need instead of only relying on their initial queries.

2. METHODS
This paper is based on user studies conducted in the course of a PhD thesis on user strategies for web search [3] and the author's experiences in conducting a number of unpublished user studies on web search at Google. The unpublished user studies include approximately 60 individual think-aloud usability test sessions.

In the published studies, the methods varied from think-aloud lab studies [8] and controlled experiments [2] to surveys [1][5][9], field usability studies [4][6], and eye-tracking experiments [7].

3. SUCCESS OF THE STRATEGIES
In the following, I will give examples of successful and less successful strategies in two different phases of web search: query formulation and results evaluation.

[1] http://www.pcworld.com/printable/article/id,130979/printable.html

[2] http://www.google.com/help/basics.html

3.1 Query Formulation and Refinement

In line with the log studies, our user studies have also shown that generally, the queries the users formulate are simple: most often, they only contain a couple of query terms and Boolean operators and other modifiers are seldom used. Furthermore, when the users formulate queries with operators and term modifiers, they often use them unnecessarily or make mistakes with them. Given these simple queries, the question is: are they enough for the searchers to find what they are looking for?

Based on our studies, the answer is: it depends. In certain tasks, a broad query with only a couple of query terms is all that is needed, but this strategy does not result in efficient performance in other tasks. For example, if the task is to find how much blood human heart pumps in one minute it is a much more efficient strategy to formulate a query *human heart pumps blood per minute* than a query *human anatomy* [8]. Similarly, when the task is to find the time difference between New Zealand and your present location, the likelihood of finding the answer quickly is much larger with the query *new zealand time zone* than with *new zealand* (examples from an unpublished usability study).

We have called these two different approaches to querying the **straight-to-information approach** and **encyclopedia approach**, respectively [1][8]. In the encyclopedia approach, the users generalize the terms from the task description (we have also observed this with the users' own tasks). These users are using search engines like they were using a paper encyclopedia: they think of a general term that describes the topic and use the search engine as an index for finding sites that are related to the topic. To find the answer to the original question, they browse to the needed information, which is often a tedious process. On the other hand, the users who employ the straight-to-information approach want to minimize the browsing: they "reverse the query terms from the documents" [5] and aim at finding the answers already in the snippets.

In the examples presented above, the users with the straight-to-information approach were more efficient in finding an answer to the tasks. However, a simple "longer and more precise is better" relationship between the queries and the success of the search does not always hold. In a usability study (unpublished), we gave the participants the task "Is there a king in Finland? If yes, what is his name?" In this case, the approach of using the specific terms from the task description (*e.g.*, king of Finland name) results in keyword based search engines returning results that all contain the terms *Finland*, *king*, and *name*. In fact, there was a king in Finland (although Finland as an independent nation did not exist back then), so the results are about the history of Finland. In this case, a successful query is broader than the original task description, *e.g., Finland* (on Google, this query returned a Wikipedia result that contained the needed information) or one that generalizes the term *king* to *government*.

Some users seem to be predominantly using either the straight-to-information or the encyclopedia approach, but the most successful users are flexible. They sometimes formulate general queries with one or two terms (tasks with a general informational goal or when they do not know good terms to use in the query) and at other times, their queries that are much longer; up to over 10 terms (fact-finding tasks where recall is not too important or when there is one known name for the document they are looking for).

Oftentimes, the first query does not provide the users with satisfactory results and the query needs to be refined. In this phase of the search process, the behavior of less and more successful searchers is again very different. Where the more successful searchers often change or modify a word or two from their original query and keep the original intent of the query similar, the less experienced users often abandon the original query completely and change the query, for example, from *energy food* to *nutrition* and then to *how much energy contain*. A more successful searchers, on the other hand, are likely to refine only parts of the query; from *children learn to walk at age* to *children learn to walk at age references* to *typically children "learn to" walk age*. In addition to these differences, the less successful searchers often have considerable difficulties in thinking of terms to use in the query if the original query was not successful: "I really do not know how I could make the search terms to be relevant for this task" (all examples from [8]).

3.2 Result Evaluation

In the query formulation phase, it seems that flexibility is the major aspect that differentiates between more and less successful searchers. Do the more successful searchers also modify their strategies in the result evaluation phase?

Log studies have shown that web searchers mostly rely on the first result page given by the search engine and if they are not happy with the results, they are more likely to refine the query than go to the next result page. Furthermore, studies have suggested that the users generally only look for a couple of results before making their selection [11].

In our user studies, we noticed that searchers sometimes spent a lot of time on the results page before selecting a result or refining the query. To understand what is happening during that time, we run an eye-tracking study [7]. This study showed that some users are very careful in evaluating the results before making the decision to refine the query or click on a result. We called these searchers "exhaustive evaluators". Some users, on the other hand, seemed to be much more efficient in making their decision to either refine the query or click on one of the results ("economical" evaluation style). Our study suggested that the economical evaluation style was more common among the more experienced computer users – this style also seemed to result in more successful task performance in certain cases (Figure 1).

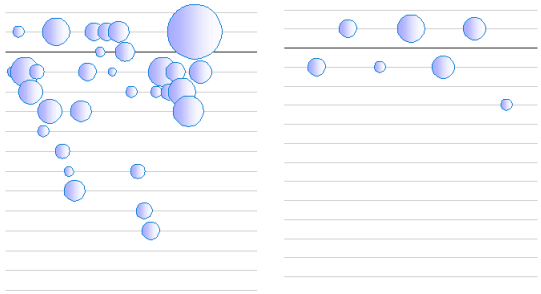

Figure 1: Scanning patterns of two users on the same search results page – exhaustive evaluation style on the left and economical evaluation style on the right. The size of the dot represents the time the user spent looking at the query box (above the black line) or one of the results (below the black line). The x-axis represents the order of scanning. In the end, both of the users decided to refine the query.

Anecdotal evidence from usability studies supports these findings. Specifically, it seems that the more successful searchers are very

fast in their evaluation of the results. If they see a relevant-looking result, they click on it or if they feel that the results are not relevant (either on the basis of the first couple of results or a very fast scanning of the complete results page), they quickly refine the query. In addition, these studies suggest that these evaluation styles extend to the evaluation of the other web pages, as well. It seems that the more successful searchers do a very fast evaluation of web pages: if the page seems to be of low quality, they quickly come back to the search engine. On the other hand, the less successful searchers seem to have little understanding on the quality of web pages and they seem to give an equal amount of attention to all the web pages they go to.

4. SUPPORTING SEARCH

In the query formulation phase, a notable factor in the users' success is their flexibility: a two-term query that results in a highly successful search experience in one task can result in a tedious session with lots of irrelevant results and poor web pages when applied to another task. Thus, the most successful searchers formulate queries that are appropriate for the task – sometimes these queries need to have several terms and even special modifiers whereas at another times, one term might be enough. In the cases where the first query does not provide relevant results, the less successful searchers seem to spend a lot of time evaluating the (irrelevant) results and web pages, and they have considerable difficulties in thinking of ways to refine the query.

In the result evaluation phase, it seems that the more successful searchers use a similar approach independent of the task. Namely, they do not spend much time evaluating the individual results nor do they go through all the results before clicking the most promising one or refining the query. The same efficiency seems to hold when they are evaluating the web pages they go to. On the other hand, the less successful searchers act as if there was a high cost associated with the result selection – they carefully choose which one to click and they invest a relatively long time in exploring each web page they go to.

Knowing that some users have clearly inefficient strategies to searching - is there anything we can do to improve their search efficiency? An example illustrates how the traditional information seeking that is based on human-to-human communication differs from the current computerized way. Imagine going to a bookstore and asking the salesperson for a book about television. Most certainly, they will not go and get you some books to see if you like them, but instead, they will ask you questions. They will ask if you want tips for buying a TV, information about TV programs, technical information about TVs, or maybe something else. After understanding what the need is, they will go and find the most relevant books (maybe also including other criteria, such as the price, in their decision on what to bring you). In the current world of search engines, when the user asks for information about television, the search engine will return a bunch of results although it has no way of knowing what the users' underlying information need is. The search engine has information on what most of the users find useful, but that may or may not apply to the user in question.

Although there are certainly advantages in using a popularity-based approach to search result ranking, I propose that we might considerably improve the searchers' experience by bringing some aspects of human-to-human communication to user-search engine interaction. In human-to-human communication, both of the parties are active in the interaction by asking questions and providing answers. Similarly, search engines could be more active in finding out what the users really want.

Another example illustrates the power of communication in web searching. In usability studies, we often ask participants to use some time to search for information they are personally interested in. Many times, the users end up being stuck – not really knowing how to proceed with the task. They have clicked on all the promising-looking results on the result page, they have used a lot of time reading the web pages, and yet, they have not found what they are looking for. In these situations, the moderator often asks: "Is there something specific that you are looking for?" Most of the time, the user can easily and precisely describe their underlying information need – and their need is often much more specific than their original query suggests. After that, the moderator often asks: "Can you think of a way to search for that?" Surprisingly, this intervention alone often helps the users in their search. It is noteworthy that the moderator is not providing any tips or new query terms to the searcher, but she or he is simply asking the user to (re-)describe their information need.

Thus, one possible way to help the users would be to have the search engine take a more active role in finding out about the users' information need – a similar role that the moderator is playing in usability studies. For example, if the user types in a one or two term query and does not click on any of the results within a certain time limit (behavior that shows that the user is most likely a less experienced user and thus, might benefit from assistance), the search engine could ask the user questions about the topic. Simple questions or prompts, e.g., "Tell me more about the topic you are interested in" or "Is there a specific aspect of [query] you are interested in?" might be efficient in eliciting more query terms from the user – just as they seem to do in usability studies when the moderator asks similar questions. This information, in turn, could be used for delivering results that more closely match the users' **underlying information need** rather than their original query.

5. DISCUSSION

In the query formulation phase, successful searchers are flexible and modify their strategies according to the task from highly specific queries with operators to broad and simple two-term queries. One user explained his/her approach for fact-finding tasks as imagining how the sentences containing the information might be formulated in the result document [5]. Although this strategy works well in fact-finding tasks, it is not as useful in tasks where it is important to find as much information as possible about the topic. In these broader tasks, the searchers will do better if they generalize the terms from the task description (employ encyclopedia approach) or possibly by formulating more sophisticated faceted search queries with Boolean operators.

To return to the questions presented in the beginning of the paper - are users happy with the simple strategies they use and would more advanced strategies make them even more successful – the answer seems to be yes to both of these questions. First of all, our interviews with users suggest that the users generally feel that they are successful in their information search. However, I believe people to adjust their information needs to the strategies they know. Along these lines, I believe that most of the web searchers use search engines for simple topical search tasks – tasks where broad queries of one or two terms are enough. In those tasks, they might feel highly successful. Most of the users may not have the motivation to challenge their search skills by trying to find

information that is more difficult to find (such as information that needs to be gathered from multiple sources or difficult fact-finding tasks) – even more importantly, they might not even we aware that search engines could provide them with more information than pages that are topically related to their queries. I believe that even those users should have the possibility to search for information for more complex search tasks if they wished – and to make their wish to be the limiting factor rather than their skills is the goal that we should try to accomplish.

6. ACKNOWLEDGMENTS

I would like to thank my colleagues at Google for inspiring discussions related to the themes of this paper. Special thanks to Kerry Rodden, Dan Russell, and Laura Granka.

7. REFERENCES

[1] Aula, A. Query Formulation in Web Information Search. In Isaías, P. & Karmakar, N. (Eds.) *Proceedings of IADIS International Conference WWW/Internet 2003*, 403-410.

[2] Aula, A. Enhancing the readability of search result summaries. *Proceedings of the Conference HCI 2004: Design for Life,* 1-4.

[3] Aula, A. *Studying user strategies and characteristics for developing web search interfaces*. Dissertations in Interactive Technology, Number 3, 2005.

[4] Aula, A. Older Adults' Use of Web and Search Engines. *Universal Access in the Information Society*, 4, 1-2 (2005), 67-81.

[5] Aula, A., Jhaveri, N., and Käki, M. Information search end re-access strategies of experienced web users. *Proceedings of the 14th International World Wide Web Conference* (Tokyo, Japan, 2005), 583-592. ACM Press, 2005.

[6] Aula, A. and Käki, M. (2005) Less is more in web search interfaces for older adults. *First Monday,* 10, 7, 2005.

[7] Aula, A., Majaranta, P. and Raiha, K-J. Eye-tracking reveals the personal styles for search result evaluation. *Proceedings of the IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT 2005)*, 1058-1061.

[8] Aula, A. & Nordhausen, K. Modeling successful performance in web searching. *Journal of the American Society for Information Science and Technology*, 57, 12 (2006), 1678-1693.

[9] Aula, A. and Siirtola, H. Hundreds of folders or one ugly pile – strategies for information search and re-access. *Proceedings of the IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT 2005)*, 954-957.

[10] Bruce, H., Jones, W., and Dumais, S. Keeping and re-finding information on the web: what do people do and what do they need? *Proceedings of ASIST 2004.*

[11] Granka, L., Joachims, T., and Gay, G. Eye-tracking analysis of user behavior in WWW search. *Proceedings of SIGIR'04,* ACM Press (2004), 478-479.

[12] Jansen, B.J. and Pooch, U. Web user studies: A review and framework for future work. *Journal of the American Society for Information Science and Technology,* 52, 3 (2001), 235–246.

[13] Jansen, B.J. and Spink, A. How are we searching the World Wide Web? A comparison of nine search engine transactions logs. *Information Processing and Management*, 42 (2006), 248-263.

Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages

Kerry Rodden
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043, USA
kerryr@google.com

Xin Fu
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599, USA
xfu@unc.edu

ABSTRACT
A mouse click is a proven indicator of a user's interest in a web search result. In this paper we explore the potential of a more subtle signal: mouse movements. We conducted a study where participants completed a range of tasks using Google, and we tracked both their eye movements and mouse movements. We discuss the relationship between these movements, and three different types of eye-mouse coordination patterns. We believe that mouse movements have most potential as a way to detect which results page elements the user has considered before deciding where to click.

1. INTRODUCTION
Web search engines are the source of most web users' interactions with information retrieval systems. Researchers have explored the potential of analyzing click patterns from search engines, both as a means of evaluating their ranking functions in the context of users' real information needs, and of gathering feedback to improve ranking for subsequent users (e.g. [1][8]). While extremely valuable, clicks do not tell the whole story of the user's interaction with the search results page. For example, they do not indicate *why* the user clicked on a particular result, or which other results they considered before making a choice.

A user's selection of a particular search result is based on the surrogate shown on the results page, which typically contains the page title, its URL, and a "snippet" showing one or more lines from it. The likelihood of a user clicking on a result is mostly dependent upon how promising they think it is, based on the surrogate. It would therefore be useful to have a better idea of which aspects of the surrogate users are paying attention to when making each decision about where to click. Also, in some cases it may be possible for the user to find the answer to a fact-finding question simply by reading the snippet, and many search engines now choose to present relevant information on the page directly, e.g. the definition of a word, or a stock quote. In both of these situations, no click would occur even though the user may have satisfied their information need.

Eye tracking can provide insights into users' behaviour while using the search results page, but eye tracking equipment is expensive and can only be used for studies where the user is physically present. The equipment also requires calibration, adding overhead to studies.

In contrast, the coordinates of mouse movements on a web page can be collected accurately and easily, in a way that is transparent to the user. This means that it can be used in studies involving a number of participants working simultaneously, or remotely by client-side implementations – greatly increasing the volume and variety of data available.

Our goal in conducting this research was to investigate the potential usefulness of tracking mouse movements on a web search results page – for example, how closely do mouse movements reflect eye movements? Do people use the mouse pointer as a marker to help them read the search results, or to help them make a decision about where to click?

Unlike this study, which focused on search results pages, previous studies on the relationship between eye movements and mouse movements have been concerned either with general web pages (e.g. [2][4][10]) or with tasks that involve locating and selecting a given target item from graphical user interface menus of various lengths (e.g. [6]). It is unclear whether findings from these studies carry over to web search results. There have been several studies involving eye tracking on web search results pages, e.g. [8][9], but these have not considered mouse actions other than clicks. Researchers of implicit relevance feedback (e.g. [5][7]) have found that mouse actions on general web pages are a potentially useful signal, but they have not studied web search results pages.

2. STUDY SETUP
2.1 Tracking Mouse and Eye Movements
To capture mouse movements, we used a method similar to that described in [2] and [3]: a web proxy server inserted a reference to a piece of Javascript code at the top of every Google search results page visited. This Javascript code captured the user's mouse coordinates (and the ID of the HTML element the mouse was over) at 100 millisecond intervals, and submitted the gathered data, with timestamps, into a MySQL database every 2 seconds (or when the user left the Google search results page).

To capture eye movements, we used a Tobii 1750 eye tracker running Clearview software, with a 17-inch screen set to a resolution of 1024x768. We used IE 6 at full size on the screen. Clearview logged each URL and saved a screenshot of every web page visited during the study.

2.2 Participants
Our 32 participants (14 male and 18 female, aged 24-61) had a range of occupations and web search experience, but all were familiar with Google. 22 were from our company's user study participant database; 10 were non-technical company employees.

2.3 Tasks
We used 16 web search tasks – a sample of which are listed in Table 1. All were of the informational type, rather than

navigational or resource-related [11]. We mostly chose closed fact-finding questions with a specific correct answer, so that the participants would have a clear idea of what was required and when they were done with the task. However, in 3 of the tasks, the user had to make a decision based on their own preference.

We provided initial queries for each task (also shown in Table 1), to ensure that each user would see the same results at first. This meant that, for fact-finding tasks, we could choose queries where the answer to the question was visible on the results page itself (e.g. in the snippet of one of the results) – because we were interested to see if users would move the mouse over the answer while looking at it. For all but one of the queries, there was at least one useful result on the initial page. We did not manipulate the visible content in any way, so as well as the 10 search results, many of the pages included other elements such as sponsored links, and extra information inserted by Google.

2.4 Procedure

This was an exploratory study, not a controlled experiment, so the participants each did the same 16 search tasks, in the same order. They started from a study home page that contained one link per task. Each of these links led to a fake Google home page, with the initial query pre-filled in the search box, and the task description inserted underneath. Participants were instructed to press the "Google Search" button once they had read and understood the description and query. This was important to ensure that the users started scanning the search results page from the same place as they would if they were entering a query into Google as normal.

As mentioned above, all users saw the same (cached) results page at first. Once they were on this first results page, it was up to them to do whatever they thought they needed to in order to complete the task – e.g. reading text on the results page itself, clicking on links, or changing the query. They could move on to the next task as soon as they felt they were done, or were ready to give up. They did this by pressing the Home button in the web browser, which returned them to their task list.

Before starting on the tasks, each participant was walked through an example and then did a practice task. Each session lasted less than an hour (typically around 30 minutes for all 16 tasks).

Table 1: 4 of the 16 tasks used in the study.

Query	Description
actress most oscars	You are a movie fan and are curious to know which actress has won the most Oscars.
lawn chair	You are going to an outdoor concert soon, and want to get yourself a lawn chair to take with you. Find one online that you would consider buying.
eschew definition	You read an article about healthy eating that listed some foods to "eschew". You want to check what that word means.
bono real name	You were watching TV and saw something about Bono, the singer. You're curious to know what his real name is.

3. RESULTS

The primary unit of analysis is a *visit* to a Google search results page. Pressing the "Google Search" button to begin a task starts the first visit of the first query. Clicking through on a result and clicking on the Back button in the web browser begins the second visit of the first query (because the results are still the same). In total there were 1216 visits to 786 queries across the 32x16 = 512 tasks. The results reported here apply to all visits in the study.

We used Clearview's default of 100ms minimum duration and 30 pixel maximum dispersion to determine eye fixations, and used the raw mouse data for our analyses. We wrote a program that automatically identified the outlines of interesting regions on each results page visited. These were the 10 search results, any sponsored links or additional pieces of information appearing with the results, and the top area of the page (including the search box, logo, etc). The remaining areas of the page were collectively treated as a single region, called "other".

3.1 Overall Eye-Mouse Relationship

3.1.1 Relative Distribution of Attention

Figure 1 shows the relative distributions of the user's attention across selected regions of the page, comparing the proportion of mouse data points in each region to the equivalents for total eye fixation duration and total number of clicks. For results 1-10, the distributions are quite similar, but it is interesting to see that the mouse spends a much higher percentage of time in the "other" regions (empty space and the bottom of the page) than the eye.

Using the full set of regions, we were interested to know how likely it was, within a single visit, that when the user moved their mouse over a region, they also looked at it. Of regions that they covered with the mouse, a mean of 76.2% (s.d. 23.4) were also fixated on during the visit. Conversely, of the regions that the users fixated on during a single visit, a mean of 64.0% of those regions (s.d. 25.7) were also covered by the mouse.

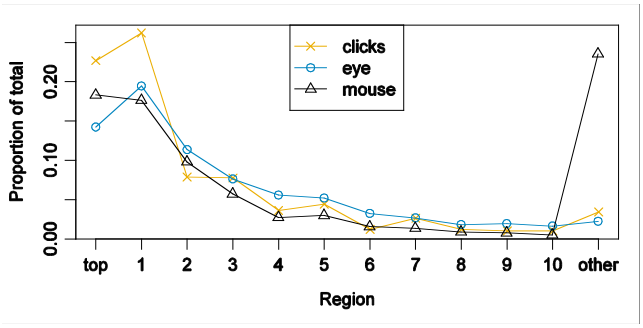

Figure 1: How distributions of eye fixation time and clickthrough relate to distribution of mouse hovering time, for the regions that were common to all pages.

3.1.2 Distance Between Mouse and Eye

In order to calculate the set of distances between mouse and eye, we matched each mouse point with the eye fixation (if any) whose duration spanned it. The overall distribution of distances is skewed, with a long tail to the right. The mean across all fixations is 257 pixels (s.d. 237); the median is 191.

It is interesting to consider the X and Y directions separately (Figure 2). For the Y direction there is a much higher peak around 0 than there is for the X direction, suggesting that mouse and eye positions corresponded more closely in the vertical direction than in the horizontal direction.

Considering the different regions on the page, we found that when the mouse was over results 1-10, the mean eye-mouse distance dropped to 194 pixels (s.d. 132). When it was over the "other" region, the mean rose to 551 pixels (s.d. 305). We have already seen that the participants were more likely to have their mouse in this area than they were to fixate on it. Combined with the data in Figure 2, this might suggest that a common behaviour is to keep the mouse in the blank area to the right of the search results, moving it downwards while scanning the results with the eyes.

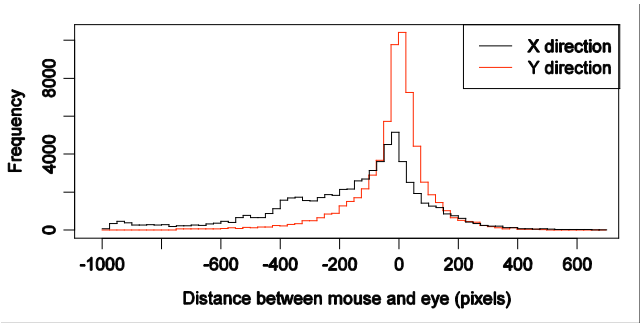

Figure 2: Histogram outlines of distance from mouse to eye, broken out separately for the X and Y directions. Each step in the histogram represents a bucket of 25 pixels. A negative X distance means that the eye was to the left of the mouse, and a negative Y distance means that the eye was below the mouse.

Further evidence for this comes from investigating how frequently the eye and mouse were in the same region at the same time. Overall, eye and mouse were in the same region for 42.2% of the mouse data points that had a corresponding eye fixation and this dropped to 6.0% if the mouse was in the "other" region. So although there is a reasonably high overlap between the regions covered by eye and mouse within a single visit to a results page, this overlap does not necessarily occur at the same time. One exception is the top of the page, where eye and mouse coincide about 70% of the time (when users were refining queries) – this is also the region with the shortest eye-mouse distances.

3.2 Eye-Mouse Coordination Patterns

Following the high-level analysis described in the previous section, we wanted to consider in more detail the interactions between eye movements and mouse movements within a visit. For each visit, we generated a visualization of the paths followed by the user's eye and mouse. This was straightforward for the eye fixations – we simply placed a circle at the point of each fixation, with area proportional to the fixation duration. To make equivalent visualizations for the mouse data, we created mouse "fixations", using the I-DT dispersion-based algorithm [12] with a minimum duration of 0ms (so no points were thrown away) and maximum dispersion of 10 pixels. We studied a sample of the visits (see http://www.rodden.org/kerry/publications/wisi07/ for example visualizations) to look for instances of the different patterns. In addition to these, we used time series plots to help us understand the relative timings of the events in more detail.

There are some patterns of eye-mouse coordination that we will not discuss further here – in general, these are movements that the user must make in order to complete their task. For example, moving the eye and mouse together to the search box in order to refine a query, or to the scroll bar in order to move further down the page.

3.2.1 Keeping the Mouse Still While Reading

In this pattern, the user holds the mouse away from the place where they are currently reading, keeping it mostly still until they have seen the result they want to click on. In general, the most common starting position of the mouse was the position of the "Google Search" button on the previous page – at the beginning of a task, users were forced to click on this button. This is also the most common starting point for the eye, but the mouse typically stays in this position for longer than the eye does. So it seems that users very often exhibit this pattern at the beginning of a visit.

As well as the starting position, the blank area to the right of the results (in the "other" region) was also a common resting place for the mouse. Users varied a lot in this regard – the mean percentage of time that they left the mouse in the "other" region ranged from 2% to 57%. Those participants who used the scroll wheel on the mouse would often scroll the page while resting the mouse in this area, resulting in a pattern of evenly-spaced mouse "fixations" in a vertical line. We also saw several pieces of evidence in the high-level data that users adopt behaviours like these at least some of the time, including the mouse being left in the "other" region much longer (relatively) than the eye spends there, and the eye and mouse being at their most distant on average when the mouse is over the "other" region.

3.2.2 Using the Mouse as a Reading Aid

In this pattern, the mouse pointer is moved around to help the user keep their place on the page while reading.

The most common form of this pattern was for the user to move the mouse pointer mostly in the vertical direction, so that it was either touching or roughly level with the region they were currently reading – for the search results, this was often in the "other" area, to the right. This form may help to explain the fact that the mouse and eye tended to be closer in the Y direction than in the X direction.

In another form of this pattern, the user moves the mouse pointer horizontally across or below the text they are currently reading. This is illustrated in Figure 3, and is characterized by sequences where the user makes short mouse movements (with short "fixations") while over a result. Although striking, this behaviour was rare in the study. Only a handful of the users ever followed a whole line of text with the mouse while reading it, and they did not do this for every task. It was more likely, however, that in cases where the answer to the task was visible on the page, users would move their mouse over the answer.

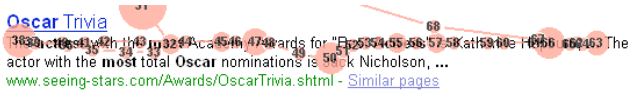

Figure 3: Example of "reading" with the mouse (user 21). The user ran the mouse pointer over part of the snippet, containing the answer to the task ("The actress with the most Academy Awards for 'Best Actress' is Katharine Hepburn").

3.2.3 Using the Mouse to Mark an Interesting Result

In this pattern, the user leaves the mouse pointer on or near the result that seems to be the most promising one they have read so far, while their eyes continue to check more results. Often, the mouse is left hovering over the title of the promising result – ready to click if the user eventually decides to select it. If another result seems more promising, the user will move the mouse on to that result, and so on. The difference between this pattern and the previous two is that the mouse is kept still for the purpose of marking an interesting result, not simply to keep it out of the way or to keep the user's place on the page while reading.

From inspection of the visualizations, and based on previous work on selection from menus [6], we speculate that users will be more likely to exhibit this pattern when the task is more difficult, and it

is not obvious to them which result is best – especially as they move further down the page. Further studies would be required to confirm this.

4. DISCUSSION

The study has given us some tentative answers to the questions we raised in the Introduction. We have found that mouse movements definitely show potential as a way to estimate which results page elements the user has considered before deciding where to click, e.g. by noting which regions were covered by the mouse during the visit, or measuring the vertical distance traversed. This has implications for evaluating the user-perceived quality of the search results (as judged from their surrogates). Mouse movements also have some potential as a method of determining whether the user has noticed the answer to their question on the results page itself, thus helping to evaluate design choices in page formatting and layout. Behaviour such as that illustrated in Figure 3 (following a line of text with the mouse while reading it), is relatively infrequent, but when it does occur, it indicates which aspects of the surrogate the user is taking into account when making a decision.

It is interesting to consider whether it would be possible to automatically identify useful patterns from mouse data alone. For example, the pattern discussed in Section 3.2.3 (using the mouse to mark an interesting result) would be particularly useful to identify, since it indicates which surrogates the user has found most relevant. However, from the mouse data alone, it is difficult to tell the difference between this pattern and that of simply moving the mouse vertically while reading. In both, the mouse pointer is still or relatively still while touching or near a result. Without the eye data, we cannot determine the exact sequence of events: did the user move their mouse to the result simply because they were in the process of reading it, or because they had already read the surrogate and decided it was relevant?

To attempt to narrow this down, we looked at cases where the user was holding the mouse pointer over the title of the result (ready to click), not just on or near the result block in general. We found that in 172 of the 1216 visits (14.1%), the user held the mouse for more than 1 second over the title of a result that they did not click on during the visit. All but one user was represented in these visits. However, manual inspection of a sample of these visits (and the associated visualizations and plots) showed that this heuristic was not enough to separate the two patterns.

In general, automated analysis of this data is complex – even with the most clear-cut examples of the patterns, it would not be straightforward to identify them automatically. Similarly, we found that the users were not easy to classify, and each one seemed to exhibit all of the patterns, to varying degrees. There is also substantial variation between users in all of the high-level measures described in Section 3.1. For example, per user, the mean distance between eye and mouse ranged from 144 to 456 pixels, and the proportion of mouse data points at which the eye and mouse were in the same region ranged from 25.8% to 59.9%.

5. FUTURE WORK

This was an exploratory study, and there is a lot of scope for more research in this area. At a minimum, additional exploratory studies could employ different sets of users and tasks (perhaps having users do their own tasks instead of prescribed ones), as well as different search results page designs.

Controlled experiments, perhaps systematically manipulating the results or the result order according to relevance, would help to confirm some of the findings presented here. In particular, it would be valuable to study the relationship between the information provided by a surrogate and the pattern of marking results with the mouse, e.g. to confirm that if the user keeps the mouse pointer over a particular result while continuing to read others, this is indeed because they saw something relevant in the surrogate.

Findings from such experiments would assist in generating reliable and valid metrics from mouse data. Such metrics would be a prerequisite to conducting any larger-scale studies (e.g. with remote users) where only mouse data can be collected.

6. ACKNOWLEDGMENTS

Xin Fu was an intern at Google when this study was conducted. We are grateful to Chris Mysen, Dan Pupius, Gill Ward, Simon Quellen Field, Laura Granka, Dave Poole, and Tiffany Griffith for their help with this work.

7. REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. 2006. Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR'06*. 3-10.

[2] Arroyo, E., Selker, T., and Wei, W. 2006. Usability tool for analysis of web designs using mouse tracks. In *CHI '06 Extended Abstracts*. 484-489.

[3] Atterer, R., Wnuk, M., and Schmidt, A. 2006. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proceedings of WWW'06*. 203-212.

[4] Chen, M.-C., Anderson, J. R., and Sohn, M.-H. 2001. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI '01 Extended Abstracts*. 281-282.

[5] Claypool, M., Le, P., Wased, M., and Brown, D. 2001. Implicit interest indicators. In *Proceedings of IUI'01*. 33-40.

[6] Cox, A.L. and Silva, M.M. 2006. The role of mouse movements in interactive search. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*. 1156-1161.

[7] Hijikata, Y. 2004. Implicit user profiling for on demand relevance feedback. In *Proceedings of IUI '04*. 198-205.

[8] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR'05*. 154-161.

[9] Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., and Gay, G. 2006. The influence of task and gender on search and evaluation behavior using Google. *Information Processing and Management, 42*(4), 1123-1131.

[10] Mueller, F. and Lockerd, A. 2001. Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI '01 Extended Abstracts*. 279-280.

[11] Rose, D. E. and Levinson, D. 2004. Understanding user goals in web search. In *Proceedings of WWW'04*. 13-19.

[12] Salvucci, D. D. and Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of ETRA'00*. 71-78.

Revisiting informativeness as a process measure for information interaction

Luanne Freund
School of Library, Archival and Information Studies
University of British Columbia
Vancouver, British Columbia, Canada
luanne.freund@gmail.com

Elaine G. Toms
Centre for Management Informatics
Dalhousie University
Halifax, Nova Scotia, Canada
etoms@dal.ca

ABSTRACT
The intent of this paper is to re-introduce and discuss the applicability of the *informativeness* concept to web-based information seeking and retrieval environments. Informativeness is rare among IR evaluation concepts, in that it focuses on the value of the process of interaction with a set of documents, rather than on the success of the algorithm matching documents to information needs. It is notable as an early attempt to bridge traditional system evaluation with the new-at-that-time perspective on the role of the user in the evaluation of the system.

Keywords
informativeness, search trails, search process

1. INTRODUCTION

> *the evaluation of information retrieval systems should be based on measures of the information provided by the retrieval process, 'informativeness' measures which take into account the interactive and full-text nature of present-day systems and the different types of questions which are asked of them. [1]*

Web-based information seeking and retrieval (IS&R) can be described as a process of interaction between people and diverse sets of digital objects, information, technologies and applications, with the primary goal of becoming informed about something. User-centred approaches have characterized the motivation for IS&R as a gap in or lack of knowledge, and have emphasized the cognitive processes associated with bridging this gap through making sense, learning and becoming informed [2, 3]. This conceptual framework is based on what Buckland [4] refers to as "information-as-process", because it emphasizes the change that takes place in the searcher through interaction with information objects.

Given that search engines are the main tools used to support IS&R activities on the Internet, it might be reasonable to think that they would provide support for the process of becoming informed. However, most search engines take a rather limited, transaction-based approach to web searching based on the "information-as-thing" approach [4]. To extend the knowledge gap metaphor, search engines are well-designed to retrieve sets of information objects that may fill the gap, but are not very well-designed to support people as they interact with the information and build bridges to span the gap. At the interface level, this limited approach to search system design has been exacerbated by the general perception that Google's "less is more" approach is optimal.

However, the lack of attention paid to the information interaction process is rooted more fundamentally in some of the key theoretical assumptions of the IR field. First among these, is that IR is primarily concerned with the relevance relationship between documents and user needs, rather than on learning, task completion or other broader process outcomes. Second, is that the goal of IR system design is to retrieve all the relevant documents and as few of the non-relevant as possible [5]. By focusing on balancing recall and precision, search systems do not provide good solutions for many searchers, who are simply seeking the shortest path to some level of information saturation. The third, expressed in the probability ranking principle, is that the best possible rank order of results is in decreasing probability of their relevance to the user. As noted by van Rijsbergen [5], this is highly problematic, as it assumes that the relevance of a document is independent of the other retrieved documents. In fact, we know that documents in a set may contain redundant or complementary information, and that the order in which the documents are viewed can have a major impact on relevance judgments [6, 7].

While it is true that search engine users have become accustomed to sifting through long lists of disconnected results, the act of filtering, sorting and making sense of them still constitutes a major cognitive challenge. It is unlikely that systems will start to address this challenge until IR evaluation frameworks are able to incorporate users and their goals in a meaningful way. Some notable work has been done in this area, focusing on using more realistic evaluation tasks, accommodating graded relevance assessments, and developing new and more realistic evaluation metrics [8-10]. Work at the TREC conferences has also begun to consider more user interaction issues, for example through the Web, HARD, and Novelty tracks, however much of the evaluation work in these tracks continues to be devoid of user participation.

2. DEFINING INFORMATIVENESS
The intent of this paper is to re-introduce and discuss the applicability of the *informativeness* concept to web-based IS&R environments. Informativeness is a user-centred concept for evaluating the effectiveness of a retrieval process, which was first proposed by Tague [11] 20 years ago. She updated it some years later in response to the move from traditional text retrieval systems to full-text interactive systems, "in which the searcher follows trails in online or ondisk databases, rather than scanning batches of 'hits'"[1]. Tague and her students developed and tested the measure, but plans to validate it in a larger study were not realized after Tague passed away in the midst of the project in 1996.

The concept of informativeness is based on a user-centred, subjective understanding of information. Tague [11] cites Fox's definition: "the information carried by a set of sentences …is the conglomerate proposition expressed by the sentences, a proposition which the originator of the sentences is in a position to know to be true…..[and] which the recipient can read and understand"[12]. Thus, the informativeness value of an information object or set of objects is determined by "the amount of information which it carries or conveys to an individual….in the context of a particular query or information need."[11]

Tague's primary focus in developing the informativeness measure was to assess the value of interactive search processes and the impact of the order in which a set of retrieved information objects are encountered. Tague explains that at the process level, "Informativeness is … related to the completeness and ordering of the search trail with respect to some expected or ideal answer set. The information provided by a text is context sensitive, in that it depends on what has already been read, on the reading order.[1]"

In considering how the informativeness of a search trail is affected by the order of display, Tague identified three types of document dependencies.

Independence: an object is independent of other objects in the set if it deals with an aspect of the information need not covered elsewhere.

Complementarity: an object is complementary if its informativeness is influenced or influences other objects in the set. In this case, the order in which objects are displayed may add or detract from the overall informativeness.

Referential (redundancy): an object is in a referential relationship with other objects in the set, if it contributes no new information and does not increase the informativeness of the search trail.

Informativeness is clearly related to the intertwined concepts of cognitive and situational relevance [13], as it depends on the extent to which information objects, both individually and in the aggregate, suit a users' cognitive and practical needs with respect to the task of becoming informed. However, the concept of informativeness is more pragmatic than relevance. Relevance claims that an information object is related or suited to the information need in some manner, which implies potential usefulness. Informativeness, on the other hand, claims that a searcher has actually interacted with an information object or collection of objects and has become informed to some extent. Thus, while relevance is well-suited to the evaluation of document surrogates, which are used to predict the usefulness of the actual documents, informativeness has more to offer as a measure of information interaction in full-text environments, in which the searcher reads and becomes informed as part of the search process.

Informativeness is also related to the concept of novelty, which is the main user-centered caveat to the probability ranking principle. Novelty has been studied to some extent by the IR research community, primarily in terms of novelty detection [14]. However, novelty is only one aspect of informativeness, since the outcome of a process of information interaction may be affected by the many different ways in which information objects are related to one another and to the searcher, such as explanation, elaboration, confirmation, etc.

Both at the level of individual information objects and at the aggregate level of an ordered set of results, informativeness is a subjective concept; a function of what individual searchers can extract, understand and make use of in order to become informed. It measures the value of a user's interaction with information as guided by ranked system output, against a user-defined search path.

3. MAKING USE OF INFORMATIVENESS

Tague reasoned that given a set of search results, searchers can determine an ordinal utility function for that set. In other words, they are capable of determining a weak ordering of information objects (with some ties) based on their relative informativeness, and of determining a stopping point in the sequence. In some cases, the information need may need to be broken down into distinct, ordered aspects to facilitate this ordering [15]. This user-defined sequence represents the optimal sequence, which is then used to determine the relative informativeness value of each item in the sequence.

Based on this formalization, Tague developed an informativeness measure that scores ranked system output compared to an optimal, user-defined sequence [1]. The measure is based on a number of assumptions:

- The amount of information that an information object delivers varies from person to person, and will depend for each person on what has been previously examined.
- A user can examine the information objects retrieved and place them on a continuum based on their relative informativeness, from their own perspective.
- The informativeness of each information object is logarithmically related to the ones previously viewed, from the user's perspective.
- The overall informativeness of an ordered set of information objects is reduced when non-pertinent documents are presented or pertinent ones were presented in an order not useful to the user.

The measure combines an initial user-defined informativeness score for each object in the set with a penalty when the system does not deliver the results in the optimal order defined by users at the time of the evaluation [15].

4. MEASURING INFORMATIVENESS

Few studies have applied the informativeness measure. Tague-Sutcliffe [16] examined the search trails of 17 online catalogue searches. In this initial study to validate the metric, she found an 82% correlation between user perception of informativeness and the value of the informativeness measure.

In 1996 Toms and Tague-Sutcliffe [17] conducted two studies to apply informativeness to browsing. In the studies, ten and twenty participants, respectively, browsed and searched in digital encyclopedias. In both studies, participants performed two tasks: implicit, in which they were asked to browse anything of interest, and explicit, in which they were assigned a task with a defined goal. After each task was completed, participants did a 'free recall' of all articles on the assumption that those remembered were the most informative. Secondly, a screen capture video was replayed so that participants could identify search goals that emerged over the course of the session. Participants then matched

goals that emerged with the informative documents and ranked the documents in two ways: in the order they would have preferred to have viewed them, and in order of informativeness for the task.

Using both the trail extracted from the screen video and the informative nodes specified by the users, informativeness was calculated for each task. Initial informativeness (i.e., the information contained in the documents) by task ranged between .78 and .95. When that result was penalized by the system delay for non-optimal ordering, informativeness dropped to between .5 and .63. In the implicit task condition, participants were more likely to prefer the order in which the system presented the documents, but this was less likely in the explicit task condition.

In these studies, measures of initial informativeness suggest that the systems are performing well at the 78 to 95% precision level. However, when penalized for not displaying the documents in the optimal order, the informativeness scores of the systems drop significantly.

The most similar measure in use in laboratory IR evaluation is Cumulated Gain, which is one of a family of measures that introduce a penalty as the rank increases [18]. Like informativeness, Cumulated Gain can also accommodate graded judgments of document value. However, in addition to using document level assessment, informativeness tests the system against user-defined optimal-paths collected in situ at the time of the evaluation. This is methodologically challenging, but has the potential to increase our understanding of how to best display search results, both in general terms, and with respect to particular user groups and situations.

5. RETHINKING INFORMATIVENESS IN A WEB WORLD

The informativeness measure was developed at a time when interactive hypertext search systems were still in their infancy, yet both as a concept and as an evaluation measure, it brings some interesting perspectives to Web-based IS&R.

Objective document informativeness

At the level of individual information objects, informativeness is primarily a subjective value, in that it will be influenced by the user's need, task, preferences, etc. However, unlike relevance which is an inherently relational concept, informativeness can also be viewed as an objective feature of a document, indicative of the raw potential a document has of informing a reader. This raises the question of how the objective informativeness potential of an object may be measured.

Tague [11] suggests that it may be a function of the topic coverage and cites work by Derr [19], who proposes combining the breadth of coverage measured by the number of propositions in a text, and the depth of coverage based on the specificity of the propositions. In the web world there are any number of additional, non-textual features that may be indicative of informativeness, such as the number of links, the depth of a page in a web site, the structuredness of text, the number of images, the genre of the page, etc [20]. Recent research on entity extraction and question answering has explored statistical methods to identify the "informativeness" or discriminating power of words and sentences within texts and collections [21, 22], and these methods might also be applicable to determining a raw informativeness value for texts.

Although measuring an information object's potential for informing a user was not the focus of Tague's work, developing such a measure could be of benefit to Web-based IS&R systems. In particular, it could be used to identify informational as opposed to navigational or transactional pages, and to weight these pages for particular types of queries.

Informativeness of a search sequence

Beyond the level of individual information objects, informativeness provides insight on the ordering of search results, one of the critical issues in search engine design. As noted above, Tague's theoretical description of search trails identifies different types of dependencies between information objects [15]. Further exploration of these relationships, both theoretical and through empirical studies of user-defined search trails, could lead to a better understanding of the role of document dependencies in human information interaction.

A major challenge in predicting search trails of optimal informativeness is their variability across diverse users and information needs. Tague [1] suggests that at a general level, trails vary along two main dimensions: length (number of information objects) and consistency (degree to which optimal trails are consistent over different users), and hypothesizes that these variables are dependent upon the nature of the information need. As examples, she notes that factual search trails are likely to be very short and quite consistent among searchers, while more complex tasks are likely to have longer and more varied trails.

While individual differences will always influence preferences for the ordering of search results, it may be possible to make use of some of these contextual patterns based on tasks, information needs, domains, etc. to predict better search trails for different situations. For example, process models and genre systems used within organizational information environments could provide guidance in ordering results, by predicting the types of documents needed by searchers and the order in which they are needed. Generic models of the information seeking process [3, 23, 24] may also be of use. Given the large amounts of user behaviour data currently available in search engine logs, a fruitful approach may be to validate model-based predictions of optimal trails with data mining techniques.

Web-based IS&R involves interaction with a much more diverse range of information objects, systems and interfaces than would have been part of the search process when the informativeness measure was developed. Given the current, richer interactive environment, it is possible that the concept needs to be extended to take into account the ways in which searchers are informed by the environment as well as the objects it contains. In this way, it would be possible to consider the role that environmental cues play in enhancing the informativeness of a search process and to consider how environmental informativeness could be improved. This reconceptualization would support holistic evaluations in keeping with the way in which people experience web searching as part of a broader interaction with the web environment.

Practical Applications

Informativeness evaluation can provide design input to search engines, with the goal of providing users with support for the process of becoming informed. In practical terms, search engines need guidance on the prediction of optimal sequences of results. These could be based on general search scenarios, or customized

to different task scenarios provided by the searcher. Search engines could then provide interfaces that allow searchers to interact with these results within an application that provides support for interaction with information. Alternately, search engines could provide toolbar tour guides, to walk searchers through static or dynamically updated paths through information. A third option is to retrieve search trails, rather than individual results, so that the hitlist would contain the starting points of the highest ranked trails, optimized for different kinds of information needs or tasks. Some search engines (trexy.com, trailfire.com) have already begun to make use of the search trail concept, by allowing users to create and share their trails with others.

6. CONCLUSION

The value of reconsidering informativeness is that it has the potential to open up new ways of thinking about search systems, which extend beyond retrieval to support the broader interaction process. It forces us to think about optimizing the paths which searchers take through information in order to maximize the amount of information they can absorb and make sense of.

Informativeness is particularly valuable due to it flexibility. It can be measured both for single units of information (sentences, texts) and for search sessions involving interactions with multiple texts. Also, it can be measured as an objective textual feature, as well as a subjective measure of the interactivity between users and texts. There is considerable potential for the application of informativeness to Web-based information interaction and there are a number of open research questions of interest to the IS&R research community.

7. REFERENCES

[1] J. Tague-Sutcliffe, "Measuring the informativeness of a retrieval process," presented at Proceedings of the 15th Annual International ACM SIGIR Conference, Copenhagen, Denmark, 1992.

[2] B. Dervin, "On studying information seeking methodologically: the implications of connecting metatheory to method," *Information Processing & Management*, vol. 35, pp. 727-750, 1999.

[3] C. C. Kuhlthau, "Inside the search process: information seeking from the user's perspective," *Journal of the American Society for Information Science*, vol. 42, pp. 361-371, 1991.

[4] M. Buckland, "Information as thing," *Journal of the American Society of Information Science*, vol. 42, pp. 351-360, 1991.

[5] C. J. van Rijsbergen, *Information retrieval*, 2nd ed. London: Butterworths, 1979.

[6] M. B. Eisenberg and C. L. Barry, "Order effects: a study of the possible influence of presentation order on user judgements of document relevance," *Journal of the American Society for Information Science*, vol. 39, pp. 293-300, 1988.

[7] M. A. Tiamiyu and I. Y. Ajiferuke, "A total relevance and document interaction effects model for the evaluation of information retrieval processes," *Information Processing & Management*, vol. 24, pp. 391-404, 1988.

[8] P. Borlund, "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems," *Information Research*, vol. 8, 2003.

[9] P. Ingwersen and K. Jarvelin, *The turn: integration of information seeking and retrieval in context*, vol. 18. Berlin: Springer, 2005.

[10] J. Kekalainen and K. Jarvelin, "Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance," presented at 4th CoLIS Conference, 2002.

[11] J. Tague, "Informativeness as an ordinal utility function for information retrieval," *SIGIR Forum*, vol. 21, pp. 10-17, 1987.

[12] C. J. Fox, *Information and misinformation*. Westport, Conn.: Greenwood, 1983.

[13] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science*, vol. 54, pp. 913-925, 2003.

[14] I. Soboroff, "Overview of the TREC 2004 novelty track," presented at Text Retrieval Conference, Gaithersburg, MD, 2004.

[15] J. Tague and R. Schultz, "Some measures and procedures for evaluation of the user interface in an information retrieval system," presented at Proceedings of the 11th Annual International ACM SIGIR Conference, Grenoble, France, 1988.

[16] J. Tague-Sutcliffe, *Measuring information: an information services perspective*: Academic Press, 1995.

[17] E. G. Toms and J. Tague-Sutcliffe, "Informativeness as a measure of browsing effectiveness," presented at Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, Copenhagen, 1996.

[18] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM transactions on Information Systems*, vol. 20, pp. 422-446, 2002.

[19] R. L. Derr, "The concept of information in ordinary discourse," *Information Processing & Management*, vol. 21, pp. 489-500, 1985.

[20] A. Tombros, I. Ruthven, and J. M. Jose, "How users assess web pages for information seeking," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 325-436, 2005.

[21] T. Roelleke, "A frequency-based and a Poisson-based definition of the probability of being informative," presented at ACM SIGIR, Toronto, Canada, 2003.

[22] J. D. M. Rennie and T. Jaakkola, "Using term informativeness for named entity extraction," presented at ACM SIGIR, Salvador, Brazil, 2005.

[23] P. Vakkari, "A theory of the task-based information retrieval process: a summary and generalization of a longitudinal study," *Journal of Documentation*, vol. 57, pp. 44-60, 2001.

[24] K. Bystrom and K. Jarvelin, "Task complexity affects information seeking and use," *Information Processing & Management*, vol. 31, pp. 191-213, 1995.

Measuring the Navigability of Document Networks

Mark D. Smucker and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{smucker, allan}@cs.umass.edu

ABSTRACT

Browsing by similarity is a search tactic familiar to most people but one that the web unevenly supports. We are interested in user interface tools that augment the web with links to help users navigate from one relevant document to other relevant documents. We propose a combination of simple metrics to measure the navigability of document networks. These measures provide for low cost evaluation of the document networks formed by similarity measures and other link creation methods.

1. INTRODUCTION

After a long and tiring search, a user finally finds a web page relevant to the user's information need. While the page is relevant, it does not fully satisfy the user's need. How should the user proceed? If the page provides links to other pages, the user can follow those links. Alternatively, the user could follow links automatically produced by a tool that examines the page's content and provides links to similar pages. Tools that allow a user to request a list of documents similar to given document support the user interface feature we call *find-similar* [9].

Find-similar provides the search user a means to travel from one document to another. In effect, find-similar links together documents into a network, and just as a traveler in the physical world needs a good road system with direct routes, the search user needs find-similar to produce links that minimize the travel time to relevant documents. As applied to the web, find-similar aims to create a more navigable network by adding additional links to the existing document network that consists of web pages and hyperlinks.

A find-similar tool embodies some document-to-document similarity method. We would like to be able to test many variations of document-to-document similarity in a low cost manner. Testing different similarity measures with users is likely to be excessively expensive and likely to show little to no difference between similarity measures. Significant differences in retrieval quality can fail to be detected in user studies [4, 12].

The field of human computer interaction (HCI) has developed many methods of automated usability testing [5]. A premise of usability testing is that an interface exists to be tested. We would like measures of document-to-document similarity quality that are largely independent of the user interface otherwise we would need to test the cross product of interfaces and similarity measures.

Furnas [1] has developed a theory of *effective view navigation* that is related to our goal of efficient navigation from relevant document to relevant document. Furnas details his theory in terms of two types of graphs: a logical graph and a view graph.[1] The logical graph represents how objects, such as documents, are truly connected to each other. Furnas gives the web with its hyperlinks as an example of a logical graph. The view graph adds directed links to each node in the logical graph and represents the ways a user who is viewing the current node can immediately get to other nodes in the view. With find-similar, we are looking at ways to augment the logical graph and create a view graph that makes it easier for a user to find relevant documents.

To achieve effective view navigation, a system needs to be both efficiently view traversable (EVT) and view navigable (VN).

To be efficiently view traversable, Furnas requires two things. The first, EVT1, is that the views should be small, in other words, the out-degree of each node should be low when considering the view graph. The second, EVT2, is that the distance from each node to each other node on the viewing graph be short compared to the size of the overall structure.

Furnas' view navigability concerns itself with the "signage" aspects a of system. Links in the network need to provide good "residue" of the objects reachable via the link. Furnas' residue is similar to Pirolli's information scent [8]. In other words, the user needs the link labeled in a manner that provides a form of lookahead. At the same time, the label must be small. Simply providing a listing of everything reachable via the link would provide good residue but would result in too large of a label.

We see Furnas' use of out-degree as an approximation of the user's cost to use the link. As such, while the links in Furnas' graphs are unweighted, we weight each link in the network proportional to the time it takes a user to discover, evaluate, and travel a link.

One of our two measures of document navigability is based on the shortest paths between relevant documents. With regard to EVT2 (shortest paths), the question for information retrieval is not how easy is it to get from one document to

SIGIR'07 Web Information-Seeking and Interaction Workshop, July 27, 2007, Amsterdam, The Netherlands.

[1]We will use the terms network and graph interchangeably. In each case, we are referring to directed graphs, which consist of nodes and directed edges. Each directed edge connects a source node to a target node [3].

another, but how easy is it to get from a relevant document to other relevant documents. The searcher cares about the time to find relevant documents and not the time to travel between arbitrary documents. With a weighted document network, shortest paths now represent the optimal path for a user to follow between two documents.

A network with paths shorter than another network may actually be less navigable. For example, a randomly constructed network of low degree can have shorts paths between most nodes in the network. No user would be expected to navigate well in a random network.

Our other measure of network navigability aims to capture the quality of the similarity measure given the neighborhood it creates for a node. Hierarchical navigation networks such as the Yahoo! or DMOZ directories of web sites are examples of the difficultly of providing good node residue to achieve Furnas' view navigability for large document collections. The links at the top of these hierarchies are broad descriptions of the content available and offer little help in selecting the correct links. While we agree with the need for good link labels, with respect to the network structure, the network should be locally navigable. We are interested in document networks linked primarily at a local level — document to document. A good similarity measure produces links from relevant documents to other relevant documents. A random network would do poorly on this measure of navigability.

We propose using these two measures in combination to evaluate the navigability of a document network. When comparing two similarity methods, the better method should produce a network that is more navigable given both measures. We next discuss the two measures in detail.

2. PROPOSED MEASURES

Given a user's information need or search topic, a perfect similarity method for find-similar makes the topic's relevant documents most similar to each other. This is a restatement of the cluster hypothesis[6]. If a user finds a relevant document, and we have a "cluster hypothesis made true" similarity method, all a user needs to do is to request similar documents and the user will retrieve all of the relevant documents.

To measure the cluster hypothesis, Jardine and van Rijsbergen plotted the distributions of relevant pairs (R-R) and relevant and non-relevant pairs (R-NR) to visually determine the extent to which the cluster hypothesis was true [6]. This same procedure was examined in more detail by van Rijsbergen and Sparck Jones [13]. Griffiths, Luckhurst, and Willet replaced the visual inspection of the distributions with a measure of separation of the two distributions called the *overlap coefficient* [2].

Voorhees [14] pointed out that the relative frequency of very similar R-NR pairs is reduced by the large number of R-NR pairs in comparison to the number of R-R pairs. As an alternative, Voorhees proposed the *nearest neighbor* test, which counted the number of relevant documents found in the n nearest neighbors of a relevant document. Voorhees set $n = 5$. Voorhees' test is equivalent to examining the precision at 5 for the ranked lists produced by using relevant documents as queries. In place of precision at 5, any other retrieval metric such as average precision could be used in a similar manner. Using average precision would result in the computation of a mean average precision (MAP) for each given topic where each relevant document for that topic acts as a query. Voorhees' methodology has an added benefit that it is a measure that is more closely mapped to user notions of distance and separation.

We use Voorhees' methodology to measure the local quality of the document network. For each relevant document, we measure the average precision given the ranking of the document's neighbors formed by taking the weighted links as each neighbor's retrieval score.

A potential problem with the above mentioned measures of the cluster hypothesis is that they fail to accommodate the triangle inequalities that make the cluster hypothesis so appealing. We want to reward a similarity measure for making it easy to get from relevant document A to relevant document C by going first from A to relevant document B and then from B to C even if the similarity measure considers A and C to be dissimilar. To capture this feature of similarity and the value of navigating from document to document, we turn to a measure of the distance between documents measured on the network.

2.1 Document Networks

In a document network, the nodes represent documents in the collection and the edges represent a user's ability to traverse from a given document to another document via some user interface.

We aim to weight the links between documents in a manner that approximates the user's cost to find that link. Given a document-to-document similarity measure embodied by an implementation of find-similar or other user interface feature, for each document in a collection, we can compute a ranking of all other documents in that collection. While at best a crude approximation of user cost, we follow traditional information retrieval metrics and set a link's weight equal to its rank.

In some cases, we will have a document network but will not know the similarity measure. An example of this is the web graph. The links on a web page can be taken to be a ranking of the other web pages. For the links in the page, the top most link is given a rank of 1 and then the next link a rank of 2 and so forth. For many web pages, it may not be obvious from the HTML or even the visual layout of links what that proper ordering of links should be. Thus, an alternative that we follow in our experiments is to give all links a weight equal to the number of links plus 1 divided by 2, i.e. the average ranking. For example, each link on a page with 9 links will get a weight of 5.

Using document rank as our distance also provides us with another benefit. If we assume that shortest paths between relevant documents avoid passing through non-relevant documents, then we can delete the non-relevant documents from the graph and obtain the same results for the shortest paths between all pairs of relevant documents. Deleting the non-relevant documents produces what we term a *relevant document network*.

We obtain a substantial computational savings by deleting the non-relevant documents to form a relevant document network. For the relevant document network, we only need to calculate similarity information for the relevant documents rather than for all documents.

If non-relevant documents were to be on the shortest paths to relevant documents, relevant documents should have non-relevant documents as common neighbors. The cluster hy-

	Non-Relevant		
	10	20	100
Minimum	0.000	0.000	0.003
1st Quartile	0.018	0.024	0.039
Median	0.036	0.044	0.069
Mean	0.057	0.066	0.091
3rd Quartile	0.066	0.080	0.119
Maximum	0.593	0.717	0.543

Table 1: The average overlap coefficient among the top $N = 10, 20, 100$ ranked non-relevant documents in the nearest neighbors of relevant documents for TREC topics 301-450. For example, the mean fraction of non-relevant documents in common is 0.066 or 6.6% for the top 20 highest ranked non-relevant documents.

pothesis says that relevant documents share something in common to make them more similar to each other. In contrast, there is a limitless set of reasons that a document is non-relevant.

As a quick test of the extent to which non-relevant documents are common neighbors of relevant documents, we took the TREC topics 301-450 and we measured the overlap of the first N non-relevant documents occurring in the ranked lists produced by using a relevant document as a query. The document collection for topics 301-450 is comprised of newswire and government documents. While not a web collection, we feel it gives insight to this issue.

Our measure of overlap was the overlap coefficient:

$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

where A is the set of N highest ranked non-relevant documents for relevant document A and similarly for document B. For each topic we computed the average overlap over all pairs of non-relevant documents and then computed summary statistics over all 150 topics. Table 1 shows that the amount of overlap is quite small with the mean overlap for $N = 20$ being 0.066 or 6.6% and three quarters of the topics have an overlap of 8% or less. Thus it appears that non-relevant documents play a role more akin to "noise" than as potentially useful stepping stones between relevant documents.

The assumption that a user will not navigate through non-relevant documents does not hold for document networks such as the web. On the web, links have a mixture of types. Some links go directly to other content rich pages while other links may go to a navigational page. Many navigation pages are not likely to be considered relevant pages in and of themselves. Imagine for example a web site that provides a find-similar link from each content page. The find-similar page is for navigational purposes and may link to a relevant page, but is not in itself a relevant page. By requiring paths to only go through relevant pages, for a similarity measure such as the web graph, we could cut off valid paths.

The relevant document network should only be used in situations where the document network is formed using a feedback-like technique such as find-similar. The relevant document network provides a reasonable upper bound on the shortest path where there is little sense in a user searching for relevant documents starting from a non-relevant document. While a non-relevant document may bridge two relevant documents, how would a user know how to decide between the good non-relevant documents and the bad ones? In a feedback situation, the user would be forced to "lie" to the system and judge a non-relevant document relevant.

2.2 Proposed Shortest Paths Measure

Given a weighted document network, we can efficiently compute shortest paths using Dijkstra's shortest paths algorithm or the Floyd-Warshall all pairs shortest paths (APSP) algorithm.

Distance on our weighted document networks represents the number of documents a user would need to examine by reading link labels such as document titles and summaries before reaching the other document. Other weighting schemes could approximate the individual costs of discovering, evaluating, and traversing links more closely.

Our proposed metric computes on a per topic basis, for each relevant document the mean reciprocal distance of all other relevant documents. Thus, the mean reciprocal distance of relevant document R_i is calculated as:

$$MRD(R_i) = \frac{1}{|R| - 1} \sum_{R_j \in R, j \neq i} \frac{1}{S(R_i, R_j)} \quad (1)$$

where R is the topic's set of relevant documents, $|R|$ is the number of relevant documents, and $S(R_i, R_j)$ is the shortest path distance from R_i to R_j. For each topic, we average the MRD over all the known relevant documents, and finally we average over all topics to produce a final metric. Because our minimum distance is 1, this metric ranges from 1 for the best possible score to 0 for the worst.

This measure is essentially the same as Latora and Marchiori's global efficiency measure [7]. Latora and Marchiori normalize the measure by dividing by the maximum possible efficiency in situations where the maximum efficiency is not 1.

3. EXPERIMENTS

We applied these two measures of navigability to three document networks: the web graph as represented by the wt10g TREC web collection, the document network formed on the same collection using a simple content based document-to-document similarity, and the combination of these two networks.

Soboroff [11] has shown the wt10g collection to have structural characteristics similar to the web. We used the TREC 2001 web ad-hoc topics numbered 501-550. Each topic defines a set of relevant documents. We do not use the topics' titles or descriptions in any way.

We constructed the web graph using the wt10g out_links file. To compute the document-to-document content similarity, we created a maximum likelihood estimated model of each document. We truncated each model to consist of only the document's 50 most probable terms. Using this model, we measure the similarity of the other documents using the KL-divergence. We used Dirichlet prior smoothing and set its parameter to 1500. We stemmed using the Krovetz stemmer and used an in-house list of 418 stop words. We used the Lemur toolkit for our experiments.

The content similarity network is a relevant document network and as such it only has links from relevant documents

	Mean Average Precision		
	Web	Content Sim.	Mix
Minimum	0.000	0.003	0.003
1st Quartile	0.000	0.045	0.040
Median	0.000	0.073	0.067
Mean	0.002	0.101	0.093
3rd Quartile	0.002	0.140	0.131
Maximum	0.022	0.375	0.375

Table 2: The mean average precision for the three document networks where "Mix" is the combination of the web and content similarity networks.

	Mean Reciprocal Distance		
	Web	Content Sim.	Mix
Minimum	0.000	0.002	0.004
1st Quartile	0.003	0.022	0.029
Median	0.004	0.034	0.040
Mean	0.005	0.064	0.071
3rd Quartile	0.006	0.052	0.061
Maximum	0.024	0.750	0.750

Table 3: The mean reciprocal distance for the three document networks where "Mix" is the combination of the web and content similarity networks.

to other relevant documents as described in Section 2.1. We only included content similarity links that had a weight of 100 or less.

Tables 2 and 3 show the results. These tables show the summary statistics across the 50 topics for each measure. For example, in Table 2 the web has at least one topic for which the mean average precision (MAP) was 0.000. A topic with a MAP measure of 0.000 means that the average relevant document has no hyperlinks to any other relevant documents. For example, topic 548 has only two relevant pages. Neither page links to each other. Thus, for topic 548, each page has an average precision of 0 and the mean average precision for the topic is 0. This does not mean there isn't a path from relevant document to relevant document. Also, it may not be the case that the worst or best score for one network is the same topic that is the worst or best for another network.

The web alone does not appear to provide good navigability either locally or globally. The content similarity links appear to be much more navigable. This echoes our other findings where we found that adding 10 content similarity links to web pages brings relevant documents closer to each other [10]. In this other work we gave all links a weight of 1 and only looked at distance on the graph between relevant documents. While a small effect, compared to content similarity alone, combining the two networks hurts the local navigability (MAP) while helping the global navigability (MRD).

4. CONCLUSION

We have proposed measuring the navigability of a document network using two measures. The nodes in the network represent the documents in the collection and the directed links represent the ability of a user to traverse from a source document to a target document. The weight of a link is set proportional to the user's cost to find, evaluate, and traverse the link. One measure captures a local and the other a global quality of the network. The local quality of a network can be measured as follows. For each relevant document, we rank a document's neighbors by their link weights and measuring the average precision of this ranking. The measure of local quality is the mean average precision for the relevant documents. The global measure captures the cost to follow the shortest path, navigating from a relevant document to another relevant document. For each relevant document, we measure the mean reciprocal distance to all other relevant documents. The overall measure is the average of these mean reciprocal distances. Together, these two measures should give us a good understanding of the navigability of a document network and allow us to design similarity methods that construct more navigable networks.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by NSF Nano # DMI-0531171. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] G. W. Furnas. Effective view navigation. In *CHI '97*, pages 367–374. ACM Press, 1997.

[2] A. Griffiths, H. C. Luckhurst, and P. Willett. Using interdocument similarity information in document retrieval systems. pages 365–373, 1997.

[3] N. Hartsfield and G. Ringel. *Pearls in Graph Theory*. Academic Press, Inc., San Diego, 1990.

[4] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR '00*, pages 17–24. ACM Press, 2000.

[5] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516, 2001.

[6] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971.

[7] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):198701, Oct 2001.

[8] P. Pirolli. *Information Foraging Theory*. Oxford University Press, 2007.

[9] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR '06*, pages 461–468. ACM Press, 2006.

[10] M. D. Smucker and J. Allan. Using similarity links as shortcuts to relevant web pages. In *SIGIR '07*. ACM Press, 2007. Poster, to appear.

[11] I. Soboroff. Do TREC web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002.

[12] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR '01*, pages 225–231. ACM Press, 2001.

[13] C. J. van Rijsbergen and K. Sparck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29:251–257, 1973.

[14] E. M. Voorhees. The cluster hypothesis revisited. In *SIGIR '85*, pages 188–196. ACM Press, 1985.

Evaluating Engagement in Interactive Search

Heather L. O'Brien
Centre for Management Informatics
Dalhousie University
Halifax, Nova Scotia, Canada
(1+)(902)494-2515
hlobrien@dal.ca

Elaine G. Toms
Centre for Management Informatics
Dalhousie University
Halifax, Nova Scotia, Canada
(1+)(902)494-8374
etoms@dal.ca

ABSTRACT
In this paper, we introduce a new approach to measuring search – engagement - a holistic metric that encapsulates the user's experience in the process of search. As an outcome measure, it aggregates the user's experience to measure a higher order construct. In addition, we discuss the challenges of measuring the engagement of search as a process, in which the process is composed of a series of engaging episodes.

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

General Terms
Measurement, Performance, Human Factors.

Keywords
Interactive search; engagement

1. INTRODUCTION
The measurement of search has had a protracted history. Initially the focus was on search results or outcomes and their relevance. In the past decades measurement has become dichotomous with document-centric relevance (i.e., topical relevance) and user-centric relevance. More recently an interest has emerged in evaluating the user experience and the activities that occur between activation of a search goal and its outcome.

At the same time, the practice of search has been influenced by developments in self-serve technologies and the emergence of user-system interactivity for the average consumer. The newer approaches now consider the journey to be at least as important as the destination, and that journey is informed by a host of variables: characteristics of the user or information consumer, the task the consumer is attempting to achieve, the resources that exist (and/or are available), the technology that is accessible to the consumer, the situation in which the consumer currently exists, and the environment in which all of the activities take place.

The challenge however becomes how to assess the users' holistic experience on that journey, and to determine the system, task, and contextual variables that contribute to it. Thus far we have been successful at measuring search outcomes. From a systems perspective, the system may be functional (or not), that is, it may be capable of delivering topically relevant to a set of query words results; from an information consumer perspective, the consumer may be satisfied with the job, or may find the results pertinent, relevant and/or useful. But no metric exists to reflect on the totality of that experience in terms of both the process and the outcome. In this paper, we propose a new metric for the assessment of search and discuss the challenges of developing the metric to assess outcome and process. This new metric, engagement, is representative of the information consumer's holistic search experience.

2. MEASUREMENT IN IR
To date, measurement of search typically measures outcomes. Those outcomes are assessed as search effectiveness (e.g., the ability to find the information sought, the user's judgment of the relevance or merit of the search results), search efficiency (e.g., the time taken to complete the search task), and user satisfaction, arguably only one affective element of the experience. These are summative measures that assess the end result. Notably, these are confounded by the multiple definitions of relevance (e.g., Sarecevic, 1997), and tend to reflect a system-centric focus on topical relevance. None of these consider search as a journey.

For the most part, past models of information retrieval have focused on the transition from beginning to end [1]. As such, the process of search has been narrowly defined as a series of actions or search states: e.g., formulating queries, examining search results and web pages; these activities are performed iteratively until the search is abandoned or completed [14], with attention given to the cognitive processes [9] or strategies [2] that accompany them. Yet, we know that interactive searching is more than outcomes – it is an experience.

In addition to metrics, part of the challenge in measuring the search process is the methods used in data collection. Typically the search is evaluated by the usability of a technology or by the consumers in relation to their search goals. In the past decade, attention has moved to transaction logs which, on an aggregate level – the search engine – are limited to what the searcher looked *at*, but not what the searcher looked *for*. These data have included, on the server side, numbers of queries typed and numbers of web pages browsed; more useful data describing the process has been collected at the client-side, e.g., mouse clicks, composition of the queries entered, websites visited, rank of search results consulted and so on.

The methods hat have had the most promise for assessing the journey are think aloud and stimulated recall protocols. *Think aloud*, a technique that requires the searcher to articulate what is going on in head, is not suitable for evaluating complex search tasks and for assessing the process used [16]. In addition, think aloud forces the user to divide their attention into two cognitively competing actions. *Stimulated recall*, on the other hand, is a retrospective account that requires the searcher to 'narrate' the process, post search [5]. This is usually done as video screen captures of the search are being replayed. Immediacy can be a factor in this case with recall accuracy diminishing over time [16].There is also the question of whether users will interpret their actions differently post-search than they did 'in the moment'.

3. A NEW METRIC FOR SEARCH

To understand the process of interactivity with applications, we investigated the holistic experience of web searchers (and contrasted their experiences with those of online shoppers, video gamers and students in an online course). We predicted that all of our users would state that they were able to accomplish their shopping or searching tasks, play the video game, or take part in an online class. We felt that parts of their experiences would stand out in their minds, not because they were able to complete a task, but because they achieved something of a higher, more experiential magnitude; we articulated this as being *engaged*.

Engagement has been deemed "a subset of flow," "flow in a more passive state," and "flow without user control" [20]. Flow, the condition "in which people are so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it" [4, p. 4], has been used to predict and design for flow experiences [6], and to understand users' reactions to and motivations for using applications [8]. We believe that engagement may share some attributes with flow, such as focused attention, feedback, control, goal- and activity-orientation, intrinsic motivation, and the creation of meaning [4]. However, we believe that users can be engaged when they are extrinsically motivated. In addition, the original theory of flow focused on the user, where the task (e.g. painting a picture) or medium (e.g. easel and brush) are irrelevant. If we are to design for engagement with computer applications, we must also determine the significance of the task and system variables in terms of the reactions they evoke in the user.

An extensive literature review was undertaken to ascertain previous uses of the engagement construct. The analysis revealed that engagement is widely used, but has no operational definition. Furthermore, there was no consensus on the attributes of engagement. From this work, we developed a theoretical model of user engagement in order to better understand users' engagement with technology. Our model suggested that engagement is a process-based interaction in which there is a beginning, middle, and end, and that, concurrent with attributes suggested in previous research, is characterized by attention, challenge, feedback, control, novelty, interest, and motivation [11].

To investigate the process of engagement, we conducted a study with seventeen adult computer users of four different technologies. Our semi-structured interviews used the critical incident technique (Flanagan, 1954). Participants were asked to recall a time in recent memory when they felt engaged during an online shopping, searching, or learning task, or while playing a video game. To assist them, semi-structured interview questions were asked about whether the activity was voluntary or mandatory, the duration and expectations for interacting with the technology, and description of the specific topics or tasks associated with the activity. Some interview questions were designed to elicit users' impressions of their experience in lieu of the attributes of engagement identified in previous research (e.g., "How focused were you on the activity?"). Other questions asked interviewees to think about and describe a specific point in the experience (e.g., "Why did you decide to stop searching the Web?"). The same semi-structured questions were administered to all participants, but their order varied to permit a natural flow of conversation between the interviewee and researcher.

We found that users achieved a higher order level of interaction that was characterized by varying degrees of attention, challenge, feedback, control, novelty, interest, and motivation. It was clear that the affective and sensory appeal of the experience was important to participants who expressed a range of emotions that were positive (e.g., feeling excited by the information one is finding during a search) as well as negative (e.g., feeling guilty at the amount of time spent browsing a website for pleasure). This experience was episodic in that participants indicate having multiple episodes of engagement. For instance, one online shopper discussed returning frequently to an online bookstore, typically going to the site to look "for a particular item, or information about a particular item" and other times to browse for pleasure, while another shopper returned to a leather goods website several times, debating whether or not she would purchase a wallet. In addition, each episode clearly had a point of engagement (e.g., shoppers and web searchers wanted to find "a particular item or information about a particular item" often times "out of personal interest", a period of sustained engagement (described by one gamer as physiological response of feeling "emotionally involved"), a point of disengagement (e.g., participants who shopped, searched the Web or played video games chose to "cut themselves off"), and a point of re-engagement. For instance, a searcher disengaged from a particular website that presented information in a cluttered way, but re-engaged with a different website that had more aesthetics and informational appeal. Participants were more likely to indicate willingness to use an application in future when the experience represented engagement.

Notably, each of the attributes was present in all applications, though it appeared that the attributes varied in their degree and manifestation by application. For example, the reasons for participants internal motivations varied according to the desire to learn (webcast viewers, web searchers), locate an item (web searchers, shoppers), or have an experience. A search had multiple episodes of engagement over the course of a search from the initiation to the cessation of the search. At the end of the search, the searcher expressed an overall perception of the engagement quotient of that experience.

Engagement, however, did not apply generally across all searches. In some cases this was due to usability issues. Of a website that kept allowing pop-ups, a Web searcher stated, "I'm like 'forget it', and I never went back to it…It's not that important. Really, if it's that much work for me to get to see it, then it's not worth it." Additionally, a user performed a fact finding search that satisfied

their information goal, but had little experiential value. Yet, for those searches that extend beyond fact-finding and do not run into usability problems, engagement was an element of the process, and, we believe, an important success factor of the experience.

It may also be that only certain components of an experience with an application are engaging. For instance, when shopping online the routine task of filling in shipping and billing information may not be as stimulating as browsing for products, and interacting with different views, styles, colours, etc. of a product. However, to prevent users from disengaging, the routine activities must be facilitated and made easy, so that the more engaging aspects of the experience resonate with users and keep them coming back.

4. ENGAGMENT QUOTIENT

4.1 Measuring Engagement

From our analysis, it was clear that engagement could be interpreted in several ways. First, participants had a sense of engagement as an aggregate variable. That is, engagement could be expressed as a perception of the totality of the experience, e.g., "that was an engaging experience." In addition, participants indicated that aspects of the experience were engaging, representing the episodic-like experience described above.

4.2 Engagement as Outcome

Our first approach was to develop a metric of engagement that represents the outcome – the totality of the experience. Currently we are developing an instrument to assess the "engagement quotient" of web searching and contrasting it with online shopping. This instrument was constructed based on the attributes uncovered in previous literature and the interview study. We generated questions from existing measures, such as the determinants of subjective experience (e.g., challenge: "I feel I have been thoroughly tested") [18] and comments of our interviewees (e.g., attention: "I was intent on what I was doing"). Our scale is undergoing a series of tests to assess its reliability and validity. First, we are administering the survey to reduce it to the most parsimonious set of items that address the attributes of engagement to determine internal consistency of the items for each attribute. Second, a large-scale survey will then be undertaken to understand how these attributes are related to each other and to an engaging outcome. Third, we will administer the survey in two experimental studies: one will assess the survey's generalizability to another search application; the other will juxtapose two interfaces with similar usability and content – one designed to be unengaging and the improved version – to determine the validity of the instrument: results are expected to be different for an engaging and non-engaging interface. The outcome from this process will be an instrument that measures the engagement quotient of a user' experience with a search system.

4.3 Engagement as Process

While our initial work has been valuable in ascertaining the composites of engagement and evaluating its relationship to system and task characteristics, the engagement quotient instrument is not a tool for measuring engagement as a process. The next phase of our research will address: How do we measure the 'ebb and flow' of engagement throughout a search session? How do we match metric with attribute? Is there a single metric that captures engagement as process in the same way that our Engagement Quotient Instrument will do in the case of outcomes? Our thinking in this regard is informed by Saracevic [13].

Saracevic [13] outlined the idea that feedback from the system or the environment can change the course of interaction, causing the user to modify or change interests, goals and strategies. He highlighted the importance of investigating these changes: "Shifts, relatively little explored events, are probably among the most important ones that occur in interaction." Similarly, the four stages of the engagement: point of engagement, engagement, disengagement, and re-engagement may occur subtly during interaction and represent a potential shift in the interaction. At the same time, a search undergoes syntactic shifts, from entering search queries to selecting from lists, to examining content.

The challenge is in first identifying when an engagement episode occurs – the point of engagement. This is not a physical point in the search process, but a conceptual point in which the interactivity. It is equally challenging to identify the disengagement point. We know that at both ends there will be a significant change in the intensity of the attributes. But how will we know when those two points occur? This may not even be something that searchers can conscious say.

The answer, we believe, lies in the use of biometric data that can be collected in a relatively unobtrusive way. Eye monitoring data holds promise in this area. Among the many measures used in eye tracking, pupil dilation has been interpreted as individual's level of arousal and interest in the content of what they are viewing [7] and blink rates have been equated with task difficulty or experience mental workload [15]. Mandryk [10] used galvanic skin response to evaluate affective reactions to video games. We are currently exploring which of this metrics are appropriate.

Traditional process measures (e.g. mouse clicks, time) may be used to identify shifts in the interaction, but they cannot be used to identify a moment of engagement. We speculate that once the points of engagement and disengagement are known, then a searcher may exhibit a pattern of input-response (e.g., mouse clicks) that can be associated with being engaged. Notably examining numerous websites and accumulating a lot of mouse clicks are generally considered inefficient. However, to the information consumer, it may be engaging, especially if the consumer was "caught up" in the interaction and highly interested and focused on the search task. Thus while excessive mouseclicks are normally associated with inefficiency, it is highly probably that the pattern may suggest a period of intense engagement.

In addition to a low level analysis of user interactivity, the answer may also lie in a more holistic examination of the process. Recent work by Oliver and Pelletier [12] addressed the issue of evaluating process using an Activity Theory approach. Activity Theory, which examines the users' intentional actions on an object as mediated by a symbolic (e.g. language) or embedded (e.g. computer) tool, was used to study earning through educational video games. [12] observed activities, actions, and operations during game play and deducted whether or not there was evidence that learning had occurred (for example, if the player tried another activity after the last one had failed). Cognitive Work Analysis (CWA), which examines the strategies people use to approach their work and the social and organizational constraints they encounter [3], may also be a useful framework to operate in when studying process. It takes into

account the user, the system, the techniques they employ to carry out tasks, and the greater social context in which they operate in. [19] proposes three levels of techniques for operationalizing CWA: input/output constraints (e.g efforts initiated to carry out the task), sequential flow (e.g. charting the sequence of procedures from task start to finish, and timeline, which places the actions in appropriate order indicates their duration. This framework may offer some insights for exploring the task aspects of search engagement. Specifically, what are the input and output constraints on users as they try to become engaged with different systems? How can examining users' sequential flow and timeline contribute to an appreciation of engagement as process, and the peaks and valleys users experience during a single encounter?

While Activity Theory and Task Analysis do not assess the actions users in conjunction with their intentions or account for prior knowledge, they do offer a useful framework to assess process, and, more specifically, the engagement during the search process. For instance, are there behavioural cues unconscious to the user, such as facial expressions that might denote concentration, which might help us to understand process? Certainly the idea of observing users is not novel, but are there ways in which we could triangulate it and other process data collected in key strokes and mouse clicks? And could this strategy be streamlined and implemented in interactive information retrieval research to standardize our methodological processes so that, as a community, we can acquire a more holistic and corroborative sense of search process?

5. CONCLUSIONS

Today we are working toward viewing information search through a contextual lens that incorporates the physical, cognitive, and affective components of the user, the nature of the task they are performing, and the usability of a system, all of which are influenced by situational demands, encapsulated by the information environment. Search under these conditions is not a simple transaction; it is a form of information interaction, an integrated process of locating information (via querying and browsing), within the parameters of a context or situated action, facilitated by previous experience and knowledge [17]. We need to consider search as a holistic experience and develop metrics that will assess that experience. Our work has identified a number of key attributes that we have combined in the concept of engagement. This concept is measured in the aggregate but it is challenging to measure as a process.

6. ACKNOWLEDGMENTS

This research was supported by SSHRC and Killam Scholarships to the first author, and grants from the Canada Research Chairs Program and NSERC (NECTAR) to the second.

7. REFERENCES

[1] Beaulieu, M. (2000). Interaction in information searching and retrieval. *Journal of Documentation, 56*(4), 431-439.

[2] Belkin, N., Cool, C., Stein, A. & Thiel, U. (1995). Cases, scripts, and information-seeking strategies. *Expert Systems with Applications, 9*(3):379–395.

[3] Benyon, D. (1992). The role of task analysis in system design. *Interacting with Computers, 4*(1), 102-123.

[4] Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper & Row.

[5] Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

[6] Finneran, C. M., & Zhang, P. (2003). A person-artefact-task (PAT) model of flow antecedents in computer-mediated environments. *International Journal of Human-Computer Studies, 59*, 475-496.

[7] Granka, L. A., Joachims, T., & Gay, G. K. (2004). Eye-tracking analysis of user behaviour in WWW search. In *Proceedings of SIGIR*.

[8] Konradt, U., & Sulz, K. (2001). The experience of flow in interacting with a hypermedia learning environment. *Journal of Educational Multimedia and Hypermedia, 10*(1), 69-84.

[9] Ingwersen, P. (1992). IR Interaction - The cognitive turn. In *Information Retrieval Interaction* (pp. 123-156): Taylor Graham Publishing.

[10] Mandryk, R. L. (2005). Evaluating Affective Computing Environments Using Physiological Measures. In *Proceedings of CHI*, Portland, OR.

[11] O'Brien, H. L., & Toms, E. G. (2005). Engagement as Process in Computer-Mediated Environments. Poster presented at the Annual Conference of *American Society for Information Science and Technology*, Charlotte, N.C.

[12] Oliver, M. & Pelletier, C. (2006). Activity Theory and learning from digital games. In Buckingham, D. & R. Willet (Eds.) *Digital Generations. Children, Young People and New Media*. Lawrence Erlbaum, pp. 67-91.

[13] Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the ASIS, 34*, 313-327.

[14] Spink, A. (1998). Study of interactive feedback during mediated information retrieval. *JASIS&*T, 45(5): 382-394.

[15] Takahashi, K., Nakayama, M., & Shimizu, Y. (2000). The response of eye-movement and pupil size to audio instruction while viewing a moving target. In *Proceedings of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, FL, USA.

[16] Toms, E.G. (1997). *Browsing Digital Information: Understanding the 'Affordances' in the Interaction of User and Text*. Unpublished Ph.D. Dissertation, University of Western Ontario.

[17] Toms, E. G. (2002). Information interaction: Providing a framework for information architecture. *JASIST*, 53(10), 855-862.

[18] Unger, L. S., & Kernan, J. B. (1983). On the meaning of leisure: An investigation of some determinants of the subjective experience. *The Journal of Consumer Research, 9*(4), 381-392.

[19] Vicente, K. J. (2000). *Cognitive Task Analysis*. Mahwah, NJ: Lawrence Erlbaum

[20] Webster, J., & Ahuja, J. S. (2004). Enhancing the Design of Web Navigation Systems: The Influence of User Disorientation on Engagement and Performance. Unpublished manuscript

Clickthrough based measures of search engine performance

Erik Graf
University of Glasgow
17 Lilybank Gardens
Glasgow, United Kingdom
graf@dcs.gla.ac.uk

Craig Macdonald
University of Glasgow
17 Lilybank Gardens
Glasgow, United Kingdom
craigm@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
17 Lilybank Gardens
Glasgow, United Kingdom
ounis@dcs.gla.ac.uk

ABSTRACT

In this paper we introduce a framework for logging user interactions with an intranet search engine. In this initial study we evaluate if, through an extension of the logged attributes, we can gain a more accurate picture of the search process, and moreover practically apply this data as a means of measuring the system's performance. We test the usefulness of our logging scheme by relating several proposed measures of system performance to a manually generated ground truth.

General Terms: Performance, Experimentation

Keywords: Query Log, Clickthrough, User Interaction

1. INTRODUCTION

Clickthrough data, stored records of user interactions with an Information Retrieval (IR) system, have been the subject of various recent research.

Early research efforts of using recorded user data have focused on using it as a means of acquiring a picture of the search process. In the domain of Web search, most of the early research in this area has concentrated on making use of this kind of data in order to uncover basic statistics of Web search. Moreover, various approaches have explored the direct integration of clickthrough data into the retrieval process to improve the precision of an IR system. In this paper, we report on a framework for the logging of user interactions in the context of an organisation's intranet, namely the Computing Science Department (DCS) of the University of Glasgow. Our proposed logging scheme is based on the introduction of a taxonomy of loggable items. The main focus of our research lies in identifying log-based measures that will allow us to infer the underlying system's performance, and improve our understanding of the users behavior. Our ultimate goal is to be able to use this knowledge to tailor the system to better meet its users needs, and to devise new system improvements. The remainder of this paper is organised as follows. Section 2 examines related work. In Section 3 we describe the underlying IR system and our logging scheme. Section 4 introduces our research hypothesis. Section 5 reports on the conducted experiments. In Section 6 we report on our results and conclude the paper in Section 7 by providing ideas for future extensions of this work.

SIGIR WISI '07 Amsterdam, The Netherlands

2. RELATED WORK

The analysis of clickthrough data as a means of determining the basic metrics of Web search has been performed on several log files obtained from commercial WWW search engines. An overview and discussion of several studies focused on analysing log files, and a proposed framework for research concerning Web search log analysis is given by [4]. The direct integration of click data into the retrieval process has been explored in various research projects aimed at increasing the performance of IR systems. Xue et al. [8] for example explored the usage of data fusion techniques to combine query data with the original document content and also the linear combination of full text score and clickthrough data score. Moreover several studies based on utilising implicit feedback in order to measure user satisfaction and user interest have been conducted. In [2] Fox et al. explored the usage of implicit feedback, including clickthrough data, that was collected via an extension to participating users' browsers as a means of measuring a user's satisfaction. Our research differs from most of the reported work in the domain of pure log file analysis in that we are applying an extended logging scheme that is tracking a total of twelve attributes as opposed to the commonly reported sets of three to four logged features. We limit our observation of user behavior to the direct interaction with the search engine and do not apply any form of browser instrumentation. We do not integrate clickthrough data directly into the retrieval process. The main focus of our research lies in acquiring knowledge that can be used to improve the retrieval process.

3. LOG MECHANISM AND UNDERLYING IR SYSTEM

Our logging framework is operating on top of the search engine of the Computing Science Department of the University of Glasgow. The underlying corpus consists of around 60 thousand documents, which are crawled and indexed on a daily basis. The engine can be accessed from the department's home page as well as from various other sites such as the home page of the Information Retrieval group, and various pages of individuals. For each result listing on the result page the engine provides the title, the URL and a query-biased snippet summarizing the document. It is of note that the corpus consists of two distinct parts: The collection of publicly accessible documents, and the collection of intranet web pages which can only be accessed by users from within the department. The majority of the Web pages in the externally visible corpus are related to either teach-

ing and course information, research, or personal and group homepages. In contrast to that, the majority of pages within the internally accessible corpus is dedicated to administrative and financial issues. The search engine is based on the Terrier Information Retrieval platform [6]. Conducting the research in the domain of our own department features several key advantages such as for example, the potential access to several key groups of people such as the authors of the Web pages, the administrators of the search engine, and the group of internal users. Moreover a high familiarity with the information expressed in the documents contained in the corpus eases manual assessment of the recorded log data.

Before introducing our logging scheme we would like to cast light on the logging sets reported in previous research studies. It is noticeable that many of those studies [1, 5, 7] do not explicitly mention the reasoning behind picking the specific set of listed attributes. Common to most reported research is the logging of the following attributes: Timestamp, unique user identifier, and query. As outlined by the reported research in the previous studies, this set allows for query-related and basic user-related analysis. Another common set of logged attributes that can be found in literature consists of the following fields: Timestamp, user identification, query, and URL of clicked document. As stated before, our initial aim lies in being able to measure system performance and to detect system anomalies. Our long term goal lies in utilizing the logged data to increase the users satisfaction by devising improvements tailored directly to the users needs, and by identifying ways to customize the search process to specific groups of individuals such as for example students, researchers, or staff. With these targets in mind we devised our set of loggable attributes. We log a total of 12 features. In our quest to identify potential attributes we have devised a taxonomy of loggable features that is shown below.

- user-related attributes: unique user id (1), internal-external (2)
- request-related attributes: query (3), submission origin (4), query submission timestamp (5)
- engine-related attributes: total number of retrieved documents (6).
- result-related attributes: snapshot of the search engine result page (SERP) (7), URL of clicked document (8), rank of clicked document (9), click occurrence timestamp (10), SERP-navigation (11), snapshot of clicked on document (12).

Our initial set of logged features is outlined in Figure 1. The bracketed numbers listed after each attribute correspond to those given in Figure 1. In our experience, structuring attributes in this way can ease the identification of potential features that could be logged. By taking a look at the taxonomy it is easy to identify further possible extensions to our initial logging scheme. Concerning user-related attributes for example, potential additional features could be the used browser and IP address. However, in this case we neglected logging these attributes out of respect to the users privacy. In the future we plan to extend the taxonomy by adding a category of corpus-related items that will log attributes such as the number of added or deleted documents, or the general level of change in the corpus.

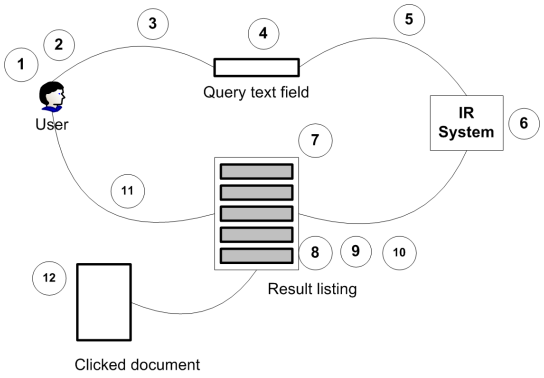

Figure 1: Visualization of the applied logging scheme

The unique user id (UID) is tracked via the usage of a cookie. The internal-external attribute logs whether a request originated from an IP address of the department or an external IP address. The logging of this attribute goes back to the fact that, as mentioned before, parts of the corpus are off limits for users with an external IP-address. The snapshots of visited SERPs are retained in the form of a compressed copy of the HTML source. The SERP-navigation attribute logs a user's browsing within the search results the engine returned. We create a log entry denoting the SERP, identified by the lowest result rank on this page, every time a user navigates forth or back within the result listings. In addition to recording the timestamp of query submission, we also record the timestamp of each click on a result. We save a compressed copy of every document a user clicked on within a result listing. Web documents are subject to change on a frequent basis, and the documents in the DCS corpus are no exception to this. Specific parts such as personal homepages of research and teaching staff, and administrative pages are subject to very frequent changes. Therefore, by retaining a snapshot of each clicked document we are able to perform more accurate retrospective analysis of the relevance of a clicked on document. In the course of the experiments that will be described in the remainder of this paper it is important to define our understanding of a user session. Following the approach of [7], a number of log entries is interpreted as belonging to the same session if all entries feature the same UID and the time interval between two consecutive entries is less than 5 minutes. He et al. [3] have explored more sophisticated ways of session identification.

4. MEASURING SYSTEM PERFORMANCE BASED ON CLICKTHROUGH

In this initial study our hypothesis is that the information contained within the clickthrough log data can be put to use in order to measure the performance of the underlying system. We manifest our hypothesis through the proposal of interaction-based metrics:

- **Percentage of sessions resulting in clicks:** This measure is based on the simple assumption that a higher percentage of sessions resulting in clicks indicates an improved performance of the system. One would assume that if a user does not bother to click

on a single document within a session it means that he or she did not deem any of the retrieved documents relevant. Indeed by interpreting a session without a click as a failure, and a session with a click, independent of the relevance of the clicked on document, as potentially successful, we hope to be able to use the ratio of failed and potentially successful sessions as an indicator of the system's performance.

- **Delta between query submission and first click:** This measure is based on the logging of the query submission timestamp and the click occurrence timestamp. We devise this measure on the assumption that a result listing containing many highly relevant documents will result in a smaller time delta between query submission and the first click occurrence as opposed to a result listing containing no or only partially relevant documents.

- **Query Reformulation:** This measure is based on the assumption that the level of reformulations within a session can serve as an indication of session success. We test the usefulness of this measure based on the intuition, that a higher level of reformulations denotes that the user deemed the returned results as not relevant to his or her query. Therefore we interpret a higher frequency of reformulations as a potential indicator of problems in the search process.

- **Ratio of clicks and undesirable actions:** Based on the intuition that from a user's point of view, a click on a result can be interpreted as a desirable action, and navigation and query reformulation as undesirable actions, we assume that the ratio of clicks versus undesirable actions can help to predict a system's performance.

In the remainder of this paper we assess the extent to which these measures are able to predict the system's performance by relating them to a manually generated ground truth (GT).

5. EXPERIMENTS

For our experiments we use a log file spanning a time period of 32 days ranging from 29th March 2007 to 29th April 2007, containing a total of 1.726 entries. An entry relates to one of the following user actions: A query submission, a click on a result, or navigation within the result pages. The average number of user actions per day is 53.9. In the examined time period there are 789 unique user ids in the log file. We believe that the high number of UIDs is caused by the fact that most of the computers that are accessible by students are configured to not retain cookies. The average number of recorded actions per user is 4.29. Applying the automatic session identification method [7] resulted in 901 sessions. During the examined time span users have clicked on 336 results. The average query length is 1.88, and the number of unique query terms is 526. The level of traffic in the observed time period is rather low, mainly due to the facts that the launch of the search engine has not been advertised within the department, and moreover the observed time period was a student holiday period. Therefore due to the low volume of available click data our reported results should be interpreted as anecdotal.

For our ground truth we manually assessed the success of sessions. A session was considered successful when at least one clicked on document within that session was assessed as being relevant with respect to the query the user entered in the search box. Concerning the assessment, the logged attributes of the request-related and result-related categories have proven to be very useful. In our experience the availability of both the snapshot of the clicked on document as well as the snapshot of the SERP eased performing this task, and potentially increased the accuracy of the assessments. In the observed time period we manually assessed the topical relevance of a total of 336 documents with respect to their queries.

6. RESULTS AND ANALYSIS

Figure 2 shows over the timespan of the experiments the percentage of sessions that resulted in clicks, which we interpret as automatically assessed successful sessions, and the percentage of manually assessed successful sessions that form our ground truth (GT). In general, we observe a good correlation between the proposed predictor and our GT. In particular, there is a strong correlation between the predictor and the GT of Spearman's ρ of 0.592. As can be seen in the Figure, in the time period ranging from the 3rd April to the 5th April, the percentage values for both measures are very low. This time period corresponds to a known outage of the system, where due to a configuration error, the system returned almost random results. A manual examination of 50 SERPs showed that the returned results were indeed random, retrieving documents that did not even contain the query terms. It is observable that the similarity of both graphs is higher for time periods with lower percentage values. This is intuitive if we take into account that interpreting sessions with a click as successful is overly optimistic.

Figure 3 shows the GT and the average daily time difference between the query submission time stamp and the first click occurrence time stamp in seconds (Reaction Time). It is visible that both graphs are not as closely correlated as the previous measure, and indeed overall correlation of ρ is only 0.214. Reaction time and the percentage of successful sessions do not seem to be directly linked to each other. An exemplary manual investigation of the logs indicates that the reaction time seems to be influenced by a couple of factors, including for example: The specific user, the type of search (i.e. navigational, informational, transactional), and the rank of the first clicked on result. Future research that will correlate the reaction time to these factors may provide more insight.

Figure 4 compares the percentage of sessions with query reformulation to our assessed GT. This predictor shows the lowest level of correlation with Spearman's ρ of 0.04. However it can be observed that for the period of the known system outage, as well as for other days (12th of April and 25th of April) that show a low percentage of successful sessions, a high percentage of sessions had reformulations. Future research that will test alternative interpretations of reformulation levels and explore the correlation of specific types of reformulations (e.g. addition of terms, removal of terms) should provide more insight.

Figure 5 shows the ratio of clicks and undesirable user actions and our assessed GT. As can be observed from the figure the level of correlation is higher than for the previous

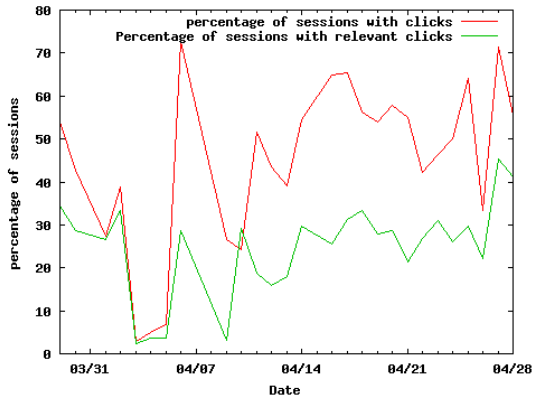

Figure 2: Comparison of successful session rates for automatic and manual classification

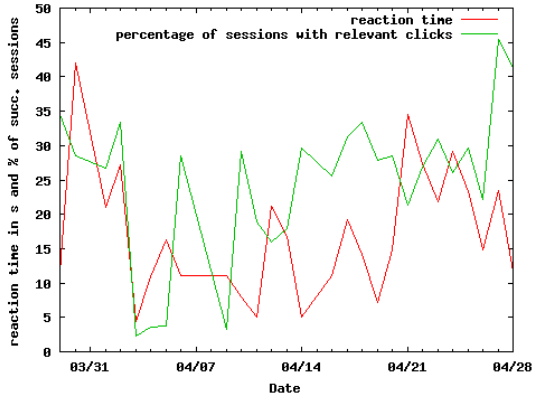

Figure 3: Time delta of query submission and first click and the percentage of manually assessed successful sessions.

two predictors with ρ of 0.357. Especially for the period of the known system outage, a high level of similarity of both graphs can be seen.

Our observation is that the measure based on the percentage of sessions resulting in clicks shows the highest level of correlation with our ground truth. It seems promising that the known system outage is also indicated by our other measures. Concluding we can say that our proposed logging scheme, proved to be very useful for generating our ground truth, and also for devising promising potential predictors of system performance.

7. CONCLUSION

In the future we wish to devise more predictors based on clickthrough evidence. Moreover the initial study presented in this work could naturally be extended to assess the correlation of the predictors with more classical IR evaluation measures such as Mean Average Precision and Mean Reciprocal Rank. In the long term we would like to explore the utilization of clickthrough data to detect difficult queries and to distinguish between specific groups of users, such as for example potential students, staff, and external researchers, in order to be able to tailor the retrieval process both to specific types of queries and specific individuals.

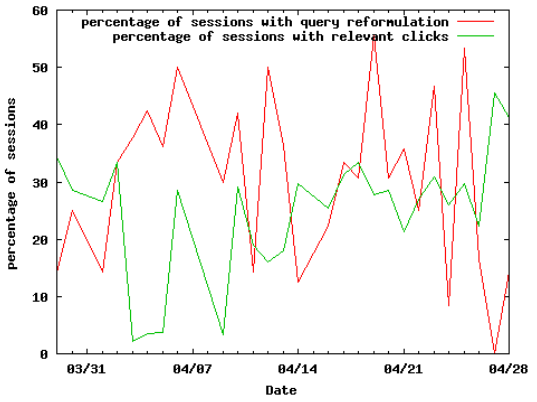

Figure 4: Comparison of percentage of sessions with reformulation and manually assessed ground truth

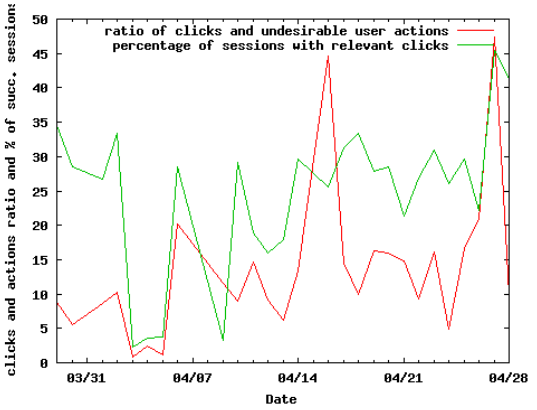

Figure 5: Comparison of the click to actions ratio and manually assessed ground truth

8. REFERENCES

[1] F. Cacheda and A. Vina. Understanding how people use search engines: a statistical analysis for e-business. In *Proceedings of the e-Business and e-Work Conference and Exhibition 2001 (pp. 319325). Venice, Italy, October*, 2001.

[2] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[3] D. He, A. Goker, and D. J. Harper. Combining evidence for automatic web session identification. *Inf. Process. Manage.*, 38(5):727–742, 2002.

[4] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):235–246, 2001.

[5] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, 2000.

[6] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of OSIR Workshop '06*, 2006.

[7] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[8] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing Web search using Web click-through data. In *Proceedings of CIKM'04*, pages 118–126, 2004.

Comparing System Evaluation with User Experiments for Japanese Web Navigational Retrieval

Masao Takaku
Research Organization of Information and Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
masao@nii.ac.jp

Yuka Egusa
National Institute for Educational Policy Research
6-5-22 Shimomeguro, Meguro-ku, Tokyo, Japan
yuka@nier.go.jp

Hitomi Saito
Aichi University of Education
1 Hirosawa, Igaya-cho, Kariya-shi, Aichi, Japan
hsaito@auecc.aichi-edu.ac.jp

Hitoshi Terai
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
terai@nul.nagoya-u.ac.jp

ABSTRACT

We conducted a search experiment targeting 31 users to investigate whether the performance evaluation metrics of IR systems used in test collections, such as TREC and NTCIR, are comparable to the user performance and subjective evaluation. We selected three systems with high, medium, and low performance values in terms of nDCG, MRR and Prec@10 metrics from among the retrieval systems that participated in the NTCIR-5 WEB task, and then selected three topics. The results of the experiment showed no signi cant di erences between these systems and topics in the completion time for each search. Furthermore, none of the results of the users' evaluations corresponded to the results of the batch system evaluations. These results indicate a need for new evaluation metrics that correspond to the users' evaluations.

1. INTRODUCTION

The performance evaluations for information retrieval (IR) systems are extremely important in today's Internet environment, where a wide variety of IR systems are provided and used. The performance evaluations of IR systems are said to have begun with the Cran eld experiments, and the eld later expanded to include evaluation experiments the use large-scale test collections, such as TREC and NTCIR.

In recent years, however, these evaluation methods have been called into question. In researches conducted by Hersh et al.[1], and Turpin and Hersh[4], it was reported that in the TREC 7-9 Interactive Track, batch evaluations did not correspond to the user evaluation results. Turpin and Shoeler[5] recently conducted more large scale tests on a simple Web information nding task, and showed that the system's MAP metrics and user performance did not correlate with each other.

This results suggest that the results of performance metrics in past system evaluations do not necessarily match the results of subjective evaluations and perception characteristics in user evaluations. However, there has been little study that has focused on this gap between the batch and user evaluations. It is necessary to gather evidence using other types of tasks or test collections to investigate why batch and user evaluation do not match, or what can be done to develop performance evaluations that are closer to the users' evaluations. Most of the previous researches have been based on TREC data, and there have been almost no studies using other large-scale test collections.

Based on the above situation, we compared user evaluations with batch evaluations in the NTCIR-5 WEB Navigational Retrieval task (Navi2)[2] for our current research. We report the preliminary results from our experiments in this paper, and introduce about the di erences and similarities between our results and those of prior researches.

2. METHODS

2.1 Subjects and Design

A total of 31 subjects (21 males and 10 females) participated in the experiment. The subjects were recruited from three universities; 12 were faculty members, 8 were graduate students, and 11 were undergraduate students. The backgrounds of the subjects varied, but the faculty members were from the nursing science eld, the graduate students were from the information science eld, and the undergraduate students were from the education science and information science elds. The average age of the subjects was 25.6 ($SD = 4.99$), and the average Internet usage time was 2.98 hours per day ($SD = 2.43$). They were unfamiliar with our dataset and this was their rst time to use it.

The experiment was conducted using a 3×3 mixed design. The rst factor was the three topics, and the second factor was the three systems (both were subject internal factors). As indicated in Table 1, the subjects were allocated into

WISI 2007 XX XXXX, XXXXX

three patterns (S_a, S_b, and S_c) combining the topics (movie, shopping, and restaurant) and systems (high, middle, and low). During the experiments, the subjects were randomly assigned to each pattern, and each pattern had ten or eleven subjects.

Table 1: Experimental design

	High	Middle	Low
Movie	S_a	S_c	S_b
Shopping	S_b	S_a	S_c
Restaurant	S_c	S_b	S_a

2.2 Materials

Three topics and three systems were selected from the NTCIR-5 WEB task for use in this experiment.

From among the systems participating in the NTCIR-5 WEB task, three systems were selected as having normalized discounted cumulative gain (nDCG), reciprocal-rank (RR), and precision at 10 (Prec@10) values corresponding to high, middle, and low (TNT-3, ORGREF-C20-P2, and ORGREF-GC1, respectively). Three topics (movie, shopping, and restaurant) were selected as having similar nDCG values within a single system (topic numbers 1196, 1296, and 1367, respectively[1]). Figure 1 shows an English translation for the shopping topic.

```
<TOPIC><NUM>1296</NUM>
<TITLE>Seiyu, online supermarket</TITLE>
<DESC>I want to visit to Seiyu's online supermar-
ket page.</DESC>
<NARR>
<BACK>I would like to go to shopping at Seiyu's
online supermarket.</BACK>
<RELE>Seiyu's online supermarket page in the offi-
cial Seiyu website is relevant.</RELE>
</NARR>
</TOPIC>
```

Figure 1: Shopping topic (English translation)

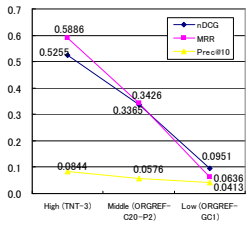

Figure 2: nDCG, MRR and Prec@10 performance measures of three runs with 269 topics

Figure 2 shows the systems' nDCG, MRR and Prec@10 values for all 269 topics from the NTCIR-5 WEB, and Fig. 3 shows these measures for each selected topic. In the NTCIR-5 WEB, graded relevance levels are assigned in relevance

[1]All topics are available at the NTCIR website: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/cdrom/WEB/NAVI2/ntcweb5-navi-frun-topics-1.euc.txt (only in Japanese)

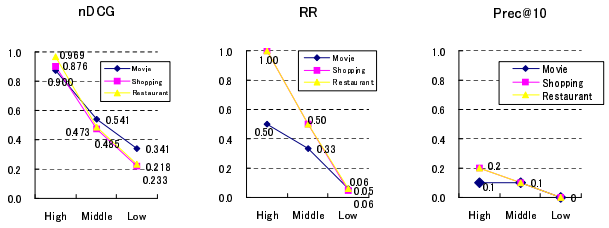

Figure 3: nDCG (left), RR (middle) and Prec@10 (right) performance measures of three runs for each topic

judgments. We calculated these evaluation metrics based on a weighted values. For nDCG, multiple relevant levels were weighted with (A, 10) and (B, 1). For MRR and Prec@10, multiple relevant values were calculated at rigid level: (A, 1), (B, 0). Additionally, duplicates of the relevant documents were judged at relevance judgments. For nDCG and Prec@10, if duplicate relevant documents were found several times, they were regarded as irrelevant except the first one found. In terms of the NTCIR-5 WEB test collection having 269 topics, as shown in the Fig. 2, there were significant differences between these three runs (high, middle, and low), and these results were confirmed by using a pair-wise t-test for the three evaluation metrics (all results at $p < .001$).

As can be seen in Fig. 3, the high run had the highest nDCG, MRR and Prec@10 values among the three topics, and the low run had the lowest nDCG and RR values among the three topics.

2.3 Procedures

Figure 4: Search result interface for a query

During the search experiment, the subjects were instructed to read a topic's description, background, and relevant criteria, and then to explore the Web, which is in reality the NW1000G-04 dataset, through our Web-based user interface for NTCIR-5 WEB run results[3] to find a relevant page. We instructed the subjects to bookmark the relevant page if they found a relevant page, and then their task for the topic would be complete. Our Web-based interface is shown in Fig. 4. We assumed that these settings could partially sim-

ulate their daily search environment in a real Web search engine.

First, the subjects were given a questionnaire on their demographics and experiences in using the Internet and computers. After an introduction to the search tasks, the subjects performed a practice search. After this, the topics were presented in random order according to the conditions of the experiment shown in Table 1. During each search task, the subjects were not informed that they were using di erent search systems each time, because we tried to assure the subjects not to have a bias against the systems, and to keep their mind neutral during the experiment. The search topics were displayed on a Web browser. When the search began, the following information was displayed: The purpose of the search (<DESC> in Fig.1), background (<BACK> in Fig.1), relevance criteria (<RELE> in Fig.1) and the link to the search result pages (SRPs). The subjects could jump to the SRPs whenever they wanted. The SRP was composed of a list of ten pages at one time, and its interface was similar to that of a usual search engine which has a title, URL, and snippets of pages (See Fig. 4). Note that we created and used the static SRPs from the search result runs submitted to NTCIR-5 WEB. The subjects looked for pages that appeared to match the topic context from this list of SRPs. The search ended when the relevant page was found, and the subjects were asked to evaluate the search. The searches for each topic were evaluated using a 5-point scale based on the following items: (1) Search di culty, (2) Satisfaction with the results, (3) Con dence in the results, (4) Appropriateness of the system for that topic, and (5) Prior knowledge of that topic.

At the end of the experiment, the subjects were informed that a di erent search system had been used for each of the three topics, and they were then asked to ll out the following two evaluations using a 3-point scale: (1) Performance of the three systems, and (2) How di cult it was to understand the search itself.

3. RESULTS AND DISCUSSIONS

3.1 Agreement with the official assessments

In general, the relevance judgments between people presented quite di ering results[6]. From our experience with Web navigational retrieval, the agreements between people are rather high for the navigational task than for the other tasks (e.g. informational task).

Table 2: Agreement rates at rigid (top) and relaxed (bottom) level between subjects' judgments and official judgments

Rigid level:

	High	Middle	Low	(total)
Movie	3/10	2/10	0/11	5/31
Shopping	7/11	8/10	8/10	23/31
Restaurant	10/10	10/11	6/10	26/31
(total)	20/31	20/31	14/31	54/93

Relaxed level:

	High	Middle	Low	(total)
Movie	9/10	8/10	11/11	28/31
Shopping	8/11	8/10	9/10	25/31
Restaurant	10/10	10/11	6/10	26/31
(total)	27/31	26/31	26/31	79/93

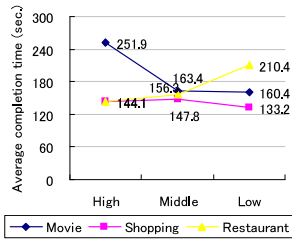

Figure 5: Average search time for each system and topic

In our experiments, the subjects reported on a relevant page. The agreement rates between the relevant pages reported by the subjects and the o cial assessments from the NTCIR-5 WEB[2] are shown in Table 2. The di erences among the systems were tested by using the Chi-square test and two-way ANOVA, but no signi cant di erences were observed. Only at the rigid level[2] was there a signi cant main e ect of the topic ($F(2, 84) = 26.887, p < 0.001$), and the Movie topic showed signi cantly lower agreement rates than those of the o cial relevance judgments ($MSe = 0.152, p < 0.001$). At the rigid level, the agreement rates on the topic Movie were rather low, although the agreements on the other topics were quite high. At the relaxed level, on the other hand, the agreement rates were high for all the topics. The low agreement rate for the Movie topic was caused by its absence in the relevance criteria in its topic description. However, most of the subjects could nd at least one relevant or partially relevant page. From the relaxed level results, our results seem to be reasonably consistent with those from the original NTCIR-5 WEB, and we could see that our experiment settings could successfully simulate the navigational retrieval settings of the original NTCIR-5 WEB.

3.2 Completion time

Figure 5 shows the subjects' average search completion time in seconds for each system and topic. We can see from this plot that for the movie and shopping topics, the high run had the longest execution time, but in the case of the restaurant topic, the search time grew longer from the high to the low run. There was no signi cant di erence, however, between the systems and topics.

These results suggest that even when the evaluation data in the NTCIR-5 WEB task is used, the system performance results based on the batch evaluations do not match the results of the user performance in the user experiments. The search completion time is one of the user performance measures, because one of the purposes of information access systems for end users is, in general, to quickly retrieve information. From this viewpoint, our results for the search completion time show that subjects can get information almost in the same period of time whether or not they use a batch high-performance system.

Turpin and Hersh[4] reported that users performed equally well on signi cantly di erent batch evaluation systems in

[2]The NTCIR-5 WEB had graded relevance judgments, in which a document was assessed as relevant, partially relevant, or irrelevant. In a rigid level analysis, partially relevant documents are seen as irrelevant documents, and in the relaxed level, they are regarded as relevant.

Table 3: Summary of subjective evaluation analysis by two-way ANOVA

	System×Topic	System	Topic						
	p-value	p-value	p-value	F(2,84)	Main e ect			MSe	p-value
Di culty	0.276	0.364	0.057+	2.972	Restaurant	>	Shopping	1.238	0.023*
Satisfaction	0.798	0.210	0.052+	3.703	Shopping	>	Movie	1.019	0.015*
Con dence	0.742	0.771	0.021*	4.058	Shopping	>	Movie	0.961	0.001**
Appropriateness	0.559	0.934	0.023*	3.961	Shopping	>	Restaurant	1.127	0.001**
Prior knowledge	0.353	0.525	0.183	1.734	—		—	—	—
Performance	0.733	0.730	0.000**	8.894	Movie	>	Restaurant	0.481	0.000**
					Shopping	>	Restaurant		0.000**
Di culty in understanding	0.843	0.105	0.013*	4.577	Movie	>	Shopping	0.523	0.003**

+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$

terms of MAP. Turpin and Shoeler[5] also reported that the MAP values did not match with their user performances, based on the automatically created rankings. The di erences between those studies and the current one are threefold. First was the dataset that was used. We used the NTCIR-5 WEB dataset, which consists of Japanese topics and mainly a Japanese Web dataset, while prior researches used a TREC dataset, which consists of English topics and newspaper articles or a Web dataset in English. Second was the tasks that were conducted. We conducted experiments for a Web navigational retrieval task, while prior researches were either for a recall-oriented task, a Q&A task, or a Web information- nding task. The third di erence was in the evaluation metrics that were used. The NTCIR WEB used a multi-graded relevance level, and reported the DCG and MRR as its o cial metrics. We used the nDCG, MRR, and Prec@10 as the system evaluation metrics, while prior studies used the MAP and Prec@n as major metrics for the batch evaluation.

Although the datasets, tasks, and metrics were changed, their results and ours are quite similar. That is, from the user experiments, we found that the di erence in batch system evaluations that were used does not directly result in the di erences in user performance, which is measured by the time taken to complete each task.

3.3 Subjective evaluation

We conducted two-way ANOVAs with topics and systems as the between-subject factors regarding the seven subjective evaluation points answered by the subjects (task di - culty, satisfaction with the result, con dence in the result, appropriateness of the system, prior knowledge of the topic, system performance, and di culty in understanding of the task).

The results from a statistical test for the systems and topics are given in Table 3. The results from this analysis showed that several signi cant di erences were found between the topics, except for the prior knowledge of a topic. However, it did prove that no signi cant di erences were found among the systems, or among the topics×systems. In summary, these results suggest that users are more aware of the di erences between the topics than they are of the di erences in the performances of the di erent systems.

From the subjects' comments, we noticed that for some topics there was di culty in nding a relevant document. In some parts, this was caused by the limitation in the experimental environment. For example, the restaurant topic was more di cult than the other topics in terms of di culty, appropriateness, and performance. The relevant pages for the restaurant topic were easily found and the relevant pages had several embedded images in them, but the NW1000G-04 dataset only gathered in text format. So, the subjects noted a poorer performance for the restaurant topic. Another example is with the movie topic. The movie topic does not have relevance criteria in its topic description. So, the subjects noted lower scores in the con dence, satisfaction, and ease in understanding for the movie topic.

4. CONCLUSION

In looking at our experimental results, in the case of the NTCIR-5 WEB task, the nDCG, MRR, and Prec@10 system performance measures did not match the users' performance and subjective evaluations. These results could be viewed as suggesting a need for the development of new evaluation metrics that more closely correspond to the user evaluations. In the future, we will analyze the subjects' tracking log data during the experiments and other supporting information. In addition, since the size of the topics we used was small, our analysis could be made more stable if we had more topics. We will test this point in the future.

5. ACKNOWLEDGMENTS

This research was partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scienti c Research of Young Scientists (B), No.17700604 and No.17700130.

6. REFERENCES

[1] W. Hersh et al. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing & Management*, 37(3):383–402, 2001.

[2] K. Oyama et al. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proceedings of NTCIR-5 Workshop Meeting*, pages 423–442, 2005.

[3] M. Takaku et al. An application of the NTCIR-WEB raw-data archive dataset for user experiments. In *Proceedings of EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)*, pages 78–81, Tokyo, 2007.

[4] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of SIGIR 2001*, pages 225–231, 2001.

[5] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of SIGIR 2006*, pages 11–18, 2006.

[6] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval e ectiveness. In *Proceedings of SIGIR 1998*, pages 315–323, 1998.

Web Page Relevance: What are we measuring? [Position Paper]

Diane Kelly
University of North Carolina
Chapel Hill, NC 27599-3360 USA
+1 919.962.8065
dianek@email.unc.edu

ABSTRACT

Relevance is central to all studies of Web information seeking, but little research has been devoted to developing and evaluating techniques for eliciting relevance judgments from users. Numerous studies have demonstrated that relevance is multidimensional and that users employ multiple criteria when making relevance judgments, but most Web information seeking studies still only elicit one-dimensional relevance judgments from users. The purpose of this *position paper* is to focus attention on the lack of adequate relevance measurement practices and to stimulate discussion about how we can do a better job of this in the future.

Categories and Subject Descriptors

H.1.2 [**Information Storage and Retrieval**]: Models and Principles – User/Machine Systems – Human factors

General Terms: Performance, Human Factors

Keywords: relevance assessment, Web searching, information seeking behavior, user studies, measurement

1. INTRODUCTION

Relevance has a long history in information retrieval (IR) [3, 11] and is central to all theories of information seeking, including Web information seeking. Researchers have identified various types of relevance across a wide-range of information seeking domains including traditional bibliographic databases; closed, full-text collections; and the Web [2, 4, 15]. Although it is well documented that relevance is multidimensional and dynamic, most studies still assess relevance with a single measure at a single point in time. Moreover, relevance is assessed in the same way regardless of the user's information seeking task. These oversimplifications of relevance lead to questions about what is actually being modeled in Web information seeking studies, especially in studies of explicit and implicit relevance feedback.

This position paper reviews and discusses the notion of relevance in the context of Web information seeking, with a particular focus on its conceptualization and operationalization. The purpose in doing so is to focus attention on the lack of good measurement practices and stimulate discussion about how we can do a better job of this in the future. Conducting information seeking studies with users is not an easy task and creating new instruments for each study is not always practical. Thus, the development of valid, reliable and sharable instruments, especially for measuring a concept as central as relevance, should be a primary concern of researchers.

This position paper is organized as follows. Section 2 describes studies of relevance in information seeking scenarios, including studies of criteria users employ when making relevance judgments. Section 3 discusses some techniques that have been used to measure relevance and provides an example from the author's own research to illustrate measurement problems. Two popular relevance feedback-based IR techniques – explicit and implicit feedback – are discussed in Section 4 to illustrate the limitations created by inadequate and incomplete relevance measurement techniques. Finally, Section 5 discusses why and how information seeking context (and task in particular) makes one-dimensional measures of relevance problematic.

2. DIMENSIONS OF RELEVANCE

Saracevic conceptualized relevance along five dimensions: (1) system or algorithm; (2) topical; (3) pertinence or cognitive; (4) situational; and (5) motivational or affective [11]. *System* or *algorithm relevance* describes the relationship between a query and the collection of information objects. This type of relevance is operationalized by a particular algorithm, and does not involve user judgment. *Topical relevance* is associated with the aboutness of a particular document. For instance, if the user's query is 'elephants,' then a document containing a discussion of elephants is topically relevant. *Pertinence*, or *cognitive relevance*, describes the relationship between a user's perception of his information need, what he currently knows about the information need and a document. This is very much related to psychological relevance [6], which considers the degree of cognitive transformation or learning that is caused by reading a document. *Situational relevance*, originally coined by Wilson [17], is concerned with the idea that relevance judgments change according to task and situation. Finally, *motivational* or *affective relevance* describes the intentions, goals and motivations of the user.

Cool, et al. [4] and Barry [2] identified additional types of relevance at a more specific level of detail. Example dimensions identified by Cool, et al. include: interest in document, quantity of information, specificity, authority, entertainment value and usefulness. Example dimensions identified by Barry include: recency, clarity, depth, and novelty. While these studies elaborated on the types of relevance criteria users employee when judging documents, they were conducted in the context of traditional document retrieval systems with traditional information seeking tasks (i.e., finding resources for a scholarly paper).

Some of the dimensions identified by Cool, et al. and Barry generalize to Web information seeking, but others do not. Tombros, et al. [15] studied the criteria used by searchers when evaluating the relevance of Web pages for three types of

information seeking tasks: background search, decision task, and many items task. Tombros, et al. found that users employed several criteria related to the quality of the page when evaluating relevance. These included depth, authority, recency, novelty and general quality, which were similar to those found by Cool, et al. and Barry. Users also employed criteria related to other aspects of Web pages including text (e.g., content, titles), structure (e.g., layout, links) and physical properties (e.g., file size, language).

Although the research discussed in this section identifies dimensions of relevance – which is helpful for conceptualizing this notion – the research does not suggest how one might operationalize (or measure) relevance. Without such studies it is difficult to know which dimensions are most important, and when and how they should be applied and assessed in the context of Web information seeking.

3. MEASURING RELEVANCE

Ultimately, a good measure is one that is valid and will reliably distinguish and discriminate among levels of a concept. Relevance research has demonstrated that users distinguish between more than just binary relevance, but what does a good relevance measure look like? *Degree of relevance* refers to the rating and indication of the relevance value of a given assessed information object. Borlund [3] provides an overview of different degrees of relevance including binary relevance (e.g., relevant, not relevance), tripartite relevance (e.g., relevance, partially relevant, not relevant), scale-based relevance (e.g., 5-, 7- and 11-point scales), and graded relevance (e.g., A, B, C).

While some research on relevance measurement has been conducted, it has primarily been in the context of traditional document retrieval scenarios. Tang, et al. [13] compared relevance scales with varying points and recommended using seven point scales. However, Spink and Greisdorf [12] identified some flaws with this work and applications of these findings have not always resulted in realistic distributions of relevance scores. Others have proposed categories of relevance and techniques that project relevance onto continuous scales. For instance, Eisenberg [6] evaluated the effectiveness of several techniques for measuring relevance based on magnitude estimation: numerical estimation, line production and force of hand grip. Eisenberg's subjects were successful at using all three techniques, but reported a preference for categorical rating scales. Eisenberg also found that magnitude scales were not biased by order of scaling or order of presentation, and relevance judgments were distributed predictably. Spink and Greisdorf compared two techniques for eliciting relevance assessments from users: a 77-mm line ranging from relevant to not relevant and a categorical measure ranging from not relevant, partially not relevant, partially relevant and relevant. In addition, users were asked to identify the levels of relevance that contributed to their ratings using Saracevic's five relevance types which were assessed with binary measures.

In my own research of users' Web information seeking behavior [9] I modeled Web page relevance after Wilson's situational relevance [17] and Cooper's subjective relevance [5], and provided users with a seven point scale to indicate how useful documents were in addressing the information tasks in which they were associated. Users completed their natural information seeking tasks in this study, which went beyond just basic searching tasks. Thus, the relevance model was much more complex since users were working on a variety of tasks. It was believed initially that a more subjective, general relevance measure was appropriate; however, this did not turn out to be the case. For instance, many documents viewed by users were done so for entertainment purposes and it is difficult to imagine that users would not rate them as useful; however, rating a document viewed for entertainment purposes as very useful is unlikely to carry the same meaning as rating a document viewed for a scholarly research as very useful. In this study, users were very liberal and subjective with their relevance judgments and rated a large portion of the documents that they found as highly relevant, which made data analysis difficult since the distribution of relevance judgments were skewed. The results of this study indicate that studies of Web information seeking with natural search tasks should elicit multiple measures of relevance, customized to specific kinds of tasks and that a seven point scale may not be the best way to indicate degree of relevance. The relationship between information seeking task and relevance is explored in more detail in Section 5.

4. RELEVANCE APPLICATIONS

Relevance feedback techniques have been studied extensively for many years [10]. Relevance feedback can be used to alter single user-system interactions, or it can be used to alter interactions over time (e.g., as part of filtering and personalization techniques). Most relevance feedback techniques make use of explicit and/or implicit feedback [8]. Examples of explicit feedback include users marking terms, passages or documents that are relevant to their information needs. Examples of implicit feedback include monitoring and using users' behaviors and interactions with documents to infer relevance.

Central to both explicit and implicit feedback is the notion of relevance. However, in studies of explicit and implicit feedback relevance is usually simplified and what is meant by relevance is often vague, general and unstated. Relevance is usually assessed using a single measure, such as binary judgments (relevant or not relevant) or a scale with a sequence of numbers ranging from relevant to not relevant. A user may mark a document relevant for a variety of reasons, but a single measure only allows the system to accept one type of relevance information and formalize it in one type of algorithm. One exception is the work of Zhang and Callan [18] which elicited several pieces of information from users of a Web news filtering service and incorporated the feedback in different ways. Zhang and Callan elicited likeness, relevance, novelty, authority, and readability from users; however, based on previous research one might argue that most of these were dimensions of relevance, rather than orthogonal measures. Nevertheless, this is a step in the right direction.

Lack of articulation of what is meant by relevance is particularly problematic in studies of implicit relevance feedback, since users' explicit relevance judgments are often used to provide a benchmark upon which to evaluate behaviors proposed as implicit feedback. Although the general idea behind the use of implicit feedback is to personalize IR interactions to the individual user, the dynamic and multidimensional nature of relevance has not been acknowledged in benchmark evaluations of implicit feedback. Studies of clickthrough data as implicit feedback [c.f., 1] present an interesting case because a seemingly simple behavior suddenly represents a very complex notion. It is often common in these types of studies for researchers to use terms such as "interest" or "preference" to describe what they are

modeling instead of "relevance," even though it is unclear if these terms really all represent different concepts. Ultimately, what is actually being modeled in these studies is still unclear because little is known about the context of users' information seeking.

5. RELEVANCE & CONTEXT

Previous research has demonstrated that relevance assessments can vary according to user- and/or context-specific attributes of the information seeking situation [c.f., 9, 14, 16]. Information seeking task is an important aspect of context that is likely to impact a user's relevance behavior. For instance, particular dimensions of relevance are likely to vary according to task; when users check the news they are likely to employ different kinds of relevance criteria than when they search for research articles to use in a scholarly paper or when they search for entertainment purposes.

Consider the following set of items that can be found on the Web: a weather report, a recipe for peanut butter pie, a YouTube video of a man catching sunglasses with his face, a description of a sweater from a retail store and a SIGIR article about ranking. What do all of these items have in common? Probably very little except that they may have all been clicked on by the same user. According to clickthrough analysis, they would all be considered equally relevant to the user. If the user were provided with a generic 'usefulness' measure and asked to explicitly assess the relevance of each of these items in relation to her information needs (e.g., checking the news, cooking, entertainment, shopping, and working on a research paper), the items might still all be marked as equally useful by the user. However, it is unlikely that they are all equally *important* or *valuable* to the user. Each item is likely to contribute different amounts to what the user knows and these contributions are likely to have varying significances. Furthermore, the implications of not finding any of these items are also likely to differ. If one is writing a review article for the IR community about ranking and misses a SIGIR paper about this topic, then this is likely to have greater consequence than missing the YouTube video for the entertainment task or the peanut butter pie recipe for the cooking task.

The relationship between relevance and value was first articulated by Cooper [5] who argued that systems should be evaluated based on subjective user satisfaction, rather than precision and recall. Cooper called a user's satisfaction with a search, *search utility*. To compute search utility, a user was first asked to indicate for every document viewed during a search how many dollars his contact with the document was worth to him. Values could be positive or negative. Once utilities were assessed for all documents viewed by the user, they were summed and divided by the total number of documents the user viewed to arrive at the search utility. Assessing the value of information objects might be one way to collect relevance assessments that are comparative across task.

The example also demonstrates the importance of developing task-centric relevance measures, especially for use in naturalistic studies. For instance, consider criteria identified by Tombros, et al.: depth, authority, recency, and novelty. The importance of these criteria is likely to change according to task. For an entertainment task, novelty is likely to be very important and depth, authority and recency are likely to be less important. For checking the weather, authority and recency are likely to be more important than depth and novelty. For other tasks, such as shopping, it is more difficult to apply these criteria. There may also be differences introduced by the types of tasks users are asked to complete – real or assigned. For instance, the importance and complexity of cognitive, situational and affective relevance is likely to differ according to whether users are completing real or assigned search tasks. It may not be possible for users to evaluate certain kinds of relevance for certain kinds of tasks and it may not even make sense to ask them to do so.

6. CONCLUSIONS

Our research does not provide us with too many choices when it comes to measuring relevance. Most researchers acknowledge the difficulty of defining and measuring relevance, but the problem of measurement still persists. Measurement difficulties have not stopped research – in fact, quite a lot has been done lately with explicit and implicit relevance feedback, but it is unclear what such studies are actually modeling. Although a number of studies have identified different dimensions of relevance or criteria that users employ when making relevance judgments, most studies have stopped short of developing and evaluating specific procedures for measuring these dimensions or criteria. While some studies have evaluated specific measurement techniques in the context of traditional document retrieval systems, it is unclear if and how these measures generalize to Web information seeking.

The goal of this position paper was to review and discuss the notion of relevance in the context of Web information seeking, with a particular focus on its conceptualization, operationalization and application. The purposes in doing so were to highlight some of the problems with using one-dimensional measures to assess relevance; to provide a focal point for discussion at this workshop on the inadequacies and consequences of poor measurement; and to stimulate discussion about what we can do to address these problems. Relevance is central to all studies of Web information seeking. It is usually necessary to make some simplifying assumptions when we design a study – otherwise we would never be able to do any research – but it is now time for relevance measurement to become one of our chief concerns.

7. REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning interaction models for predicting Web search result preference. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, Seattle, WA, 3-10.

[2] Barry, C. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*, 149-159.

[3] Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science, 54*(10), 913-925.

[4] Cool, C., Belkin, N. J., Frieder, O., & Kantor, P. (1993). Characteristics of texts affecting relevance judgments. In M. E. Williams (Ed.), *Proceedings of the Fourteenth National Online Meeting*, 77-83.

[5] Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of

evaluation. *Journal of the American Society for Information Science, 24*, 87-100.

[6] Eisenberg, M. (1988). Measuring relevance judgments. *Information Processing & Management, 24*(4), 373-389.

[7] Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*, 602-615.

[8] Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole (Eds.) *New Directions in Cognitive Information Retrieval*. Springer Publishing: Netherlands (pp.169-186).

[9] Kelly, D. (2006). Measuring online information-seeking context, part 2. Findings and discussion. *Journal of the American Society for Information Science & Technology, 57*(14), 1862-1874.

[10] Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, 18*(2), 95-145.

[11] Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pors (Eds.), *Integration in Perspective. Proceedings of the Second International Conference on Conception of Library & Information Science (CoLIS 2)*, Denmark, 201-218.

[12] Spink, A. & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Science & Technology, 52*(2), 161-173.

[13] Tang, R., Shaw, M., & Vevea, J. L. (1999). Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science, 50*(3), 254-264.

[14] Taylor, A. R., Cool, C., Belkin, N. J., & Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management, 43*, 1071-1084.

[15] Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess Web pages for information seeking. *Journal of the American Society for Information Science & Technology, 56*(4), 327-344.

[16] Wen, L., Ruthven, I., & Borlund, P. (2006). The effects on topic familiarity on online search behaviour and use of relevance criteria. *Proceedings of the 28th European Conference in Information Retrieval (ECIR 2006)*, London, UK.

[17] Wilson, P. (1973). Situational relevance. *Information Storage & Retrieval, 9*, 457-469.

[18] Zhang, Y., & Callan, J. (2005). Combining multiple forms of evidence while filtering. *Proceedings of Human Language Technology Conference & Conference on Empirical Methods in Natural Language Processing (HLT 2005)*, Vancouver, Canada, 587-595.

Position paper: User interactions with results summaries

Frances Johnson
The Institute for Information Research
Manchester Metropolitan University
Geoffrey Manton, Manchester M15
6LL
(+44) 161 247 6156
F.Johnson@mmu.ac.uk

ABSTRACT
This position paper on web information seeking and interaction draws on information seeking models to broadly describe the searcher's interactions and the functionality of the retrieved results page as supporting a process of concept forming. Viewing search as developing an information need enhances the supporting function of the presentation of the search results, beyond the more traditional function of relevance spotting. User studies to investigate the effectiveness of novel interfaces supporting search are essential, but there is a need for basic research into the nature of search and its relation, specifically, with results presentation.

Keywords Search Interface, Summary Presentations, Information Seeking

1. INTRODUCTION
This position paper relates to the workshop's theme and to the author's research on modeling users' conception of search and the design of system components to support users' interaction during information seeking. Various student projects (carried out in the Department and its associated Research Institute of Information Research during 2006-07) have provided indication that the web environment and search engines are giving rise to new styles of interaction and information seeking, especially among the student population. New styles are reported in published research projects such as Nicholas et al [1] whose logs analysis showed a dynamic form of information seeking behaviour (isb) with information gathered horizontally moving from site to site. The authors termed this as bouncing or flicking. The reported use and students' preference for search engines [2], especially Google, when seeking course related information has, however, prompted some alarm

among academics. The concern is that students may not be required to employ critical thinking skills in finding information, resulting in the retrieval of superficial information and/or information that does not connect to anything else [3]. Google's popularity is unsurprising given its widespread use for personal queries, for example of a transactional nature, and its familiar 'minimal' interface of search box and ranked results offering easy access. The concern surrounding students' usage lies primarily with the need to judge the quality of the information retrieved but also, it would seem, with the possibility that its ease of use influences the student's perception of and approach to search. Yet some of our earlier research exploring students' mental models of search engines revealed that a fairly sophisticated model of search - as a process - was held by the participants [4]. Further, more recent interviews with students (albeit a small number) indicated their use of Google as only one of several tools and, its strategic use at the early stages of search to learn about the topic and/or to increase confidence in ability to search on other databases. It is in this context that it seems important that further research aims to better understand the users' conception of search and the possible impact of the systems' conceptual interface in supporting search processes.

2. SEARCH AND THE INTERFACE
The challenge for the design of novel interfaces to support users' interaction during information seeking is posed by the fact that search rarely is a single interaction, but a process, and is exacerbated by the diversity of the user population and tasks. In modern retrieval environments it is likely that the search tool is used at any point in this process, possibly for which the system was not purposefully designed. It is possible to derive this from a brief overview of some of the key models in isb [5-8] which in common describe (pre web) information seeking as a process involving sub processes of: the recognition of a information need, its definition, selection of a source, formulation of the query, examine results, reflect/iterate or stop. Interfaces are, in the main, designed for the input of the query and the output of the results on which the user identifies item(s) sought and/or makes some relevance judgment, possibly to modify the query with the intention to retrieve better results from the collection. Within this model, empirical data on users' information seeking behaviour (query formulation and relevance assessments) has informed the design of the supporting search interface. For example, back in 1997 significant use was made of different windows in the DLITE interface [9] recognising the need to provide different functionality and to make distinct the user tasks of controlling the search process and reading detailed bibliographic information about the retrieved documents. Further interfaces designed to support sub processes in the interaction have focused on the visual presentation of the retrieved results. The Nirve interface

[10], for example, displays in a 3D format query term frequency and co-occurence and Tilebars [11] further displays query term distribution in the retrieved results. Thus the display of the retrieved items could be seen to have the aim of directly assisting the user in the use of results pages, in both *retrieving*: the user is identifying appropriate or interesting items, and *relevance spotting*: the user is seeking to determine the relation between the query and the retrieved item.

2.1 Concept forming on the results page

Search in the web environment and on search engines, as has been touched on, would relate more to Bates' berry picking model [12] or Kuhlthau's seven stages [13] which characterise search as a fluid and dynamic process in which the searcher may start in a very uncertain state, with limited knowledge and is expected to learn about the topic and the query itself as the search progresses. Kuhlthau's model delineates a stage of exploration where the user is seeking information in a stage of uncertainty as the information need is not yet identified. Similarities can be drawn to the search plan stage which forms part of the traditional search intermediaries' training. Although this takes place at a later stage where the need has been identified it involves the identification and conceptualization of the query. This is taught as a process of concept forming or a concept dialogue requiring the searcher to identify the concepts of the query, the terms and the aspects to use in the subsequent implementation and manipulation of the online search. The web and search engine environment appears to be used for this purpose, and its interactivity possibly facilitates searchers in concept forming and query identification. The extent to which this represents a new style of users' interaction remains to be explored. Nevertheless support for such interaction at the interface does present a greater significance to the functionality of the results page as a tool used to identify and formulate a query. Search interface design, with regards to the presentation of search results and for certain types of queries, may also target the function of *concept spotting*: the summaries provide, in the mind of the searcher, a relation between the assimilated results and the information need expressed in the query.

Search interfaces generally provide little support for the dynamic 'middle' interactive stage of search in which the user is engaged in *relevance spotting* and *concept spotting*. Furthermore until recently few studies have evaluated the effectiveness of different search results presentations. White et al [14] found query biased summaries were more effective than general summaries in assisting users gauge document relevance. Tombros & Sanderson [15] had similar findings and attributed this to fact that they indicated the context within which potentially ambiguous query terms were used. Clustering of search results also goes some way to prompting the user to think about the impact of their query and to disambiguate or refine it in selecting a folder of grouped results. The presentation of clustered retrieved results or its variation in the form of diversification (effectively displaying results from each possible cluster) in the ranked page of retrieved results appears to close the gap between the computer and user model of search working as the human brain on the lines of "like" and "different from", without always achieving consensus. Whether these developments are intended or actually achieve an information seeking dialogue in the mind of the searcher, early indication is that they have a positive effect in supporting the user's evolving query. Further evaluation of the effectiveness of these interfaces is called for. Joho and Jose [16-18], for example, compare the effectiveness of an interface to present faceted groupings from the surrogate record of a selected item as an alternative to clustering. They also investigated the effect of additional representations in the search results presentation, such as top ranking sentences and thumbnail images, and a browsing interface in which each of the three top ranked sentences (trs) for a document could in turn be supplemented by new trs from the top 30 urls. These interfaces offer the users different functionality and the users were reported to have welcomed this. The researchers also indicated a positive effect on the users' query reformulation and search results browsing (relevance judging and viewing of retrieved documents) but called for careful consideration in the selection of additional representations.

This brief review serves to demonstrate the theme of the workshop that styles of user interaction during information seeking behaviour has an important (possibly reciprocal) impact of the design of the features and functionality of the search interface. As a position paper it serves to highlight the need for further investigation of the users' conceptions of the search activity during interaction. The search environment, the user perspective and search models indicate the emerging requirement for search engines to support the user in some form of a concept dialogue. Our further research aims towards this end in the investigation of the functionality of the results page and the summary presentations during search.

3. REFERENCES

[1] Nicholas, D., Huntington, P., Williams, P. and Dobrowolski, T. Re-appraising information seeking behaviour in a digital environment: Bouncers, checkers, returnees and the like. *Journal of Documentation.,* 60, 1(2004), 24-43.

[2] Fast, K.V., and Campbell, G.D. "I still like Google": university student perceptions of searching OPACs and the web. *Proceedings of the American Society for Information Science and Technology*., 41, (1), 2004. 138-146.

[3] Brabrazen, T. BA(Google): Graduating to information literacy. *Keynote paper at the IDATER on-line conference.* (Loughborough University, August 2004)

[4] Crudge, S. E. and Johnson, F. C. Using the repertory grid and laddering technique to determine the user's evaluative model of search engines. *Journal of Documentation.,* 63, 2(March. 2007).

[5] Wilson T.D. (1999). Models of information behavior research. *Journal of Documentation*, 55, 2(1999), 249–262.

[6] Ellis, D. A behavioural model for information retrieval systems design. *Journal of Information Science*., 15, (1989), 237-347.

[7] Ellis, D., Cox, D., & Hall, K. A comparison of the information seeking patterns of researchers in the physical and social sciences., *Journal of Documentation*.,49, 4(1993), 356-369.

[8] Kuhlthau, C.C., Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42, 5(1991), 361-371.

[9] Cousins, S. B.. Paepcke,, Winograd, T., Bier, E. A. and Pier K. The digital library integrated task environment (DLITE). In *Proc. of the 2nd ACM International Conference on Digital Libraries*, (Philadelphia, PA, USA, July 1997) 1997, 142-151

[10] Sebrechts, M.M., Vasilakis, J., Miller, M.S., Cugini, J.V., & Laskowski, S. (1999). Visualization of Search Results: A Comparative Evaluation of Text, 2D and 3D Interfaces. In Hearst, M.A. et al. (Eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999,* 3-10

[11] Marti A. Hearst. TileBars: Visualization of term distribution information in full text information access. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Denver, CO, May 1995) 1995, 59-66,

[12] Bates. M.J., The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 13, 5(1989), 407-431,

[13] Kahlthau, C.C.: *Seeking Meaning: A process approach to library and information services*. Norwood, N.J: Ablex 1993.

[14] White, R.W., Jose, J.M., Ruthven.I.:A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*., 39, 5(2003), 707-733.

[15] Tombros, A., Sanderson. M. Advantages of query biased summaries in information retrieval. In*: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Melbourne, Australia, 1998) ACM press, 1998.

[16] Joho, H. and Jose, J.M Effectiveness of additional representations in the search result presentation on the web. In: *Abstract Booklet of the 1st International workshop on adaptive information retrieval* (University of Glasgow, 14th Oct, 2006), 2006, 30-32.

[17] Joho, H. and Jose, J.M. Using local context to slice and dice the search results on the web. In: *Abstract Booklet of the 1st International workshop on adaptive information retrieval* (University of Glasgow, 14th Oct, 2006), 2006, 32-34.

[18] Birbeck, R.D., Joho, H. and Jose, J.M. A sentence based ostensive browsing and searching on the web. In: *Abstract Booklet of the 1st International workshop on adaptive information retrieval* (University of Glasgow, 14th Oct, 2006), 2006, 34-36.

Position Paper: Towards Evaluating the User Experience of Interactive Information Access Systems

Leif Azzopardi
Department of Computing Science
University of Glasgow
leif@dcs.gla.ac.uk

ABSTRACT

The sequence of documents a user accesses while using an Information Access System will heavily influence the user's experience of the system. In this position paper, I discuss the idea of using these sequences to evaluate one facet of the overall user experience. Such sequences can be considered a generalization of rankings used to evaluate Information Retrieval System, but can be generated by any Information Access System. While this view provides greater flexibility for evaluating and comparing different Information Access Systems, it also raises many other issues regarding their application for evaluation. This paper provides an overview of some of these issues along with possible directions for using sequences of documents to evaluate Information Access Systems.

1. INTRODUCTION

Evaluating interactive Information Access Systems (IAS)[1] is a major challenge within Information Retrieval (IR). Largely, IR has been concerned with the effectiveness of a system defined by traditional measures such as mean Average Precision [8]. In order to obtain such measurements, a highly controlled batch retrieval experiment is conducted, where the performance of the system is measured by evaluating the ranked list of documents returned in response to each query. In this experiment, it is assumed that given the ranked list of documents for a particular query the user will perform the following; starting at the first document in the ranked list, assess it for relevance, then continue to next the document, assess it for relevance, and so on, until some cut off around one thousand documents. This ranked list of documents provides the basis for measuring the performance of an IR system.

[1]I use the phrase Information Access, as opposed to Information Retrieval, because I want to consider broader interactions than just retrieval. For instance, interactions based on searching and browsing [2], clustering [5], orienteering [7], etc.

SIGIR'07 Web Information-Seeking and Interaction (WISI) Workshop,
July 27, 2007, Amsterdam, The Netherlands.
ACM 0-89791-88-6/97/05.

In an interactive retrieval scenario, the way in which a user interacts with a system is significantly different to the assumed interactions in batch retrieval. While a ranked list of documents may be presented to the user in response to their query, the user will determine the order in which the documents are retrieved. They may choose not to view certain documents, they may curtail their search at any time, they may reformulate their queries, etc. For example, in web search, a searcher may retrieve the first document, then the fifth document, where they then browse to another document, before returning to the results pages to select another document. This interaction with the system builds up a sequence of documents that are accessed as a result of interaction with the system. Such interaction generally leads to: (1) only a few documents being accessed (instead of hundreds) per query, (2) the order in which documents are retrieved may not be in a linear fashion given a ranking, (3) documents accessed by the user are not bounded or constrained to only those results in the ranked list presented because links facilitate navigation, (4) multiple queries may be submitted for a particular topic, and (5) multiple topics may be expressed within the same session. Consequently, the assumptions under the batch experiment are not appropriate, because the interaction means the user dictates what is assessed or not, according to which documents they retrieve.

An attempt to bridge this gap between interactive and batch experiments was proposed by Leuski [5] who was motivated by Bookstein's [1] view of the information retrieval process. Bookstein argued that the retrieval process is one where the user examines a sequence of retrieved documents, and where the system provides feedback by adjusting the documents presented to the user. Then depending on the feedback and adjustments, the system presents different documents and the user selections create a particular sequence of retrieved/accessed documents. Leuski [5] took this view on board when evaluating cluster based retrieval systems. First he defined different "search strategies" to represent different user behavior, and then he simulated the subsequent interaction with the system. For each query the resulting interaction produced a sequence of documents accessed, which was considered to be the ranked list given the query. This meant that cluster based systems could be compared to standard retrieval systems using the same measures, because both output a ranking from the course of interaction.

While this provides a novel solution for comparison purposes, in this paper I consider Bookstein's view in a broader context. That is if the IR process is a sequence of docu-

ments retrieved, then we should perform evaluation on this sequence. I argue that since the sequence of documents retrieved through the course of interaction is common to all Information Access Systems (as opposed to rankings which are more IR system based), then these sequences form the basis for measuring the performance of any Information Access System. This is because the primary goal of an IAS is to provide access to relevant and useful documents, so it is then reasonable to assume that the sequence of documents retrieved will play a major role in determining the user experience of the IAS. The reason being, that the sequence of interactions with the system determines the user's perception of the system's performance which defines their user experience [6]. The sequence of documents retrieved by the IR process is a subsequence of all interactions, but arguably captures the a predominate facet of the user's experience with respect to the goal of an IAS.

Using sequences provides many different evaluation possibilities, but also raises many issues regarding their usage for the evaluation of IASs. The remainder of this paper is as follows: the next section introduces sequences as a way to capture the User Experience, before discussing their application in evaluating IASs and the issues that arise when considering their application in Section 3.

2. USER EXPERIENCE AS A SEQUENCE

User Experience (UX) is a term used to describe the overall experience and satisfaction a user has when using a system/product and includes "all aspects of the user's interaction with the product: how it is perceived, learned, and used" [6]. The field of user experience design is a highly multi-disciplinary field, incorporating facets from psychology, computer science, graphic design and industrial design. User experience design pertains to the creation of the architecture and interaction models which impact upon a user's perception of a digital device or system. Consequently, there has been a move towards developing interactive systems with the goal of designing for the "user experience". And so, the aim of any such system is to maximize the user experience of the system, and for very good reasons. A negative user experience can be expensive, resulting in lost revenue, diminished customer loyalty, loss of word of mouth advertising and even a damaged brand identity. This is especially important for web search providers where competition is fierce. A poor UX may result in the searcher switching web search providers. This provides a strong motivation for maximizing the UX of searcher because of the potential loss in advertising revenue.

The UX of an IAS is affected by other many factors aside from the sequences of documents accessed. For instance, the information cues presented to the user by the system (types of summaries, snippets, recommendations, sponsored links, etc), the presentation paradigm (cluster based, ranked list, graphs, etc), user interface (color, font, layout, etc), mode of interaction (voice, mouse, keyboard, etc), and so forth will all contribute to the overall satisfaction and experience of the user of the IAS. For example, if the IAS was very difficult to use for some reason making it virtually impossible to access information then this would seriously degrade the user experience. The focus of this paper, as previously mentioned, is on only one particular facet of the UX because, I argue, it is the main factor in defining this experience and is a common factor shared by all Information Access Systems. Other possible factors are assumed to be either system dependent (such as presentation paradigm), or affect the UX independently (such as ease of use).

So given the following high level user interactions with a web search engine (as the IAS), where the initial state is from the IAS home page and user performs the following actions: submits a query q_1, examines the results page, selects a result, examines the document d_1, returns to the results page, selects another result, examines the document d_2, selects a link (from d_2), examines the document d_3, returns to IAS home page, then submits a query reformulation q_2, examines the results page, select a result, examines the d_4, and so on. The subsequence of interest is the sequence of documents accessed over the course of interaction with the system (i.e. $d_1, d_2, d_3, d_4, ...$) and the conditions under which this access arose.

Table 1 shows an example sequence, where the search history, is decomposed into a number of search sessions. If we think in terms of a web search engine, then, such sessions maybe be broken down into a number of topic searches, where each topic search may consist of several interactions such as querying, finding similar documents, relevance feedback, and so forth, or even include browsing. In this example, Topic k_1, has two queries issued, before the user moves on to the next topic, and when searching for information on Topic k_3, the user browses from d_{12} to d_{13} and then onto d_{14}. Queries and topics can also be inter weaved and split over different search sessions. Assuming there is a judgment on the utility/relevance of a document associated with each document accessed in sequence, then the performance of this user experience can be evaluated.

Having information about the context of each access in the sequence means that evaluation could be performed by decomposing the sequence into a number of smaller context specific sequences. For instance, query, topic or session boundaries could be used to determine the average performance given these contexts. In the case of query boundaries, for instance, traditional IR measures such as precision could be applied. Each sequence given a query would be used as a ranked list (as in [5], see subsection 3.1).

While, the sequences can be used to perform traditional ranking based evaluation, sequences offer many different evaluation alternatives. This is because the focus of evaluation is user oriented (i.e. based on the documents actually retrieved), as opposed to system oriented (i.e. based on the documents returned and assumed retrieved). One alterative is to consider how the user experience changes over the sequence. For instance, by placing a window over the sequence to define artificial boundaries, then the user's experience over the course of interaction could be evaluated. The main advantage is that regardless of the type of IAS, for instance a browsing based interface [5, 2] where there is no query, different types of systems can be evaluated and compared. Another dimension which sequences bring to the evaluation is the transition between document accessed (i.e. moving from retrieving d_1 to retrieving d_2). Considering the transitions means that other factors such as the effort required to accessed each document can be included in the evaluation process. For instance the effort involved in using a particular IAS to facilitate the access due to reading the results page, traversing the clustering, etc. So using sequences for evaluation purposes introduces at least two different ways in which to consider measuring the UX of an IAS; in which traditional IR measures could be applied

Information Access Path $\Longrightarrow$																		
User Experience Contexts	Search History s_1																	
	Session s_1										Session s_2							...
	Topic k_1						Topic k_2				Topic k_3					Topic k_4		...
	q_1				q_1'		q_2				q_3			d_{12}	d_{13}	q_4		...
Documents	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9		d_{10}	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}		...
Judgments	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9		j_{10}	j_{11}	j_{12}	j_{13}	j_{14}	j_{15}		...

Table 1: Example Information Access Path

or extended, or new measures which are more suitable or appropriate for the scenario could be developed.

3. DISCUSSION

The introduction of sequences provides a different perspective on the evaluation of IASs. However, there are many issues that need to be considered, such as how the sequences relate to batch retrieval (rankings) and other IAS, the difference between relevance and utility, the dependance/independence of documents/judgments, what documents to include in the sequence, what is a document, and what measures to use. The remainder of this paper is a preliminary discussion on some of these issues.

3.1 Sequences in Traditional IRS Evaluation

Before progressing any further, let's consider a sequence in the context of traditional IRS Evaluation [8]. As previously mentioned, in the standard batch retrieval experiment a fixed mode of interaction is assumed; once a query is executed and the ranked list of documents is presented it is assumed that the user starting at the first document in the list, assesses each document in turn, one after another. The sequence defined, is determined by the ordering of the document in the ranked list (i.e. $d_1, d_2, d_3, \ldots, d_n$). The ranking is therefore a specialization of a sequence, to which a whole host of evaluation measures that can be applied to determine the performance of the system in response to the particular query, such as Precision, Recall, Average Precision, Discounted Cumulative Gain, bpref, etc [8].

3.2 Mapping different styles of Interaction

While the case is trivial for the batch experiment we can easily define a sequence for any IAS. As we previously mentioned Leuski [5] suggested how cluster based retrieval can be mapped into a ranked list (which can be considered as the sequence for a particular query/topic using such an IAS). Different rankings were created by simulating different "search strategies", applicable to a cluster based presentation paradigm. Different search strategies[2] resulted in different sequences of documents, which were used as ranked lists in order to compare different retrieval systems. The same idea applies for other interactive scenarios. Below are a few examples, but mapping different scenarios raises various issues.

Relevance feedback in Batch retrieval. When there are iterations of relevance feedback, the sequence built up is composed of all the accessed documents. Documents accessed in successive iterations of the relevance feedback are appended to the sequence. Assuming that the sequence includes a unique set of documents retrieved then the outcome of doing so will produce a ranked list. In fact, this ranking would be equivalent to using the "residual collections" technique proposed in [9]. Applying the Residual collections technique excludes the documents that have already be assessed by the user from subsequent rankings and is one solution to address the problems of evaluating relevance feedback (i.e. "total performance" and "ranking effect"). While this restriction on the sequence is suitable for relevance feedback evaluation, other types of interaction will result in the same document accessed multiple times during a sequence. An open question is how to evaluate a such a sequences?

Browsing based retrieval. Browsing based systems like the ostensive browser [2] also produce a sequence. The documents assessed during the browsing form the path of interaction in the document space which lead to a sequence of documents being accessed. If the query is explicated to the system then the sequences could be used to form rankings determined by the browsing (as in [5]). If not, the sequences still enable the evaluation of the IAS, despite the query being unknown. For instances windows over the sequence could be used to measure the performance.

Since browsing based strategies assume that the information need is more dynamic and changeable in nature, then the judgments a user makes will be based upon the order in which the documents are assessed and the user's current information need. This type of interactive system especially motivates considering how the (in)dependance between documents and (in)dependence between judgments should be treated. Is it sensible to assume independence as done in tradition IR evaluation and what does this mean for the evaluation of an IAS?

Search and browse. The combination of searching and browsing resulting from interactions such as relevance feedback, query reformulation, find similar, etc can all be mapped to a sequence. In a web setting, as already described, the documents a user accesses from those presented, are added to the sequence. Since, web pages link to other documents, then should documents subsequently accessed also be added to the sequence? One could argue that these documents also affect the user's experience of the IAS, because as a result of using the IAS these documents are indirectly accessed. For instance, in a recent studies it was found that two predominate web search strategies were target search and orienteering [7]. In orienteering, searchers queried the search engine to acquire an entry page from which they could navigate to the desired information. Here, the web search engine and the provider of the entry site will both affect the user's information access experience. So including documents accessed

[2]A search strategy defines/describes the interactions of the (simulated) user with the system, for instance, a user might select the best-first, breadth-first, depth-first, randomly, etc.

indirectly from a web search engine may be inappropriate. However, if it is the user's total information access experience that is under study then all documents accessed would be appropriate to include in the sequence.

While documents that lead to relevant documents are not totally relevant, they are certainly useful, and would presumably enhance the user's experience. But, how should such useful pages be considered when evaluating the user's experience of the IAS? If a document is useful because it links to a relevant document, then how should a document which contains a set of search results, be considered? If as just another document which is accessed during the course of interaction then the evaluations measures need to cater for the different utility of the different types of documents, and how the utility changes through the course of subsequent interactions with the document.

3.3 Issues from Interaction

A major challenge when evaluating an IAS is dealing with the change of state in the user as they interact with the system and access different documents. Through the interaction the user learns about the information space which may affect their perception and judgment of subsequent documents retrieved. This "learning effect" creates many problems for the evaluation process. As a result the judgments made are unlikely to be independent meaning that re-using the judgements may not be possible because the judgments are dependent on the order in which the documents are retrieved. For instance, a non-relevant document may provide cues that help to reformulate the query, but seeing several other similar non-relevant documents (while providing confirmation) does not provide any more utility. On the other hand, a partially relevant document seen after a highly relevant document may be less useful, than if the order was reversed. So while it may be relevant it may not be as useful. Accounting for such effects given the change in ordering is an open problem and limits the re-usability of the judgments. So unlike batch experiments where judgments are typically assumed independent of each other [8], in the interactive setting this is not the case, and limit the judgments only to the context in which they appear.

As opposed to the independence of judgements, the dependence between documents is an important factor for systems to consider when ranking and presenting results to users [10]. This is especially the case in web search where queries are often highly ambiguous. For instance, in [3], they advocate that a system should return documents that are as presumed to be highly relevant, but which are also different to the documents already returned. This is so the chance of returning at least one document which is relevant to the user's information need is increased. Under traditional IR evaluation, this kind of system would be penalized for returning many documents which are not relevant to the specific information need. However, by evaluating the system using a sequence the system would not be penalized for presenting a list of novel, but irrelevant options. While on one hand this is appropriate given the interaction, it also raises other issues concerning how to consider the list of results presented and the effort required to select the novel but relevant document from the list of results. Specifically, is the list of results that a system presents a document itself (a pseudo document)? In which case how relevant and/or useful is this document and should it be considered within the sequence?

Another issue resulting from the interaction, is the possibility of building sequences that contain non-unique document sequence i.e. documents are revisited (as could be the case when browsing). In the batch retrieval experiment it assumes that only unique documents appear on the path. While this assumption that could be applied in an interactive setting, it is reasonable to assume that during interaction a document may be accessed multiple times. Each time the document is accessed a new judgment about the relevance/utility of the document would be generated. For example, if the search results were included in the sequence, then we could imagine that such documents would be revisited multiple times. While this document may not be relevant, it may be very useful and of high utility initially, but not necessarily relevant. On subsequent accesses the utility of the document would presumably decrease. This raises the questions, what affect the utility and relevance of the documents have on the user experience and how does it affect their judgments?

As mentioned above there appears to be a difference between utility and relevance in the context of a sequence. Relevance appears to be a specific relationship between a query, document and user at a particular point in time, where as utility appears to be a more general relationship between a document and a user at a particular point in time. This distinction appears in cases such as the previous example where the document has utility but is not relevant. However, this distinction is not so clear in the case of a document which is relevant and therefore highly useful. If one is happy to accept such a distinction then we could consider two aspects of user experience w.r.t the documents retrieved, effectiveness based on relevance and usefulness based on utility of the sequence in order to evaluate the UX of the IAS.

Besides the problems associated with the independence of judgments, and the utility or relevance of judgments, a further problem is how to capture the utility/relevance of the documents that a user interacts with. While explicit judgments could be obtained, in an operational setting, implicit mechanism for inferring the relevance/utility of the documents would be more appropriate (and cost effective). Possibly, surrogate or implicit indicators [4] such as time spent on page to suggest relevance, or the number of times a document is viewed, could be used to infer the utility of a page. Whether methods to infer the utility of a document in an implicit manner is possible is also an open problem.

Since we have a sequence of documents accessed, as a result of some interaction with the system, then it is sensible to consider a frame of reference for measuring performance. As previously mentioned, this could be defined by the query/topic boundaries, but in highly interactive systems, the number of documents associated to any one sequence for a given query is typically very short which may not be adequate for a robust measurement. On the other hand, the sequence of interaction may not result from direct querying but by browsing (or some other means), where the information need may be changing. Consequently, there will not always be a clear boundary in which to segment the stream of interactions. A solution is to define a window which we can place over the stream of document accesses. This window represents the current experience of the user, and total (or overall) user experience is obtained by the sliding of the window along the stream of interactions.

Now given the window over the sequence of documents

which defines a subsequence from positions k to $k+n$, there are a number of options for measurement. The obvious approach is to apply existing measures to the subsequence. For instance, the precision of the window which defines the proportion of relevant information retrieved given the set of retrieved in the sequence. Because, the window may cross different query/topic boundaries, we are not interested in the performance of that query, but performance of the user's experience of the system providing relevant information as the user interacts with the system. And so the natural extension of the window precision, would be the average window precision defined by sliding the window along the sequence to capture the user experience.

Another set of measures could be devised based on the transitions in the sequences. For instance, the effort required in accessing one document to another document in the sequence would impact upon the user experience. Measures which consider the effort to relevance ratio, or effort required before relevance could be conceived. In an interactive system the effort could be quantified in various ways such as number of clicks or amount of time and provide another dimension to the evaluation of IAS, beyond traditional IR evaluation.

4. SUMMARY

In this paper, I have suggested that evaluating interactive Information Access Systems should be based on the sequence of documents retrieved by the user through the course of interaction with the system. This view of evaluation mirrors Bookstein's view of the information retrieval process and also provides a natural way to capture the user's information access experience. While this view raises many issues, sequences offer many different possibilities for the evaluation of Information Access Systems which go beyond traditional IR evaluation.

Acknowledgements

I would like to thank Jana Urban and Hideo Joho for their helpful feedback and discussion about this work, and also the anonymous reviewers for their constructive comments.

5. REFERENCES

[1] A. Bookstein. Information Retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34:331–342, 1983.

[2] I. Campbell. Interactive evaluation of the ostensive model using a new test-collection of images with multiple relevance assessments. *Journal of Information Retrieval*, 2:89–114, 2000.

[3] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436. ACM Press, 2006.

[4] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Intelligent User Interfaces*, pages 33–40, 2001.

[5] A. Leuski. Evaluating document clustering for interactive information retrieval. In *Proc. of the 10th Int. Conf. on Information and knowledge management*, pages 33–40, 2001.

[6] Donald A. Norman. *The Design of Everyday Things*. Doubleday, New York, 1988.

[7] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422. ACM Press, 2004.

[8] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Massachusetts 02142, 2005.

[9] C. Cirillo Y.K. Chang and J. Razon. Evaluation of Feedback Retrieval Using Modified Freezing, Residual Collection, and Test and Control Groups. In G. Salton, editor, *The SMART RE- TRIEVAL SYSTEM*.

[10] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, 2003.