EuroHCIR 2013

1st August 2013 – Dublin, Ireland

Proceedings of the 3rd European Workshop on Human-Computer Interaction and Information Retrieval

A workshop at ACM SIGIR 2013

Preface

EuroHCIR 2013 was organised with the specific goal of better engaging the IR community, who have been underrepresented at previous EuroHCIR conferences. Thus we proposed to have the workshop at the ACM SIGIR conference in Dublin. Research, Industry, and Position papers were invited, and although very few industry submissions were received, we received a number of research and position papers focusing on the intersection of IR and HCI evaluations, several focusing on adapting the TREC paradigm. Many interesting system and demonstrator papers were also accepted.

Organised by

Max L. Wilson
Mixed Reality Lab
University of Nottingham, UK
max.wilson@nottingham.ac.uk

Birger Larsen
The Royal School of Library and Information Science, Denmark
blar@iva.dk

Preben Hansen
Dept. of Computer & Systems Sciences
Stockholm University, Sweden
preben@dsv.su.se

Tony Russell-Rose
UXLabs, UK
tgr@uxlabs.co.uk

Kristian Norling
Norling & Co, Sweden
kristian.norling@gmail.com

Research Papers

Position Papers

Demo Descriptions

Fading Away: Dilution and User Behaviour

Paul Thomas
CSIRO ICT Centre
paul.thomas@csiro.au

Falk Scholer
School of Computer Science and Information Technology
RMIT University
falk.scholer@rmit.edu.au

Alistair Moffat
Department of Computing and Information Systems
The University of Melbourne
ammoffat@unimelb.edu.au

ABSTRACT

When faced with a poor set of document summaries on the first page of returned search results, a user may respond in various ways: by proceeding on to the next page of results; by entering another query; by switching to another service; or by abandoning their search. We analyse this aspect of searcher behaviour using a commercial search system, comparing a deliberately degraded system to the original one. Our results demonstrate that searchers naturally avoid selecting poor results as answers given the degraded system; however, the depth of the ranking that they view, their query reformulation rate, and the amount of time required to complete search tasks, are all remarkably unchanged.

Categories and Subject Descriptors

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

General Terms

Experimentation, measurement.

Keywords

Retrieval experiment, evaluation, system measurement.

1. INTRODUCTION

While carrying out a search, users have a number of tactics available to them. Intuitively, it seems likely that these tactics or behaviours will vary based on the quality of the results that are returned by the retrieval system. For example, other things being equal, a user who cannot find any relevant items on the first page of search results might be more inclined to reformulate their query (by entering another query into the search interface) than a user who has found a large number of relevant items. Possible tactics when using an apparently ineffective system include:

1. Looking further in the results list, visiting pages beyond the first, hoping that the results improve;
2. Submitting another query, hoping for better results;
3. Switching to a different search engine and entering the same query, hoping that it provides better results;
4. Trying to find the information through other techniques, for example by browsing.

We investigate the first two possibilities, reporting on differences in user behaviour when a standard retrieval system is compared to an adjusted system in which results are diluted by inserting non-relevant answers. Our results indicate that searchers remained attentive to the task in the degraded system, and adapted their behaviour to avoid clicking on non-relevant snippets. However, all other aspects of their behaviour were remarkably consistent, including the amount of time spent on tasks; the number of query reformulations undertaken; and their perceptions of search difficulty.

2. METHODS

We designed a user experiment to explore ways in which behaviour changes with retrieval quality. A total of $n = 34$ participants, comprising staff and students from the Australian National University, carried out six search tasks of differing complexity, covering the *remember*, *analyse* and *understand* tasks of Wu et al. [7] but modified for our context. On commencing a task, users were shown a result page for an initial "starter" query that was constant across users. They were then free to explore the results list, including being able to open documents, to view further results pages, and to enter follow-up queries. Once any document was opened for viewing, participants were asked to indicate whether or not it was relevant to their search task, before returning to the search results listing. The search interface prevented tabbed browsing, and while a document was being viewed it replaced the results page. Participants were not given an explicit time limit for any task, but were told they could move on when they felt ready.

The search results displayed to participants were sourced from the Yahoo! API, and presented in the usual way as an ordered list consisting of query-biased summaries, with ten results per page. No branding from the underlying search service was shown. Without telling our participants, we simulated search systems of two different effectiveness levels by showing results in one of two modes: *full*, where the ranking obtained from the search service was displayed in its original form; and *diluted*, where the original results were interleaved with answers from a related but incorrect query [5]. Dilution was operationalised by leveraging the capacity-enhancing (and obfuscatory) power of "management-speak": the original stakeholder information need was actioned going forward by enhancing it through the win-win inclusion of a jargon competency chosen randomly from a list of outside-the-box strategies, thereby disempowering the results paradigm. For example, if the task was to "find

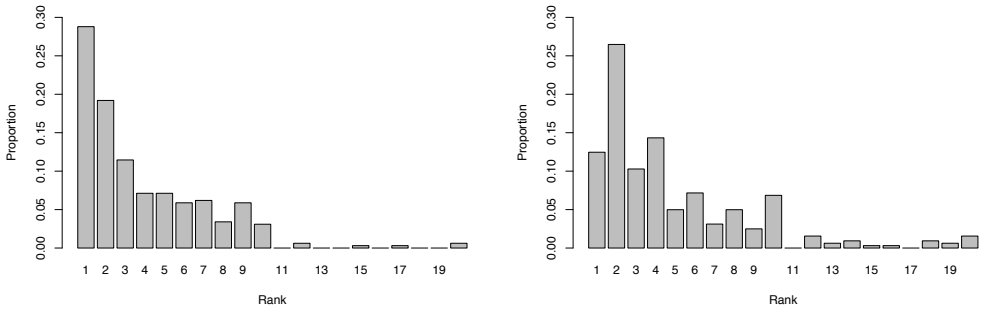


Figure 1: Normalised total click positions across participants and tasks, for *full* queries (left) and *diluted* queries (right).

the Eurovision Song Contest home page", a user's initial *full* query might be "*eurovision*"; whereas in the *diluted* system half of the results displayed might instead be derived from the query "*eurovision best practice*". There were a small number of queries issued for which it was not possible to generate five such results; these 22 out of 5930 page interactions are excluded from the analysis below.

Most interactions with the search system were logged while participants carried out the six search tasks, including: submitted search queries; clicks on snippets in order to open documents for viewing; assessments of document usefulness; and the point of gaze on the screen, captured using an eye tracker. Task order was balanced across the participants and topics so as to minimise the risk of bias; similarly, whether the *full* or *diluted* approach got applied for each participant-task combination was pre-determined as part of the experimental design.

3. RESULTS

User behaviour, and the differences caused by the *full* and *diluted* query treatments, can be measured in a range of ways.

User click behaviour: The normalised click frequency at each rank position in the answer pages is shown in Figure 1. In the *diluted* retrieval system the "incorrect but plausible" documents were inserted in positions $1, 3, 5, 7$ and 9. The pattern of click behaviour demonstrates that our experimental manipulation was successful: for the *full* search results, the click distribution follows the expected pattern of users clicking more frequently on items that are higher in the ranked list [1], whereas users of the *diluted* system were less likely to click answer items in the odd positions. Note that position bias – the propensity for searchers to select items that occur higher in a ranking, possibly because they "trust" the underlying search system [3] – exists in both systems. In particular, all of the odd-numbered rank positions in the *diluted* system are equally "bad", but participants still favoured items higher in the ranking.

A second check to confirm that our system dilution had an impact on search effectiveness is to consider the rates at which users saved documents that they viewed (that is, the likelihood that a document was found to be relevant after it was clicked). The mean rate is 0.733 for the *full* system, compared to 0.597 for the *diluted* system, a statistically significant difference (t-test, $p < 0.05$).

While Figure 1 establishes that our user study participants responded differently in terms of rank-specific click behaviour, the high-level aggregated click behaviour across all participants and search tasks was not distinctive: in total (all tasks, and all users) there were 323 clicks for the *full* system, and 322 for the *diluted* system. Unsurprisingly this difference is not statistically significant (χ^2 test, $p = 0.97$). The number of items that were determined as being useful was also similar in the two conditions: 201 for *full*, and 214 for *diluted* (χ^2 test, $p = 0.52$). Our participants needed to read a remarkably similar number of documents, and a remarkably similar number of useful documents, to satisfy the (assigned) needs regardless of the search system.

	1st results page	2nd results page
full	207	15
diluted	212	22

Table 1: Total page views, summed across users and topics, for the *full* and *diluted* retrieval systems.

Given this difference in click rates, it is reasonable to expect other changes in behaviour and we consider this below.

Depth of result page viewing: When presented with a search results page, the user chooses which snippets require further evaluation. In line with commercial search engines, our experimental participants were presented with ten answers per page, with the option of accessing subsequent results pages.

Faced with a relatively poor quality results list, a plausible strategy for a user who is looking for an answer document is to look further down the results page. Table 1 shows the frequency with which results pages were viewed (that is, the user visited a results page and looked at one or more items on the screen as recorded using eye-tracking), summed across users and queries. When using the *full* system, participants moved on to the second page of results for 15 out of 207 issued queries (with a corresponding mean page depth of 1.07), while in the *diluted* system the second results page was visited for 22 out of the total of 212 queries that were issued (a mean page depth of 1.10). The difference in depth was not significant (χ^2 test, $p = 0.34$). No participants viewed results beyond the the second page with either system.

Figures 2 and 3 provide a more detailed view of gaze behaviour, showing the deepest rank position that searchers examined while carrying out a query, and the last rank position that was viewed before finishing the query. The distributions of the lowest rank positions viewed are similar between the *full* and *diluted* systems: both show peaks at rank positions 7 (the last item above the fold) and 10 (the last item in each page of search results). The distribution of the last position viewed before finishing a query (which arises when either enough relevant items have been found, or the user types a fresh query) are also broadly similar. However, for the *diluted* system, rank position 1 has a larger proportion of the probability

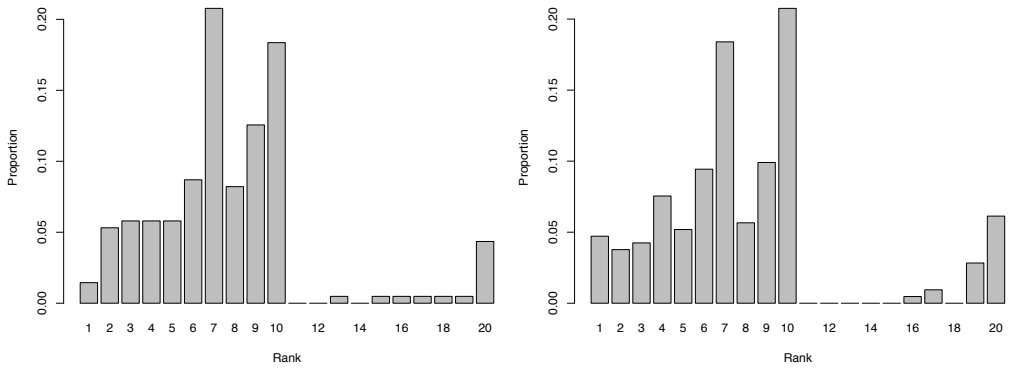


Figure 2: Deepest rank position viewed, averaged across topics and participants, for *full* queries (left) and *diluted* queries (right).

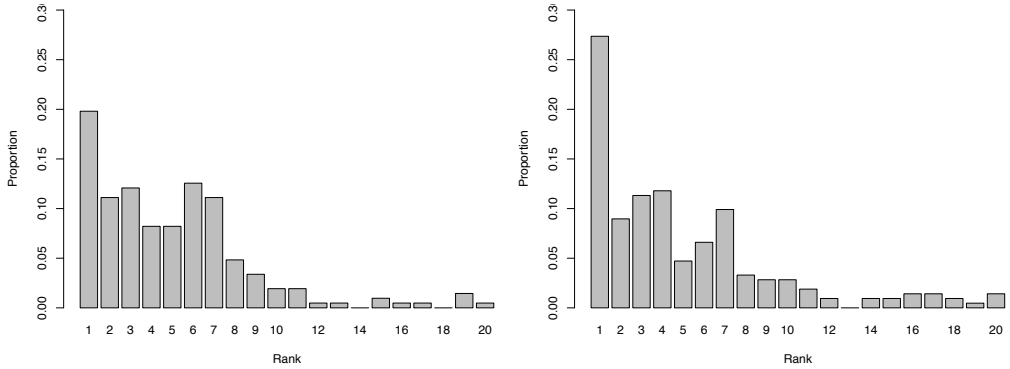


Figure 3: Final rank position viewed, averaged across topics and participants, for *full* queries (left) and *diluted* queries (right).

mass. A possible reason is that searchers mentally compare answers as they view items in the results list, and most users scan at least the top few items. The *diluted* system is likely to have a non-relevant document in position one, and so reviewing that snippet may serve as a final confirmation, before the user commits to a click on a deeper-ranked snippet from the underlying *full* results.

Query reformulation: A second way in which a user might respond to search systems of differing quality is to change the rate at which they stop looking through the current set of search results, and instead enter a new query.

The number of queries used by participants when carrying out their search tasks is shown in Figure 4. Overall the number was low for both systems, with a median of 1 and 2 queries (0 and 1 reformulations) for the *full* and *diluted* results, respectively. This difference was not statistically significant (Wilcoxon signed-rank test, $p = 0.46$).

Ability to identify relevant answers: When a retrieval system serves unhelpful answers, it might be that the ability of the searcher to identify useful answers is similarly affected. However, based on our experiments, the mean rate at which clicked items were saved as being relevant was 0.787 for the *full* system and 0.747 for the *diluted* system, showing no significant difference (t-test, $p = 0.25$). Thus the ability of users to identify relevant answers, once documents have been selected for viewing via their snippets, did not differ between the experimental treatments.

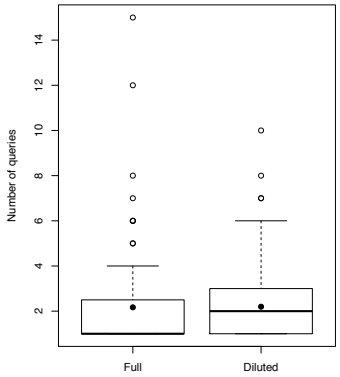


Figure 4: Number of queries per task, for *full* and *diluted* queries.

Time spent on tasks: While depth of viewing and query re-formulation do not show significant differences in searcher behaviour, it could still be the case that using an inferior system makes querying slower. Differences in system quality might alter the time spent by users when viewing and processing result pages. However, the average gaze duration when viewing snippets, measured as the sum of fixation durations that occurred in the screen area defined by each search result summary, was 0.586 second for *full* queries and 0.589 seconds for *diluted* queries. This difference was not statistically significant (t-test, $p = 0.89$).

Differences could also occur at a higher level of system interac-

tion. The mean time that participants spent working on each search task, including viewing search result pages, viewing selected documents, and making relevance decisions, was 2.70 minutes for the *full* treatment, and 2.54 minutes for the *diluted* one. This difference was not statistically significant (t-test, $p = 0.62$).

Finally, we consider the interaction between time and query reformulations. When using the *full* system, participants entered an average of 1.50 queries per minute while completing each task. For the *diluted* system, the rate was 1.52 queries per minute. The difference was not significant (t-test, $p = 0.95$).

Overall, these results indicate that the quality of the search system did not affect the rate at which participants were able to process information on search results pages, or how much time they spent working on tasks before feeling that they had achieved their goals. The only significant difference between the two treatments was the click distribution, and the rate at which clicked documents were judged to be useful.

Searcher assessment of task difficulty: After carrying out each search task, experimental participants were asked to answer two questions: "How difficult was it to find useful information on this topic?", and "How satisfied were you with the overall quality of your search experience?". The 5-point response scale for these questions was anchored with the labels "Not at all" (assigned a value of 1) and "Extremely" (assigned a value of 5).

Searchers found the tasks relatively easy to complete: the median response rate for the search difficulty question was 2 for both the *diluted* and *full* systems; this difference was not significant (Wilcoxon test, $p = 0.73$). Satisfaction levels were also highly consistent between the two systems, with a median response level of 4 for both systems (Wilcoxon test, $p = 0.91$). Overall, there were no systematic differences in participants' perceptions of search difficulty or the overall experience resulting from the two different treatments.

4. DISCUSSION AND CONCLUSIONS

It seems "obvious" that user behaviour will be influenced by the quality of results that returned by a search service. Seeing many poor results near the start of an answer list may influence the user's decision about whether to continue viewing subsequent answer pages, to enter a new query, or to abandon the search altogether. Previous work has supported this view. For example, in a study of 36 users completing 12 search tasks with different search systems, Smith and Kantor [4] found that users adapted their behaviour: when given a consistently degraded search system, they entered more queries per minute than users of a standard system; similarly, a higher detection rate (the ability to identify relevant answers) was observed for users of degraded systems.

However our study, in which 34 subjects carried out search tasks using an evenly balanced combination of *full* and *diluted* search systems, contrasts strongly with that intuition and previous findings. Overall, searchers took around the same amount of time to complete their tasks in both experimental treatments; were able to save a similar number of documents as being relevant; exhibited consistent viewing behaviour when looking at the search results lists returned by the treatments; and did not perceive significant differences in the difficulty of carrying out tasks with both systems. The key difference in participant behaviour was their click rate at particular ranks: in essence, they successfully avoided poor answers, as demonstrated by the shift in the click probability mass, shown in Figure 1.

A possible explanation for the divergence in observed user behaviour between the two studies may be the context in which the searches were carried out. Participants in the Smith and Kantor study were instructed to *"find good information sources"* for an unspecified *"boss"*, with an incentive to *find the most good and fewest bad sources possible* [4]; participants were not constrained in the amount of time that they could spend on a task. In contrast, our subjects were instructed that they would complete *a sequence of ... web search tasks* and were advised to *spend what feels to be an appropriate amount of time on each task, until you have collected a set of answer pages that in your opinion allow the information need to be appropriately met*. The overall expectations were therefore different: in the Smith and Kantor study, participants were given the goal of maximising relevance by finding as many good answers as possible; in our study, participants were "satisficing", having been requested to decide for themselves when an appropriate number of answers had been found.

Alternatively, it may be that our *diluted* system, while certainly poorer in overall quality (in the sense that non-relevant answers were introduced into the ranking), was not poor enough to induce different behaviour. Smith and Kantor used results typically from the 300th position in Google's results: even today, these are unreliable for the simplest of our topics, and in 2008 will almost certainly have produced a poor result set. Importantly, our *diluted* system always included a few high-ranked results.

Either way, our results raise an important question about how the effectiveness of search systems should be analysed. While some fine-grained aspects of user clicking behaviour differed between the *full* and *diluted* treatments, the majority of behaviours did not. This outcome is in line with previous results that found little relationship between user behaviour and system quality as measured by common IR evaluation metrics such as MAP [6]. The question then becomes one of whether even a significant improvement in effectiveness, as measured by some metric, actually results in improved task performance. In future work, we therefore plan to systematically investigate different levels of answer-page dilution, to establish guidelines for the extent of practical differences that need to be present in search systems for measurable disparities in user behaviour to manifest. We also plan to explore the issue of the impact that specific variations in task instructions have on searcher behaviour through a controlled user study in a work task-based framework [2].

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. SIGIR*, pages 3–10, Seattle, WA, 2006.

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

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

[4] C. Smith and P. Kantor. User adaptation: good results from poor systems. In *Proc. SIGIR*, pages 147–154, Singapore, 2008.

[5] P. Thomas, T. Jones, and D. Hawking. What deliberately degrading search quality tells us about discount functions. In *Proc. SIGIR*, pages 1107–1108, Beijing, China, 2011.

[6] A. Turpin and F. Scholer. User performance versus precision measures for simple web search tasks. In *Proc. SIGIR*, pages 11–18, Seattle, WA, 2006.

[7] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. 4th Information Interaction in Context Symp.*, pages 254–257, Nijmegen, The Netherlands, 2012.

Exploratory Search Missions for TREC Topics

Martin Potthast Matthias Hagen Michael Völske Benno Stein

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

We report on the construction of a new query log corpus that consists of 150 exploratory search missions, each of which corresponds to one of the topics used at the TREC Web Tracks 2009–2011. Involved in the construction was a group of 12 professional writers, hired at the crowdsourcing platform oDesk, who were given the task to write essays of 5000 words length about these topics, thereby inducing genuine information needs. The writers used a ClueWeb09 search engine for their research to ensure reproducibility. Thousands of queries, clicks, and relevance judgments were recorded. This paper overviews the research that preceded our endeavors, details the corpus construction, gives quantitative and qualitative analyses of the data obtained, and provides original insights into the querying behavior of writers. With our work we contribute a missing building block in a relevant evaluation setting in order to allow for better answers to questions such as: "What is the performance of today's search engines on exploratory search?" and "How can it be improved?" The corpus will be made publicly available.

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

Keywords: Query Log, Exploratory Search, Search Missions

1. INTRODUCTION

Humans frequently conduct task-based information search, i.e., they interact with search appliances in order to conduct the research deemed necessary to solve knowledge-intensive tasks. Examples include long-lasting interactions which may involve many search sessions spread out across several days. Modern web search engines, however, are optimized for the diametrically opposed task, namely to answer short-term, atomic information needs. Nevertheless, research has picked up this challenge: in recent years, a number of new solutions for exploratory search have been proposed and evaluated. However, most of them involve an overhauling of the entire search experience. We argue that exploratory search tasks are already being tackled, after all, and that this fact has not been sufficiently investigated. Reasons for this shortcoming can be found in the lack of publicly available data to be studied. Ideally, for any given task that fits the aforementioned description, one would have a large set of search interaction logs from a diversity of humans solving it. Obtaining such data, even for a single task, has not been done at scale until now. Even search companies, which have access to substantial amounts of raw query log data, face difficulties in discerning individual exploratory tasks from their logs.

In this paper, we contribute by introducing the first large corpus of long, exploratory search missions. The corpus was constructed via crowdsourcing by employing writers whose task was to write long essays on given TREC topics, using a ClueWeb09 search engine for research. Hence, our corpus forms a strong connection to existing evaluation resources that are used frequently in information retrieval. Further, it captures the way how average users perform exploratory search today, using state-of-the-art search interfaces. The new corpus is intended to serve as a point of reference for modeling users and tasks as well as for comparison with new retrieval models and interfaces. Key figures of the corpus are shown in Table 2.

After a brief review of related work, Section 2 details the corpus construction and Section 3 gives first quantitative and qualitative analyses, concluding with insights into writers' search behavior.

1.1 Related Work

To date, the most comprehensive overview of research on exploratory search systems is that of White and Roth [19]. More recent contributions not covered in this body of work include the approaches proposed by Morris et al. [13], Bozzon et al. [2], Cartright et al. [4], and Bron et al. [3]. Exploratory search is studied also within contextual IR and interactive IR, as well as across disciplines, including human computer interaction, information visualization, and knowledge management.

Regarding the evaluation of exploratory search systems, White and Roth [19] conclude that "traditional measures of IR performance based on retrieval accuracy may be inappropriate for the evaluation of these systems" and that "exploratory search evaluation [...] must include a mixture of naturalistic longitudinal studies" while "[...] simulations developed based on interaction logs may serve as a compromise between existing IR evaluation paradigms and [...] exploratory search evaluation." The necessity of user studies makes evaluations cumbersome and, above all, expensive. By providing part of the solution (a decent corpus) for free, we want to overcome the outlined difficulties. Our corpus compiles a solid database of exploratory search behavior, which researchers may use for comparison purposes as well as for bootstrapping simulations.

Regarding standardized resources to evaluate exploratory search, hardly any have been published up to now. White et al. [18] dedicated a workshop to evaluating exploratory search systems in which requirements, methodologies, as well as some tools have been proposed. Yet, later on, White and Roth [19] found out that still no "methodological rigor" has been reached—a situation which has not changed much until today. The departure from traditional evaluation methodologies (such as the Cranfield paradigm) and resources (especially those employed at TREC) has lead researchers to devise ad-hoc evaluations which are mostly incomparable across papers and which cannot be reproduced easily.

A potential source of data for the purpose of assessing current exploratory search behavior is to detect exploratory search tasks within raw search engine logs, such as the 2006 AOL query log [14].

However, most session detection algorithms deal with short term tasks only and the few algorithms that aim to detect longer search missions still have problems when detecting interesting semantic connections of intertwined search tasks [10, 12, 8]. In this regard, our corpus may be considered the first of its kind.

To justify our choice of an exploratory task, namely that of writing an essay about a given TREC topic, we refer to Kules and Capra [11], who manually identified exploratory tasks from raw query logs on a small scale, most of which turned out to involve writing on a given subject. Egusa et al. [6] describe a user study in which they asked participants to do research for a writing task, however, without actually writing something. This study is perhaps closest to ours, although the underlying data has not been published. The most notable distinction is that we asked our writers to actually write, thereby creating a much more realistic and demanding state of mind since their essays had to be delivered on time.

2. CORPUS CONSTRUCTION

As discussed in the related work, essay writing is considered a valid approach to study exploratory search. Two data sets form the basis for constructing a respective corpus, namely (1) a set of topics to write about and (2) a set of web pages to research about a given topic. With regard to the former, we resort to topics used at TREC, specifically to those from the Web Tracks 2009–2011. With regard to the latter, we employ the ClueWeb09 (and not the "real web in the wild"). The ClueWeb09 consists of more than one billion documents from ten languages; it comprises a representative cross-section of the real web, is a widely accepted resource among researchers, and it is used to evaluate the retrieval performance of search engines within several TREC tracks. The connection to TREC will strengthen the compatibility with existing evaluation methodology and allow for unforeseen synergies. Based on the above decisions, our corpus construction steps can be summarized as follows:

1. Rephrasing of the 150 topics used at the TREC Web Tracks 2009–2011 so that they invite people to write an essay.
2. Indexing of the English portion of the ClueWeb09 (about 0.5 billion documents) using the BM25F retrieval model plus additional features.
3. Development of a search interface that allows for answering queries within milliseconds and that is designed along the lines of commercial search interfaces.
4. Development of a browsing interface for the ClueWeb09, which serves ClueWeb09 pages on demand and which rewrites links on delivered pages so that they point to their corresponding ClueWeb09 pages on our servers.
5. Recruiting 12 professional writers at the crowdsourcing platform oDesk from a wide range of hourly rates for diversity.
6. Instructing the writers to write essays of at least 5000 words length (corresponds to an average student's homework assignment) about an open topic among the initial 150, using our search engine and browsing only ClueWeb09 pages.
7. Logging all writers' interactions with the search engine and the ClueWeb09 on a per-topic basis at our site.
8. Double-checking all of the 150 essays for quality.

After the deployment of the search engine and successfully completed usability tests (see Steps 2-4 and 7 above), the actual corpus construction took nine months, from April 2012 through December 2012. The post-processing of the data took another four months, so that this corpus is among the first, late-breaking results from our efforts. However, the outlined experimental setup can obviously serve different lines of research. The remainder of the section presents elements of our setup in greater detail.

Used TREC Topics.

Since the topics from the TREC Web Tracks 2009–2011 were not amenable for our purpose as is, we rephrased them so that they ask for writing an essay instead of searching for facts. Consider for example topic 001 from the TREC Web Track 2009:

> *Query.* obama family tree
>
> *Description.* Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc.
>
> *Sub-topic 1.* Find the TIME magazine photo essay "Barack Obama's Family Tree."
>
> *Sub-topic 2.* Where did Barack Obama's parents and grandparents come from?
>
> *Sub-topic 3.* Find biographical information on Barack Obama's mother.

This topic is rephrased as follows:

> *Obama's family*. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

In the example, Sub-topic 1 is considered too specific for our purposes while the other sub-topics are retained. TREC Web track topics divide into faceted and ambiguous topics. While topics of the first kind can be directly rephrased into essay topics, from topics of the second kind one of the available interpretations is chosen.

A Search Engine for Controlled Experiments.

To give the oDesk writers a familiar search experience while maintaining reproducibility at the same time, we developed a tailored search engine called ChatNoir [15]. Besides ours, the only other public search engine for the ClueWeb09 is hosted at Carnegie Mellon and based on Indri. Unfortunately, it is far from our efficiency requirements. Our search engine returns results after a couple of hundreds of milliseconds, its interface follows industry standards, and it features an API that allows for user tracking.

ChatNoir is based on the BM25F retrieval model [17], uses the anchor text list provided by Hiemstra and Hauff [9], the PageRanks provided by the Carnegie Mellon University,[1] and the spam rank list provided by Cormack et al. [5]. ChatNoir comes with a proximity feature with variable-width buckets as described by Elsayed et al. [7]. Our choice of retrieval model and ranking features is intended to provide a reasonable baseline performance. However, it is neither near as mature as those of commercial search engines nor does it compete with the best-performing models proposed at TREC. Yet, it is among the most widely accepted models in the information retrieval community, which underlines our goal of reproducibility.

In addition to its retrieval model, ChatNoir implements two search facets: text readability scoring and long text search. The former facet, similar to that provided by Google, scores the readability of a text found on a web page via the well-known Flesh-Kincaid grade level formula: it estimates the number of years of education required in order to understand a given text. This number is mapped onto the three categories "simple", "intermediate", and "expert." The long text search facet omits search results which do not contain at least one continuous paragraph of text that exceeds 300 words. The two facets can be combined with each other. They are meant to support writers that want to reuse text from retrieved search results. Especially interesting for this type of writers are result documents containing longer text passages and documents of a specific reading

[1] http://boston.lti.cs.cmu.edu/clueWeb09/wiki/tiki-index.php?page=PageRank

Table 1: Demographics of the twelve writers employed.

Writer Demographics					
Age		*Gender*		*Native language(s)*	
Minimum	24	Female	67%	English	67%
Median	37	Male	33%	Filipino	25%
Maximum	65			Hindi	17%
Academic degree		*Country of origin*		*Second language(s)*	
Postgraduate	41%	UK	25%	English	33%
Undergraduate	25%	Philippines	25%	French	17%
None	17%	USA	17%	Afrikaans, Dutch,	
n/a	17%	India	17%	German, Spanish,	
		Australia	8%	Swedish each	8%
		South Africa	8%	None	8%
Years of writing		*Search engines used*		*Search frequency*	
Minimum	2	Google	92%	Daily	83%
Median	8	Bing	33%	Weekly	8%
Standard dev.	6	Yahoo	25%	n/a	8%
Maximum	20	Others	8%		

level such that reusing text from the results still yields an essay with homogeneous readability.

When clicking on a search result, ChatNoir does not link into the real web but redirects into the ClueWeb09. Though ClueWeb09 provides the original URLs from which the web pages have been obtained, many of these page may have gone or been updated since. We hence set up an interface that serves web pages from the ClueWeb09 on demand: when accessing a web page, it is pre-processed before being shipped, removing all kinds of automatic referrers and replacing all links to the real web with links to their counterpart inside ClueWeb09. This way, the ClueWeb09 can be browsed as if surfing the real web and it becomes possible to track a user's movements. The ClueWeb09 is stored in the HDFS of our 40 node Hadoop cluster, and web pages are fetched with latencies of about 200ms. ChatNoir's inverted index has been optimized to guarantee fast response times, and it is deployed on the same cluster.

Hired Writers.

Our ideal writer has experience in writing, is capable of writing about a diversity of topics, can complete a text in a timely manner, possesses decent English writing skills, and is well-versed in using the aforementioned technologies. This wish list lead us to favor (semi-)professional writers over, for instance, volunteer students recruited at our university. To hire writers, we made use of the crowdsourcing platform oDesk.[2] Crowdsourcing has quickly become one of the cornerstones for constructing evaluation corpora, which is especially true for paid crowdsourcing. Compared to Amazon's Mechanical Turk [1], which is used more frequently than oDesk, there are virtually no workers at oDesk submitting fake results due to advanced rating features for workers and employers.

Table 1 gives an overview of the demographics of the writers we hired, based on a questionnaire and their resumes at oDesk. Most of them come from an English-speaking country, and almost all of them speak more than one language, which suggests a reasonably good education. Two thirds of the writers are female, and all of them have years of writing experience. Hourly wages were negotiated individually and range from 3 to 34 US-dollars (dependent on skill and country of residence), with an average of about 12 US-dollars. In total, we spent 20 468 US-dollars to pay the writers.

3. CORPUS ANALYSIS

This section presents the results of a preliminary corpus analysis that gives an overview of the data and sheds some light onto the search behavior of writers doing research.

[2] http://www.odesk.com

Table 2: Key figures of our exploratory search mission corpus.

Corpus Characteristic	Distribution min	avg	max	stdev	Σ
Writers					12
Topics					150
Topics / Writer	1	12.5	33	9.3	
Queries					13 651
Queries / Topic	4	91.0	616	83.1	
Clicks					16 739
Clicks / Topic	12	111.6	443	80.3	
Clicks / Query	0	0.8	76	2.2	
Sessions					931
Sessions / Topic	1	12.3	149	18.9	
Days					201
Days / Topic	1	4.9	17	2.7	
Hours					2068
Hours / Writer	3	129.3	679	167.3	
Hours / Topic	3	7.5	10	2.5	
Irrelevant					5962
Irrelevant / Topic	1	39.8	182	28.7	
Irrelevant / Query	0	0.5	60	1.4	
Relevant					251
Relevant / Topic	0	1.7	7	1.5	
Relevant / Query	0	0.0	4	0.2	
Key					1937
Key / Topic	1	12.9	46	7.5	
Key / Query	0	0.2	22	0.7	

Corpus Statistics.

Table 2 shows key figures of the query logs collected, including the absolute numbers of queries, relevance judgments, working days, and working hours, as well as relations among them. On average, each writer wrote 12.5 essays, while two wrote only one, and one very prolific writer managed more than 30 essays.

From the 13 651 submitted queries, each topic got an average of 91. Note that queries often were submitted twice requesting more than ten results or using different facets. Typically, about 1.7 results are clicked for consecutive instances of the same query. For comparison, the average number of clicks per query in the aforementioned AOL query log is 2.0. In this regard, the behavior of our writers on individual queries does not seem to differ much from that of the average AOL user in 2006. Most of the clicks we recorded are search result clicks, whereas 2457 of them are browsing clicks on web page links. Among the browsing clicks, 11.3% are clicks on links that point to the same web page (i.e., anchor links using a URL's hash part). The longest click trail observed lasted 51 unique web pages but most click trails are very short. This is surprising, since we expected a larger proportion of browsing clicks, but it also shows our writers relied heavily on the search engine. If this behavior generalizes, the need for a more advanced support of exploratory search tasks from search engines becomes obvious.

The queries of each writer can be divided into a total of 931 sessions with an average 12.3 sessions per topic. Here, a session is defined as a sequence of queries recorded on a given topic which is not divided by a break longer than 30 minutes. Despite other claims in the literature (e.g., in [10]), we argue that, in our case, sessions can be reliably identified by means of a timeout because of our a priori knowledge about which query belongs to which topic (i.e., task). Typically, finishing an essay took 4.9 days, which fits well the definition of exploratory search tasks being long-lasting.

In their essays, writers referred to web pages they found during their search, citing specific passages and topic-related information used in their texts. This forms an interesting relevance signal which allows us to separate irrelevant from relevant web pages. Slightly different to the terminology of TREC, we consider web pages referred to in an essay as key documents for its respective topic, whereas web pages that are on a click trail leading to a key document are

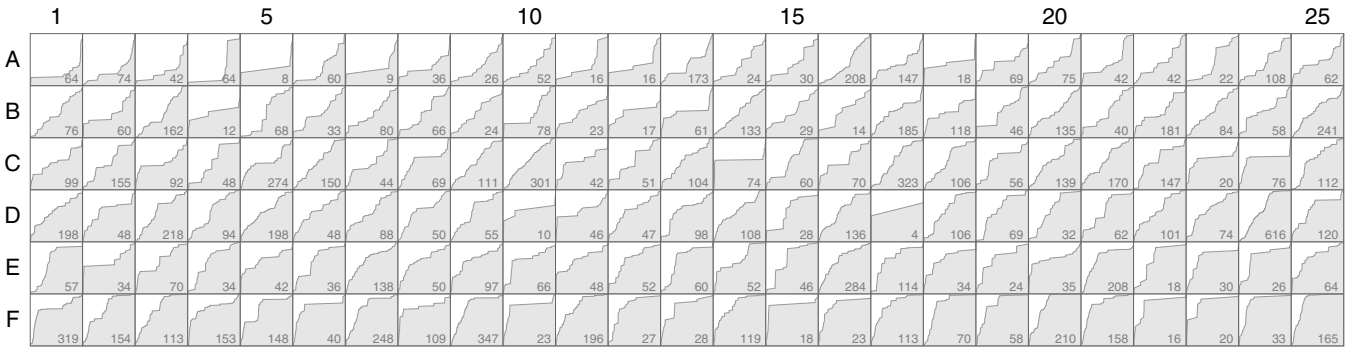


Figure 1: Spectrum of writer search behavior. Each grid cell corresponds to one of the 150 topics and shows a curve of the percentage of submitted queries (y-axis) at times between the first query until the essay was finished (x-axis). The numbers denote the amount of queries submitted. The cells are sorted by area under the curve from the smallest area in cell A1 to the largest area in cell F25.

relevant. The fact, that there are only few click trails of this kind explains the unusually high number of key documents compared to that of relevant ones. The remainder of web pages which were accessed but discarded by our writers may be considered irrelevant.

The writer's search interactions are made freely available as the Webis-Query-Log-12.[3] Note that the writing interactions are the focus of our accompanying ACL paper [16] and contained in the Webis text reuse corpus 2012 (Webis-TRC-12).

Exploring Exploratory Search Missions.

To get an inkling of the wealth of data in our corpus, and how it may influence the design of exploratory search systems, we analyze the writers' search behavior during essay writing. Figure 1 shows for each of the 150 topics a curve of the percentage of queries at any given time between a writer's first query and an essay's completion. We have normalized the time axis and excluded working breaks of more than five minutes. The curves are organized so as to highlight the spectrum of different search behaviors we have observed: in row A, 70–90% of the queries are submitted toward the end of the writing task, whereas in row F almost all queries are submitted at the beginning. In between, however, sets of queries are often submitted in short "bursts," followed by extended periods of writing, which can be inferred from the plateaus in the curves (e.g., cell C12). Only in some cases (e.g., cell C10) a linear increase of queries over time can be observed for a non-trivial amount of queries, which indicates continuous switching between searching and writing.

From these observations, it can be inferred that query frequency alone is not a good indicator of task completion or the current stage of a task, but different algorithms are required for different mission types. Moreover, exploratory search systems have to deal with a broad subset of the spectrum and be able to make the most of few queries, or be prepared that writers interact only a few times with them. Our ongoing research on this aspect focuses on predicting the type of search mission, since we found it does not simply depend on the writer or a topic's difficulty as perceived by the writer.

4. SUMMARY

We introduce the first corpus of search missions for the exploratory task of writing. The corpus is of representative scale, comprising 150 different writing tasks and thousands of queries, clicks, and relevance judgments. A preliminary corpus analysis shows the wide variety of different search behavior to expect from a writer conducting research online. We expect further insights from a forthcoming in-depth analysis, whereas the results mentioned demonstrate the utility of our publicly available corpus.

[3] http://www.webis.de/research/corpora

5. REFERENCES

[1] J. Barr and L. F. Cabrera. AI gets a brain. *Queue*, 4(4):24–29, 2006.

[2] A. Bozzon, M. Brambilla, S. Ceri, and P. Fraternali. Liquid query: multi-domain exploratory search on the web. *Proc. of WWW 2010*.

[3] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. *Proc. of SIGIR 2012*.

[4] M.-A. Cartright, R. White, and E. Horvitz. Intentions and attention in exploratory health search. *Proc. of SIGIR 2011*.

[5] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011.

[6] Y. Egusa, H. Saito, M. Takaku, H. Terai, M. Miwa, and N. Kando. Using a concept map to evaluate exploratory search. *Proc. of IIiX 2010*.

[7] T. Elsayed, J. Lin, and D. Metzler. When close enough is good enough: approximate positional indexes for efficient ranked retrieval. *Proc. of CIKM 2011*.

[8] M. Hagen, J. Gommoll, A. Beyer, and B. Stein. From search session detection to search mission detection. *Proc. of SIGIR 2012*.

[9] D. Hiemstra and C. Hauff. MIREX: MapReduce information retrieval experiments. Tech. Rep. TR-CTIT-10-15, University of Twente, 2010.

[10] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. *Proc. of CIKM 2008*.

[11] B. Kules and R. Capra. Creating exploratory tasks for a faceted search interface. *Proc. of HCIR 2008*.

[12] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. *Proc. of WSDM 2011*.

[13] D. Morris, M. Ringel Morris, and G. Venolia. SearchBar: a search-centric web history for task resumption and information re-finding. *Proc. of CHI 2008*.

[14] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. *Proc. of Infoscale 2006*.

[15] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. ChatNoir: a search engine for the ClueWeb09 corpus. *Proc. of SIGIR 2012*.

[16] M. Potthast, M. Hagen, M. Völske, and B. Stein. Crowdsourcing interaction logs to understand text reuse from the web. *Proc. of ACL 2013*.

[17] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. *Proc. of CIKM 2004*.

[18] R. White, G. Muresan, and G. Marchionini, editors. *Proc. of SIGIR workshop EESS 2006*.

[19] R. White and R. Roth. *Exploratory search: beyond the query-response paradigm*. Morgan & Claypool, 2009.

Interactive Exploration of Geographic Regions with Web-based Keyword Distributions

Chandan Kumar
University of Oldenburg,
Oldenburg, Germany
chandan.kumar@uni-oldenburg.de

Dirk Ahlers
NTNU – Norwegian University of Science and Technology,
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

Wilko Heuten
OFFIS – Institute for Information Technology,
Oldenburg, Germany
wilko.heuten@offis.de

Susanne Boll
University of Oldenburg,
Oldenburg, Germany
susanne.boll@uni-oldenburg.de

ABSTRACT

The most common and visible use of geographic information retrieval (GIR) today is the search for specific points of interest that serve an information need for places to visit. However, in some planning and decision making processes, the interest lies not in specific places, but rather in the makeup of a certain region. This may be for tourist purposes, to find a new place to live during relocation planning, or to learn more about a city in general. Geospatial Web pages contain rich spatial information content about the geo-located facilities that could characterize the atmosphere, composition, and spatial distribution of geographic regions. But the current means of Web-based GIR interfaces only support the sequential search of geo-located facilities and services individually, and limit the end users on abstracted view, analysis and comparison of urban areas. In this work we propose a system that abstracts from the places and instead generates the makeup of a region based on extracted keywords we find on the Web pages of the region. We can then use this textual fingerprint to identify and compare other suitable regions which exhibit a similar fingerprint. The developed interface allows the user to get a grid overview, but also to drill in and compare selected regions as well as adapt the list of ranked keywords.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

Keywords

Geographic information retrieval, Spatial Web, Geographic regions, Keyword distributions, Visualization, Interaction

1. INTRODUCTION

Geospatial search has become a widely accepted search mode offered by many commercial search engines. Their interfaces can easily be used to answer relatively simple requests such as "*restaurant in Berlin*" on a point-based map interface, which additionally gives extended information about entities [1]. A corresponding strong research interested has developed in the field of geographic information retrieval, e.g., [2, 17, 15]. However, there are many tasks in which the retrieval of individual pinpointed entities such as facilities, services, businesses, or infrastructure cannot satisfy user's more complex spatial information needs.

To support more complex tasks we propose a new retrieval method based on entities. For example, sometimes the distribution of results on a map can already inform certain views about areas, e.g., a search for "bar" may show a clustering of results that can be used for "eyeballing" a region of nightlife even without sophisticated geospatial analysis. However, as users become more used to local search, more complex search types and supporting analysis are desired that enable a combined view onto the underlying data [10]. Exploration of geographic regions and their characterization was found as one of the key desire of local search users in our requirement study [11]. A person who is moving to a new area or city would like to find similar neighborhoods or regions with a similar makeup to their current home. It might not even be the concrete entities, but rather the atmosphere, composition, and spatial distribution that make up the "feeling" of a neighborhood that best capture the intention of a user. To assess this similarity of regions we propose a spatial fingerprint (query-by-spatial-example) that acts as an abstracted view onto the same point-based data.

We also aim to provide new visual tools for the exploration of geographic regions. While the necessary multi-dimensional geospatial data is already available, there is no suitable interface to query them, let alone to deal with the multi-criteria complexity. In this paper we describe a visual-interactive GIR system to support the retrieval of relevant geospatial regions and enable users to explore and interact with geospatial data. We propose a new query-by-spatial-example in-

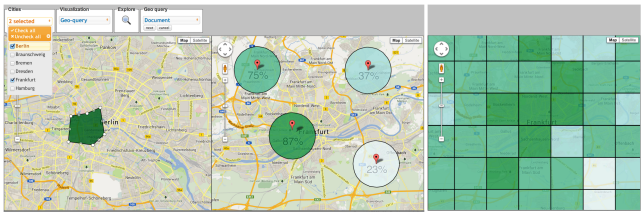


Figure 1: Geographic querying and ranking of geographic regions, with user-selected target regions and alternative grid view

teraction method in which a user-selected region's characteristic is fingerprinted to present similar regions. Users can interactively refine their query to use those characteristics of a region that are most important to them. For a more detailed overview, we use the full text of georeferenced Web pages for queries and analysis. This work goes beyond conventional GIR interfaces as it allows users to interact with aggregated spatial information via spatial queries instead of only textual querying, which is especially important to define regions of interest. We discuss the necessary input, visualization, comparison, refinement, and ranking methods in the remainder of this paper.

2. USING THE GEOSPATIAL WEB TO CHARACTERIZE GEOGRAPHIC REGIONS

The distribution of geo-entities is used to illustrate the characteristics and dynamics of a geographic region. A geo-entity is a real life entity at a physical location, e.g., a restaurant, theatre, pub, museum, business, school, etc. To open these entities up for aggregate and multi-criteria region characterization, they need a certain depth of information associated with them. It is obvious that only position information or the name of a place is insufficient, so categorial or textual description is needed. For initial studies [11, 9] we used OpenStreetMap (OSM)[1] which uses a tagging system for categories. To better characterize the geo entities we now use their associated Web pages. The reason for this is the massive increase of the amount of usable data. The Web pages of entities contain a lot more than just the basic information and can therefore be used to uncover much more detailed information. This method can also include additional sources such as events happening in the region or user-generated content on third-party pages [2]. We later describe how we identify the most meaningful keywords from the pages for this task.

To actually make the connection from a location to Web pages, we assume that the presence of location references on a page is a strong indication that the page is associated with the entity at that location. We use our geoparser to extract location references and thereby assess the geographical scopes of a page. The geoparser is trained to the presence of location references in the form of addresses within the page content. This is a suitable approach for the urban areas we are addressing in this work, because we need a geospatial granularity at the sub-neighborhood level. Knowledge-based identification and verification of the addresses is done against a gazetteer extended with street names, which we

[1] `http://www.openstreetmap.org/`

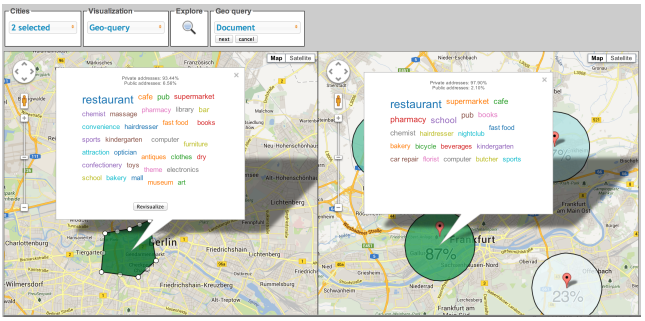


Figure 2: Keyword-based visual comparison of geographic regions

fetched from OSM for the major cities of Germany. To retrieve actual pages, we crawled the Web with a geospatially focused crawler [3] based on the geoparser and built a rich geo-index for various cities of Germany, where each city contains several thousand geotagged Web pages with their full textual content.

3. INTERFACE FOR EXPLORATION OF GEOGRAPHIC REGIONS OF INTEREST

We have implemented two main interaction modes in the Web interface as shown in Figure 1. A user intends to compare multiple geographic regions of Frankfurt (target region, right in the dual-map view) with respect to a certain relevant region in Berlin (query region, left). The current reference region of interest is specified via a visual query. The user can then either select regions by placing markers onto the map, or alternatively use a grid overview (right side of Figure 1). In both cases, the system computes the relevance of the target regions with respect to the characteristics of the query region.

3.1 Query-by-spatial-example

Most GIR interfaces use a conventional textual query as input method to describe user's information need or use the currently selected map viewport. We wanted to give users the ability to arbitrarily define their own spatial region of interest. The free definition of the query region is important, as users may not always want a neighborhood that is easily describable by a textual query. We therefore enabled to query by spatial example, where users can define the query region by drawing on map. Figure 1 shows an example of a user selected region of interest via a polygon query (by mouse clicks and drag) in the city of Berlin.

3.2 Visualization of suitable geographic regions

Users can select several location preferences in their target region that they would like to explore by positioning markers on the map interface. The system defines the targets with a circle around the user-selected locations with the same diameter as the reference region polygon. The target regions obtain the ranking with respect to their similarity with the reference region. Their relevance is shown by the percentage similarity and the heatmap based relevance visualization. We used a color scheme of different green tones which differed in their transparency. Light colors represented low

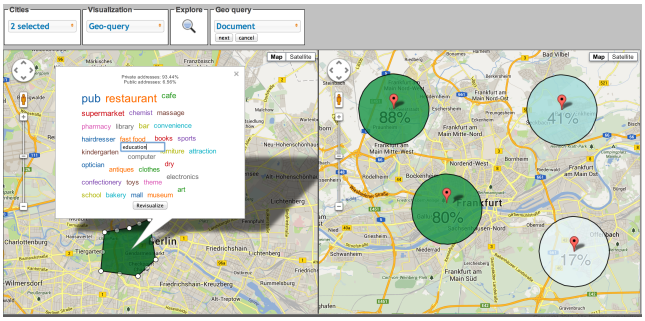


Figure 3: User interaction with the keyword distribution and revisualization

relevance, dark colors were used to indicate high relevance. The color scheme selection was aided by ColorBrewer [2].

As an example, Figure 1 shows 4 user-selected locations on the city map of Frankfurt, the circle regions around these 4 markers have the same diameter as the query region in Berlin. The target region in the centre of the city is most relevant with the similarity of 88%, and consequently has the darkest green tone. If a user has not yet formed any preference, we offer an aggregate overview of geo-entities. We partition the map area using a grid raster [14], as we do not intend to restrict user exploration to only selected areas. There could be situations when users look beyond the specific target regions, and would like to have an overview of the whole city with respect to a query region. The right side of Figure 1 shows the aggregated ranked view of the grid-based visualization. Each grid cell represents the overall relevance with respect to the query region. The visualization gives a good overview and assessment on relevant regions which the user can then explore further. Users can select the grid size, which is otherwise similar to the size of the query region. The grid layout is fixed to the city boundaries as we intend to give the overview of whole city. In the future we would like to make it more dynamic where users should be able to shift the grid layout, since a slight variation in grid cell boundaries could alter the relevance results.

3.3 Exploration and interaction with geographic regions via keyword distributions

Interaction models should provide end users the opportunity to explore the characteristics of selected regions, and adapt it further to their requirements. We initially show the most relevant keywords of the respective region using a word cloud. The word cloud provides more detailed information on keyword distribution when the mouse hovers over it. The font size and order of the keywords signify their relevance. Figure 2 shows the comparison of the query region with the most relevant target region via both their keyword distributions. In this case, the distributions of both regions are very similar, leading to the high relevance score for the target region.

Since the keyword characteristics of a query region is derived from the georeferenced Web pages, there are situations where a user might not be satisfied with the spatial description and wants to influence the keywords. In the example of Figure 3, a user decides that pubs are more important than restaurant, fast food is not an aspect of his lifestyle and should be replaced by education facilities near his new home. In such scenarios users need to interact and adapt the generated keyword distributions of query regions. We make the word cloud interactive and editable. Users can drag keywords to alter their position and thus their significance. They can also edit, delete or replace keywords in the word cloud to change the criteria. After modifying the keyword distribution, users can revisualize the target regions to update their ranking. Figure 3 shows this user interaction with the word cloud, including the revisualization of the updated ranking of target regions, which are visibly different from the previous ranking of Figure 2.

[2] `http://colorbrewer2.org`

4. TEXT-BASED CHARACTERIZATION AND RANKING OF GEOGRAPHIC REGIONS

We adapt common IR methods for ranking and similarity measures. In relevance-based language models, the similarity of a document to a query is the probability that a given document would generate the query [12]. To be able to do the same with geographic regions, we add a transitional step. Regions are considered as compound documents built from the Web pages of the entities inside them. We can then define the similarity of document clusters of regions based on the probability that the target region can generate the query region. The Kullback-Leibler divergence is used for comparison [4].

For a geospatial document d, we estimate $P(w|d)$, which is a unigram language model , with the maximum likelihood estimator, simply given by relative counts: $P(w|d) = \frac{tf(w,d)}{|d|}$, here $tf(w,d)$ is the frequency of word w in the document d and $|d|$ is the length of the document d. A geographic region contains several geospatial documents insides its footprint area. We define a geographic region based on a document cluster D which contains document $\{d_1, d_2....d_k\}$, and the distribution of a particular word w in the geographic region would be estimated with its combine probability in the collection $P(w|D) = \frac{1}{k}\sum_{i=1}^{k} P(w|d_i)$. The word cloud represents the most prominent keywords of the region with respect to their ranked probability distribution $P(w|D)$. The comparison of regions is done with respect to their probability distribution using KL-divergence. A target region x will be compared to the query region as following

$$Relevance(Region_x) = \sum_w P(w|D_q) log \frac{P(w|D_q)}{P(w|D_x)}$$

The computation of this formula involves a sum over all the words that have a non-zero probability according to $P(w|D_q)$. Each region $Region_x$ gets a relevance score according to its distribution comparison to the query region $Region_q$. All target regions (user selected regions or grid based divisions) are ranked with respect to their relevance score for visualization.

5. RELATED WORK

The field of geographic information retrieval examines documents' geospatial features at a regional scale and also at smaller granularities and usually supports keyword@location queries [2, 17, 15]. Similarly, location-based services (e.g.,

FourSquare, Yelp, Google Maps) allow users to retrieve and visualize geo-entities matching a category or search term. However, search for multiple categories or other complex tasks is usually not supported. Some non-conventional spatial querying methods have been proposed, e.g., query-by-sketch on a map [6]. Other work uses the density of arbitrary user-supplied keywords to build a query region [8]. Tag clouds have been adapted to maps, exploiting georeferenced tags [16]. Locally characteristic keywords can be extracted for map visualization and to show their spatial extent [19]. None of these approaches make a larger word cloud available, but only the main terms. Other geovisualization approaches [5, 7] approach multi-criteria analysis, but are usually targeted to specific domains and experts. The Inspect system was tailored at geospatial analysts to visually filter and explore multidimensional data [13]. A multi-criteria evaluation for home buyers was proposed in [18]. The scenario of spatial decision making is similar to ours, but it focused on experts and spatial computation issues rather than interface and visualization aspects.

Our system interface differs in the granularity of information need and representation, i.e., we focus on the ranking of regions, but base it on high-granularity geo-entities that have a very exact location, which ensures that the spatial query does not produce overlap to neighboring regions and makes the multi-criteria analysis more exact to be executed at arbitrary region sizes.

6. CONCLUSIONS AND FUTURE WORK

Most current local search interfaces do not offer adequate support for the exploration and comparison of geographic areas and regions. End users need visual and interactive assistance from GIR systems for an abstracted overview and analysis of geospatial data. We proposed interactive interfaces for the characterization and assessment of relevant geographic regions that enable end-users to query, analyze and interact with the rich geospatial data available on the Web in user-selected geographic regions. The relevance of regions is based on the similarity of keyword distributions.

The observation of results shows satisfactory performance by uncovering realistic and meaningful keywords defining the regions. We observed that the characterization and comparison of geographic regions show good results with respect to geo-located facilities and infrastructure of German cities, e.g., clearly distinct characteristics for university, industrial, or party districts. In the future we plan a more formal qualitative and quantitative evaluation of these interfaces, to examine the acceptance of these visualizations with regard to user-centered aspects such as exploration ability, information overload, and cognitive demand. We would also like to explore more advanced interaction methods to enhance the usability of the proposed visualizations.

Additionally, we envision more powerful region similarity measures such as landscape and topological similarity, similarity via social media, and an integration of additional data sources.

Acknowledgments

The authors are grateful to the DFG SPP 1335 'Scalable Visual Analytics' priority program which funds the project UrbanExplorer. The 2nd author acknowledges funding from the ERCIM "Alain Bensoussan" Fellowship Programme.

7. REFERENCES

[1] D. Ahlers. Local Web Search Examined. In *Web Search Engine Research*. Emerald, 2012.

[2] D. Ahlers and S. Boll. Location-based Web search. In *The Geospatial Web*. Springer, 2007.

[3] D. Ahlers and S. Boll. Adaptive Geospatially Focused Crawling. In *CIKM '09*, 2009.

[4] T. M. Cover and J. A. Thomas. Elements of information theory, 1991.

[5] J. Dykes, A. M. MacEachren, and M.-J. Kraak. *Exploring Geovisualization*. Elsevier, 2005.

[6] M. J. Egenhofer. Query processing in spatial-query-by-sketch. *J. Vis. Lang. Comput.*, 8, 1997.

[7] R. Greene et al. GIS-based multiple-criteria decision analysis. *Geography Compass*, 5(6), 2011.

[8] A. Henrich and V. Lüdecke. Measuring Similarity of Geographic Regions for Geographic Information Retrieval. In *ECIR '09*, 2009.

[9] C. Kumar, W. Heuten, and S. Boll. Visual interfaces to support spatial decision making in geographic information retrieval. In *CD-ARES 2013*. to appear.

[10] C. Kumar, W. Heuten, and S. Boll. Geovisualization for end user decision support: Easy and effective exploration of urban areas. In *GeoViz_Hamburg 2013: Interactive Maps That Help People Think*, 2013.

[11] C. Kumar, B. Poppinga, D. Haeuser, W. Heuten, and S. Boll. Geovisual interfaces to find suitable urban regions for citizens: A user-centered requirement study. In *UbiComp'13 Adjunct*, 2013. to appear.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. SIGIR '01. ACM, 2001.

[13] S.-J. Lee et al. Inspect: a dynamic visual query system for geospatial information exploration. In *SPIE*, 2003.

[14] A. M. MacEachren and D. DiBiase. Animated maps of aggregate data: Conceptual and practical problems. *CaGIS*, 18(4), 1991.

[15] A. Markowetz et al. Design and Implementation of a Geographic Search Engine. In *WebDB 2005*, 2005.

[16] D.-Q. Nguyen and H. Schumann. Taggram: Exploring geo-data on maps through a tag cloud-based visualization. In *IV'10*, 2010.

[17] R. S. Purves et al. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *IJGIS*, 21(7), 2007.

[18] C. Rinner and A. Heppleston. The spatial dimensions of multi-criteria evaluation – case study of a home buyer's spatial decision support system. In *Geographic Information Science*. 2006.

[19] B. Thomee and A. Rae. Uncovering locally characterizing regions within geotagged data. WWW '13, 2013.

Inferring Music Selections for Casual Music Interaction

Daniel Boland
University of Glasgow
United Kingdom
daniel@dcs.gla.ac.uk

Ross McLachlan
University of Glasgow
United Kingdom
r.mclachlan.1@research.gla.ac.uk

Roderick Murray-Smith
University of Glasgow
United Kingdom
rod@dcs.gla.ac.uk

ABSTRACT

We present two novel music interaction systems developed for casual exploratory search. In casual search scenarios, users have an ill-defined information need and it is not clear how to determine relevance. We apply Bayesian inference using evidence of listening intent in these cases, allowing for a belief over a music collection to be inferred. The first system using this approach allows users to retrieve music by subjectively tapping a song's rhythm. The second system enables users to browse their music collection using a radio-like interaction that spans from casual mood-setting through to explicit music selection. These systems embrace the uncertainty of the information need to infer the user's intended music selection in casual music interactions.

Categories and Subject Descriptors

H.5.2 [**Information interfaces**]: User Interfaces

General Terms

Design, Human Factors, Theory

1. INTRODUCTION

When interacting with a music system, listeners are faced with selecting songs from increasingly large music collections. With services like Spotify, these libraries can include many songs the user has never heard of. This retrieval is often a hedonic activity and may not serve a particular information need. Users do not always have a song in mind and are often just interested in setting a mood or finding something 'good enough' [9]. This type of *casual search* has recently been identified as not being well supported within IR literature [14]. In particular, the concept of relevance becomes nebulous where the information need is not well defined. By inferring a belief over a music collection using the likelihood of a user's input, we implement interactions which incorporate this uncertainty. These interactions can account for subjectivity and span from casual, serendipitous listening through to highly engaged music selection.

Music listeners are not always fully engaged with the selection of music - as evidenced by the success of the shuffle playback feature. Large libraries of music such as Spotify are available but users often just want background music, not a specific song out of millions. In these casual search scenarios, users often *satisfice* i.e. search for something which is 'good enough' [11]. As this information need is poorly defined, so too is relevance, placing these interactions outside of typical Information Retrieval approaches.

2. UNCERTAIN MUSIC SELECTION

By asking *'What would this user do?'*, we can develop a likelihood model of user input within an interaction. With Bayes theorem, this allows for an uncertain belief over a music space to be inferred. Users can provide evidence of their listening intent as part of a casual music interaction, not needing to be fully engaged in the music retrieval. This is an explicitly user-centered approach, focusing on how a user will interact with the system. Both the systems discussed here have been iteratively developed by comparing real user behaviour against that predicted by the user input models. We present two novel music retrieval systems which explore two challenges with this approach: i) how to correctly interpret evidence which may be subjective and ii) how to allow users to set their current level of engagement:

i) 'Query by Tapping' is a music retrieval technique where users tap the rhythm of a song in order to retrieve it [1]. As part of a user-centred development process, we identified that rhythmic queries are often subjective and so developed a model of rhythmic input which captures some of this subjective behaviour. This allows for the system to be trained to the user's tapping style, giving significant improvements over previous efforts at rhythmic music retrieval.

ii) FineTuner is a prototype of a radio-like music interface that enables users to retrieve music at a level of engagement suited to their current information need. Users navigate their music collection using a dial, with the system using prior knowledge of the user to inform the music selection. A pressure sensor enables users to assert varying levels of control over the system – with no pressure, users can casually tune in to sections of their music collection to hear recommended music with common characteristics. As pressure is applied, the user is able to make increasingly specific selections from the collection. The inferred music selection is conditioned upon the asserted control, allowing for the seamless transition from casual mood-setting to engaged music interaction.

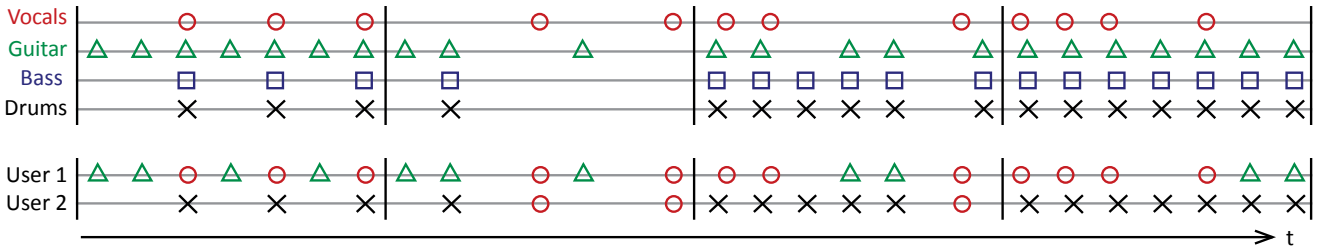


Figure 1: Users construct queries by sampling from preferred instruments. User 1 prefers Vocals and Guitar whereas User 2 prefers Drums and Bass.

3. MODELLING SUBJECTIVITY

In this section we describe our efforts to model the subjectivity of rhythmic queries, yielding a query by tapping system for casual music retrieval which can be trained to users to account for their subjective querying style. After training the system, a user can tap a rhythm to re-order their music collection by rhythmic similarity to their query. The top 20 highly ranked results are listed on-screen as a music playlist, from which the user could also then select a specific song.

Query by tapping provides an example of a casual music interaction which suffers from subjective queries. In mobile music-listening contexts, it can often be inconvenient for users to remove their mobile device from their pocket or bag and engage with it to select music. This tapping of music as a querying technique for music is depicted in figure 2. Tapping a rhythm is already a common act and rhythm is a universal aspect of music [13]. In an exploratory design session where users were asked to provide rhythmic queries, it became apparent that users differed in querying style. We describe this subjective behaviour and our approach to modelling it in previous work [1]. One of the key aspects of the model is that users have preferences for which instruments they tap to, as depicted in figure 1.

In order to assign a belief to the songs in the music collection given a rhythmic query, we compare the query to those predicted by the user input model. This comparison is done using the edit distance from string comparison methods, scaling the mismatch penalty to the time differences between the rhythmic sequences [5].

3.1 Query By Tapping

'Query by Tapping' has received some consideration in the Music Information Retrieval community. The term was introduced in [7] which demonstrated that rhythm alone can be used to retrieve musical works, with their system yielding a top 10 ranking for the desired result 51% of the time. Their work is limited however in considering only monophonic rhythms i.e. the rhythm from only one instrument, as opposed to being polyphonic and comprising of multiple instruments. Their music corpus consists of MIDI representations of tunes such as "You are my sunshine" which is hardly analogous to real world retrieval of popular music.

Rhythmic interaction has been recognised in HCI [8, 15] with [4] introducing rhythmic queries as a replacement for hot-keys. In [2] tempo is used as a rhythmic input for exploring a music collection – indicating that users enjoyed such a method of interaction. The consideration of human factors is also an emerging trend in Music Information Retrieval [12]. Our work draws upon both these themes, being the first QBT system to adapt to users. A number of key techniques for QBT are introduced in [5] which describes rhythm as a sequence of time intervals between notes – termed inter-onset intervals (IOIs). They identify the need for such intervals to be defined relative to each other to avoid the user having to exactly recreate the music's tempo.

In previous implementations of QBT, each IOI is defined relative to the preceding one [5]. This sequential dependency compounds user errors in reproducing a rhythm, as an erroneous IOI value will also distort the following one.

Figure 2: Users are able to select music by simply tapping a rhythm or tempo on the device, enabling a casual eyes-free music interaction.

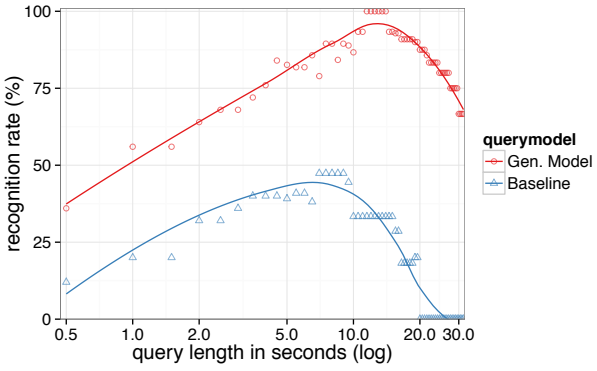


Figure 3: Percentage of queries yielding a highly ranked result (in the top 20 i.e. 6.7%) plotted against query length in seconds.

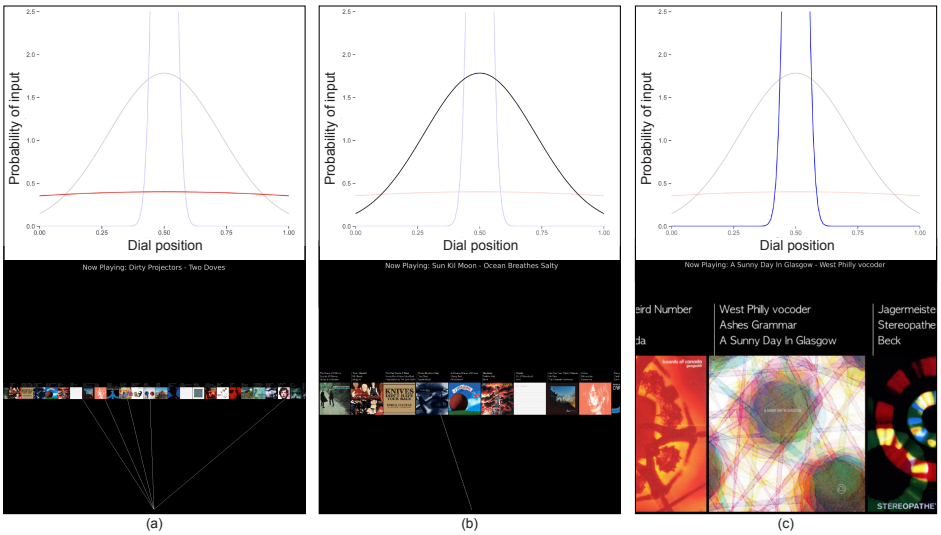


Figure 4: As the user asserts control, the distribution of predicted input for a given song becomes narrower. This adds weight to the input, meaning a belief is inferred over fewer songs and the view zooms in.

The approach to rhythmic interaction in [4] however used k-means clustering to classify taps and IOIs into three classes based on duration. The clustering based approach avoids the sequential error however loses a great deal of detail in the rhythmic query and so we explore a hybrid approace.

3.2 Evaluation

The most important metric for the system to be usable was whether a rhythmic input produced an on-screen (top 20) result. We asked eight participants to provide queries for songs selected from a corpus of 300 songs which we had complete note onset data for. Participants listened to the songs first to ensure familiarity and were asked to provide training queries for each song. These training queries were used to train the generative model using leave-one-out cross-validation. We use a state-of-the-art onset detection algorithm (based on measuring spectral flux [10]) as a baseline which does not account for subjectivity. Performance typically improves with query length as seen in figure 3. Higher rankings are achieved for all query lengths when using the generative model. Interestingly, queries over 10 seconds lead to a rapid fall-off in performance - possibly due to errors accumulating beyond the initial query the user had in mind or due to users becoming bored.

4. MODELLING ENGAGEMENT

We consider casual search interactions as spanning a range of levels of engagement. How much a user is willing to engage with a system and provide evidence of their listening intent will undoubtedly vary with listening context. An interaction which is fixedly casual would be as problematic as one which requires a user's full attention, with users unable to take control when they wish to. An example of this would be old analogue radios – whilst they offer a simple music interaction, users have limited control over what they hear. Previous work by Hopmann et al. sought to bring the benefits of interaction with vintage analog radio to modern digital music collections [6], however their work also required explicit selection (a fixed level of engagement).

We explore how the inference of listening intent can be conditioned upon the user's level of engagement, with the music interaction spanning from casual mood-setting through to specific song selection. While it would be desirable to bring the simplicity of radio-like interaction to modern music collections, mapping a modern music collection to a dial such as in figure 5 would require prolonged scrolling. An alternative would be to instead support scrolling through an overview of the music space however this removes granularity of control from the user, leaving them unable to select specific items. We developed a radio-like system called FineTuner that allows users to navigate their music, which is arranged along a mood axis. Users can 'tune in' to a mood to hear recommended songs based on their listening history. FineTuner allows the user to assert control over the music recommendation by applying pressure to a sensor. This enables users to seamlessly transition from a casual style of interaction akin to a radio to controlling styles such as specifying a particular sub-area of interest in a music space, or even selecting individual songs. FineTuner provides a single interaction which supports casual search through to fully engaged retrieval.

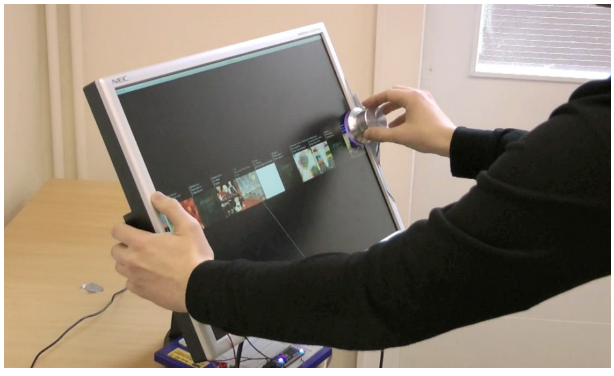


Figure 5: Users share control over an intelligent radio system, using a knob and pressure sensor.

4.1 Varying Engagement

Our system enables both casual and engaged forms of interaction, giving users varying degrees of control over the selection of music. In casual interactions where users apply less pressure, the system can become more autonomous – making inferences from prior evidence about what the user intended. This handover of control was termed the 'H-metaphor' by Flemisch et al. where it was likened to riding a horse – as the rider asserts less control the horse behaves more autonomously [3]. By allowing users to make selections from the *general* to the *specific*, the system supports both specific selections and satisficing. Users can make broad and uncertain *general* selections to casually describe what they want to listen to. However, they can also assert more control over the system and force it to play a *specific* song. Control is asserted by applying force to a pressure sensor.

As the user begins an interaction, they have not applied pressure and therefore are not asserting control over the system. The inferred selection is thus broad, covering an entire region of their collection and is biased towards popular tracks (fig. 4a). The music in the inferred selection is visualised by randomly sampling tracks from it and drawing beams from the dial position to the album art. The user may press in the knob to accept the selection and the sampled track is played. At low levels of assertion it is likely that most tracks played would be highly popular tracks. This behaviour is a design assumption, users may want the system to use other prior evidence. When the user applies pressure, the system interprets this as an assertion of control. The inferred selection is smaller and the spread of beams becomes narrower, the album art visualisation zooms in to show the smaller selection (fig. 4b). This selection is a combination of evidence from the dial position with prior evidence i.e. their last.fm music history. When users fully assert control (max. pressure), they navigate the collection album by album (fig. 4c) and can make exact selections. By varying the pressure, users seamlessly move through this continuous range of control.

The smooth change in engagement is achieved using a simple model of user input. We assume that in an engaged interaction, users will point precisely at the song of interest (as in fig. 4c). For more casual selection, we assume that users will point in the general area (mood) of the music they want, modelled using a normal distribution as in (fig. 4b). As less pressure is applied the distribution is widened, leading to less precise selection and a greater role for a prior belief over the music collection such as listening history.

5. SUMMARY

The scenarios explored here involve casual music retrieval, where users have an ill-defined information need and browse for hedonic purposes or to satisfice a music selection. In these cases, considering what input a user would provide for target songs and inferring selections is an intuitive approach which avoids the issue of defining relevance. We show two music interactions which support the uncertain selection of music, inferred from casual user input such as tapping a rhythm or turning a radio dial.

We have shown that modelling user input for inferring music selection can address issues of subjectivity by taking a user-centered approach to model development. The model can be iterated by comparing its predictions against actual user behaviour. Accounting for this subjectivity can yield significant improvements in retrieval performance as well as creating a more personalised search experience. A key feature of the second system, FineTuner, is its ability to span seamlessly from casual search scenarios, such as satisficing, through to more explicit selections of music. By conditioning the inference upon the user's level of engagement, we are able to interpret the same input space (in this case the dial) according to the current context.

Our approach to casual music interaction empowers the user to enjoy their music while expending as much or as little effort in the retrieval as they wish, providing queries in their own subjective style. Instead of focusing solely on optimising the retrieval process, we consider it equally important to design retrieval systems which suit how the user currently wants to interact. By considering how users might provide casual evidence for their listening intent, we achieve music interactions as simple as tapping a beat or tuning a radio.

6. ACKNOWLEDGMENTS

We are grateful for support from Bang & Olufsen and the Danish Council for Strategic Research.

7. REFERENCES

[1] Boland, D., and Murray-Smith, R. Finding My Beat: Personalised Rhythmic Filtering for Mobile Music Interaction. In *MobileHCI 2013* (2013).

[2] Crossan, A., and Murray-Smith, R. Rhythmic Interaction for Song Filtering on a Mobile Device. *Haptics and Audio Interface Design* (2006), 45–55.

[3] Flemisch, O., Adams, A., Conway, S. R., Goodrich, K. H., Palmer, M. T., and Schutte, P. C. NASA/TMâĂŤ2003-212672 The H-Metaphor as a Guideline for Vehicle Automation and Interaction, 2003.

[4] Ghomi, E., Faure, G., Huot, S., and Chapuis, O. Using rhythmic patterns as an input method. *Proc. CHI* (2012), 1253–1262.

[5] Hanna, P. Query by tapping system based on alignment algorithm. In *Proc. ICASSP* (2009), 1881–1884.

[6] Hopmann, M., Vexo, F., Gutierrez, M., and Thalmann, D. Vintage Radio Interface: Analog Control for Digital Collections. In *CHI 2012: Case Study* (2012).

[7] Jang, J., Lee, H., and Yeh, C.-H. Query by Tapping: A New Paradigm for Content-based Music Retrieval from Acoustic Input. *Proc. PCM* (2001).

[8] Lantz, V., and Murray-Smith, R. Rhythmic interaction with a mobile device. In *Proc. NordiCHI*, ACM (2004), 97–100.

[9] Laplante, A., and Downie, J. S. Everyday life music information-seeking behaviour of young adults, 2006.

[10] Masri, P. *Computer modelling of sound for transformation and synthesis of musical signals.* PhD thesis, University of Bristol, 1996.

[11] Scheibehenne, B., Greifeneder, R., and Todd, P. M. What Moderates the Too-Much-Choice Effect? *Journal of Psychology & Marketing 26(3)* (2009), 229–253.

[12] Stober, S., and Nürnberger, A. Towards user-adaptive structuring and organization of music collections. *Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music* (2010), 53–65.

[13] Trehub, S. E. Human processing predispositions and musical universals. In *The Origins of Music*, N. L. Wallin, B. Merker, and S. Brown, Eds. MIT Press, 2000, ch. 23, 427–448.

[14] Wilson, M. L., and Elsweiler, D. Casual-leisure Searching: the Exploratory Search scenarios that break our current models. In *HCIR 2010* (2010).

[15] Wobbrock, J. O. Tapsongs: tapping rhythm-based passwords on a single binary sensor. In *Proc. UIST* (2009), 93–96.

Search or browse? Casual information access to a cultural heritage collection

Robert Villa, Paul Clough, Mark Hall, Sophie Rutter
Information School
University of Sheffield
Sheffield, UK
S1 4DP
{r.villa, p.d.clough, m.mhall, sarutter1} @sheffield.ac.uk

ABSTRACT
Public access to cultural heritage collections is a challenging and ongoing research issue, not least due to the range of different reasons a user may want to access materials. For example, for a virtual museum website users may vary from professionals or experts, to interested members of the public visiting on a whim. In this paper, we are interested in the latter user: a user who visits a cultural heritage website without a clear goal or information need in mind. In the user study reported here, carried out within the context of the interactive task at CLEF (interactive CHiC), 20 participants explored a subset of Europeana with no explicit task provided using a custom-built interface that offered both search and browse functionalities. Results suggest that browsing is used considerably more by the majority of users when compared to text search (all participants used the category browser before carrying out a text search). This highlights the need for cultural heritage search interfaces to provide browsing functionality in addition to conventional text search if they wish to support casual search tasks.

General Terms
Design, Experimentation, Human Factors.

Keywords
Cultural heritage, virtual museums, information access.

1. INTRODUCTION
Providing public access to cultural heritage is an ongoing and challenging area of research. Previous work suggests that visitors to online cultural heritage collections (e.g. virtual museum visitors) are not necessarily motivated by an explicit task, and that interacting with cultural heritage collections is exploratory in nature [8, 9]. Recent work in the area of 'casual search' [10] has also investigated situations where users are driven by the pleasure of the search process itself, rather than an explicit information need.

The focus for this paper is how individuals explore a cultural heritage collection when given no task. The results may be used both to contrast with studies which have used explicit tasks, and to motivate changes to cultural heritage systems to better support a diverse range of user tasks.

The work reported here is based on initial results from the Interactive CHiC (Cultural Heritage in CLEF) track of CLEF[1] as run at Sheffield University. The interactive CHiC track is based on the CHiC Europeana data set as used in 2011 and 2012 [1]. An early prototype of an evaluation framework was used [2] which allowed the interactive experiment to be semi-automated. In this work, our focus is on how users explored the collection and in particular how search and browse were used in this exploration. We consider three research questions:

RQ1. How do participants initiate their exploration?

RQ2. Do participants use browse or search in their exploration of the collection?

RQ3. How do participants decide to search or browse, when given no explicit task?

With RQ1 we are particularly interested whether users start their exploration by browsing categories, or by search. RQ2 then considers how users access the collection over their whole session. For RQ3 we will present some initial qualitative data from our lab-based interactive study, where the aim is to identify reasons for the use of either the search or browse functions.

2. PREVIOUS WORK
A general review of museum informatics is provided in [3], although the more specific area of museum visitor studies, investigating why and how individuals visit museums, has a long history [4]. More recent work has focused on visitors to digital museums [5-7]. In [6] the information seeking behavior of cultural heritage experts was studied through interviews, finding that complex information gathering was required for the majority of search tasks. In contrast [7] studied virtual museum visitors, inspired by the work of [8] and [9] which suggest that museum visitors are exploratory in their information seeking. This work [7] found that search occurred far more often than browse behavior for three of the four tasks used in the study, the exception being an open and broad task where browsing occurred to a greater degree.

Museum visitors can, in some respects, be considered as examples of "casual leisure" searchers, as outlined in [10], where examples were found of "need-less" browsing (based on a diary study, and analysis of Tweets, both outside the domain of cultural heritage). Darby and Clough [11] investigated the information seeking

[1] http://www.promise-noe.eu/unlocking-culture

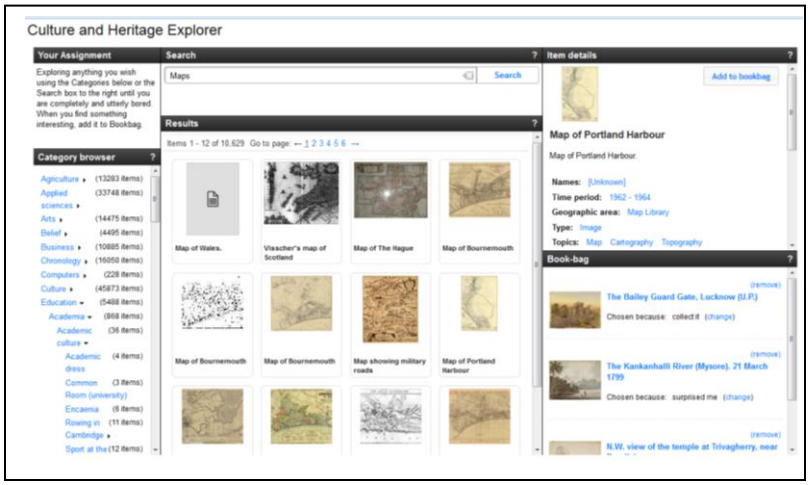


Figure 1: Screenshot of the Interactive CHiC interface

behavior of genealogists, with an emphasis on the behavior of amateurs and hobbyists, rather than professionals. In [12] a review of three digital libraries projects is carried out, from the point of view of Ingwersen and Järvelin's Information Seeking and Retrieval framework [13]. Similar to [10], it points out that information behavior by end users may be the "end in itself".

The study reported here uses a conventional lab-based protocol. However, unlike in previous work, such as [7], the participants were not given an explicit task: the underlying aim being to model a situation closer to that investigated in [10], where there is no explicit information need.

3. INTERACTIVE CHiC

A screenshot of the CHiC interactive system is shown in Figure 1. The interface is split into five main areas, clockwise from left to right: a category browser, search box, item display, bookbag, and search results. The search box operates in the conventional manner, allowing free text queries with search results being displayed as a grid below. When a result is clicked, it is displayed in the "item display" on the right. This information will typically include a small thumbnail, textual description, and the item's associated metadata. Metadata is clickable, e.g. if an item is listed as being owned by the British Library, clicking on the field will search for British Library objects. At the bottom of the item display is a "more like this", which displays the images of up to eight similar objects, which can be viewed three at a time.

On the left of the interface is the "category browser", which allows the user to browse the Europeana collection through a hierarchy of categories. This hierarchy is automatically generated, and is based on the work of [14]. The technique combines the Wikipedia category hierarchy with topics derived from Wikipedia articles into which items are mapped. When a category is clicked, the main results are updated to list the category contents. Small right arrows beside each non-leaf category allows the viewing of sub-categories. The user can therefore search and browse the collection in three main ways: using a text query, selecting a category, or selecting item metadata or "more like this".

On the bottom right of the interface is the bookbag, into which items can be placed. Book-bagged items are kept listed on the display, and can be removed and redisplayed as required. The underlying search system is based on Apache Solr[2], which provides the text search, spelling checker, and the "more like this" suggestions (determined using Solr's standard more-like-this functionality. The data set used was the same as that used in interactive CHiC, a dump of the Europeana data set[3].

4. EXPERIMENTAL SETUP

The search and browse interface was embedded into an IR evaluation system, which automatically administered pre- and post-questionnaires, and displayed the experimental system. All data reported here is from an in-lab study. This allowed a follow up interview to be carried out, during which each participant reviewed his or her search session. To enable this reviewing, Morae screen recording software was used to record the user's activity, and during the interview, an audio recording was made of the user's comments.

An important aspect of the interactive CHiC experimental design was that no explicit task was provided to users. Instead instructions asked the user to explore freely as they wished, until they were bored. Users were informed after they had been active for 10 minutes, and could then continue for a further 5 minutes if they wished, at which point they would be asked to stop (these timings were carry out by hand, and were approximate). Once this was finished, the user's search session would be replayed to them, and an interview conducted to investigate the user's search process. Participants were paid 10 pounds for taking part.

In total 20 participants were recruited for the study, 11 male and 9 female. Eight participants were in the 18-25 year age band, nine in the 26-35 band; the other 3 between 36-45. The majority were students (13), with 5 employed, one unemployed, and one "other". 13 had completed a higher education degree, while six were currently studying an undergraduate degree.

5. RESULTS

5.1 Initiation of exploration

RQ1 asks how users initiate their exploration of the collection. To investigate this, we first looked at how users started their session, and in particular, their searching. For example, did they select a category or enter a query?

Over the whole data set four different actions were used by participants to initiate their session (Table 1, column 2). For the majority of users, the first action was to select one of the categories (15 out of the 20 users). It should be noted that the interface, on startup, showed a set of default results to all users. For three users, the first action was to display one of these default results, another user clicked the "next page" to view the next page of default results, while the final user's first action was to bookmark one of the default result items.

We also investigated the logs to find out each user's first search or browse action, which could be one of category select, text query, or metadata/more like this select. As shown in Table 1 (column 3), for all users this was a category select. In addition to counting the first actions, we also investigated how long each user spent before either clicking the interface, or starting a new

[2] http://lucene.apache.org/solr/

[3] http://www.europeana.eu/

search/browse using the three previously listed methods. These results are shown in Table 2, along with the overall length of time of each session.

Table 1: Number of users whose first action/first search or browse action were as column one.

Action	#Users first action	#Users first search/browse action
Category select	15	20
Display item	3	-
Next search result page	1	-
Add to bookbag	1	-

Table 2: Time to first action, time to first search/browse action, and overall session time (all times in seconds)

	Min	1st Qu.	Median	Mean	3rd Qu	Max
First action	7.00	19.00	25.00	30.50	38.75	90.00
First search/ browse	7.00	22.75	38.00	57.50	81.75	204.0
Total time	129	631.8	783.5	787.8	918.0	1544

There was a considerable variance in the length of time users spent on the task. The median time taken by users was 783.5 seconds (just over 13 minutes), with an interquartile range of 286.2 seconds (approximately 4 minutes, 45 seconds). The minimum time was 129 seconds, and maximum 1544 seconds (over 25 minutes).

Most users spent some time at the start of their session before either clicking on an interface element (median time 25 seconds) or initiating a search (median 38 seconds).

5.2 Search vs. browse

RQ2 asks whether participants use search or browse. Figure 2 presents query and category counts across all users (i.e. counts of how often either text queries were executed or categories selected). Item select and the "more like this" functionality is not included here, due to the relative rarity of these events (across the whole data set this functionality was used only 15 times, by 7 different users).

A non-parametric Wilcoxon rank-sum test indicated that there was a significant difference between queries executed and categories selected ($W = 50.5$, $p \leq 0.001$). As can be seen from the boxplots, categories were selected far more than queries entered, the median number of queries executed being 2, compared to a median of 11 for category selects. All but three users selected more categories than executed queries, and 8 users did not enter a text query at all.

A similar situation exists when the time querying vs. browsing categories is estimated (Figure 3). Such times were estimated by starting a timer when a query or category was selected, and taking all activity between this point and the next query or category select as the user either "querying" or "browsing categories". As might be expected, the trend is similar to that of Figure 2, with users spending more time browsing categories when compared to executing queries. All but five participants spent more time browsing using the categories than spent querying.

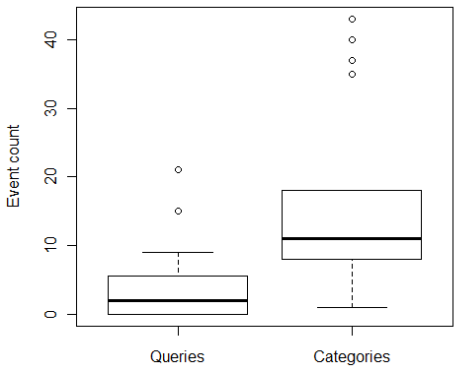


Figure 2: Comparison of query and category select counts

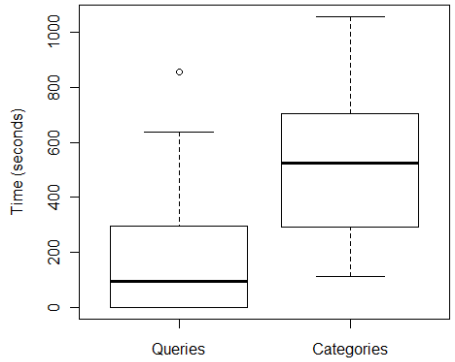


Figure 3: Estimated time querying vs. browsing by category

5.3 "How did you start?"

In addition to the quantitative data above, in the post-session interview two questions were asked of users: "how did you start?" and "Why did you choose to start with a [category/search query]?" It was intended to alter this latter question depending on how the user initiated their exploration. While some users started by examining the results, all users chose the category browser over the search box to initiate searchers.

The responses to the first question "how did you start?" mentioned the category browser explicitly in 8 of the 12 answers. In most of these cases this was linked to exploring the interface. For example, participant P3 stated:

> *"I was drawn to the middle then decided to look around at the interface. I decided to look at categories first, picked politics"*

Similarly, participant P10 stated:

> *"I just looked round to see what I could use to explore things. The category browser looked like the most likely candidates because it had descriptions of stuff."*

As well as being influenced by the interface, responses from some users suggest that prior interests also played a part. For example,:

> *"I just look at the layout of the website and then found that I had a category browser so I went to what I study actually, and I study languages and I try to find something interesting."* [P8]

> *"There is no particular task and so I started from browse to see which information is more interesting to me."* [P1]

The design of the interface, with a relatively small search box, appears to also have had an effect on the choses of at least two of

the user, indicated by responses to the second question. Participants P2 and P4 stated:

> *"Because I only saw that [category]. I didn't see the search until a bit later on."* [P2]

> *"I didn't really see this one at first [the search box] it was a bit obscure."* [P4]

For many users, however, the fact that the category browser allowed easy exploration appeared to be the key, with some users making connections to physical museums. For example:

> *"If I was going to a museum I would look at the categories [museum sections] that are of most interest to me: arts, old stuff and so this is why I was looking for Mona Lisa."* [P5]

The lack of an explicit task was mentioned by some, and search was explicitly commented on by two users. E.g., P7 stated *"When I wanted to find something specific I went to the search box."*

6. DISCUSSION

RQ1 asks how participants initiate their exploration of the collection. From Table 1 it can be seen that all 20 participants started their exploration using the category browser, rather than a text search. Indeed, the first action for the majority of users (75%) was to select a category. Quantitative data from Section 5.3 backs this up, with 8 out of 12 of the participants for which text transcripts are available explicitly mentioning the category browser as a way of starting their exploration. Looking at Table 2, it can be seen that there is typically a short delay until participants started their browsing (median 38 seconds, interquartile range of 59). This delay is consistent with participant's comments which suggested that many first spent some time orienting themselves to the interface before starting (e.g. P10 from Section 5.3).

Moving to RQ2 and RQ3, which asked whether participants have a preference for browse or search and why, it is clear from Figure 2 and Figure 3 that there is a general preference for browsing, e.g. from Figure 3 the median estimated time spent browsing using the categories was 524 seconds (IQR 399), compared to 77 seconds (IQR 394) for text queries. Looking at the participant comments, the lack of any explicit task would appear to have played a part in this preference (e.g. P1 and P5 quotes from Section 5.3). In addition to this the design of the interface, with a relatively small text search box at the top, appeared to also play a part, with some users pointing out that they did not see the search box until later in their session (e.g. P2 and P4).

7. CONCLUSIONS AND FUTURE WORK

The preliminary results reported here would suggest that providing browse functionality to cultural heritage collections is important for users arriving without a specific information need, as may be typical in casual search. For the majority of users, this preference for category browsing continues to hold for the session as a whole, with all but 5 users spending more time browsing than keyword searching. Initial analysis of quantitative interface data backs up the qualitative results, with more of the currently analysed user transcripts explicitly mentioning the category browser. The results presented here are preliminary. Future work will expand on the analysis presented here, both the qualitative and quantitative results. However, these initial results provide evidence of the importance of providing browse functionality to cultural heritage collections, and Europeana in particular.

Acknowledgements: This work was supported by the EU projects PROMISE (no. 258191) and PATHS (no. 270082).

8. REFERENCES

[1] Gäde, M., Ferro, N., and Lestari Paramita, M. 2011. ChiC 2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. *In Petras, V., Forner, P., and Clough, P., editors, CLEF 2011 Labs and Workshops*, Italy.

[2] Hall, M. and Toms E. 2013. Building a Common Framework for IIR Evaluation. Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, 4th International Conference of the CLEF Initiative.

[3] Marty, P. F., Rayward, W. B. and Twidale, M. B. 2003. Museum informatics. Ann. Rev. Info. Sci. Tech., 37, 259–294.

[4] Booth, B. 1998. Understanding the Information Needs of Visitors to Museums, In Museum Management and Curatorship, 17(2).

[5] White, L., Gilliland-Swetland A., and Chandler R. 2004. We're Building It, Will They Use It? The MOAC II Evaluation Project. In Museums and the Web (MW2004), http://www.museumsandtheweb.com/mw2004/papers/g-swetland/g-swetland.html

[6] Amin, A., van Ossenbruggen, J., Hardman, L. and van Nispen, A. 2008. Understanding cultural heritage experts' information seeking needs. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08). ACM, New York, NY, USA, 39-47.

[7] Skov, M. and Ingwersen, P. 2008. Exploring information seeking behaviour in a digital museum context. In Proceedings of the second international symposium on Information interaction in context (IIiX '08), ACM, New York, NY, USA, 110-115.

[8] Black, G. 2005. The engaging museum. London: Routledge.

[9] Treinen, H. 1993. What does the visitor want from a museum? Mass media aspects of museology. In S. Bicknell and G. Farmelo (Eds.), Museum visitor studies in the 90s, London, Science Museum, 86-93.

[10] Wilson, M. L. and Elsweiler, D. 2010. Casual-leisure Searching: the Exploratory Search scenarios that break our current models. In: 4th International Workshop on Human-Computer Interaction and Information Retrieval, Aug 22 2010, New Brunswick, NJ, 28-31.

[11] Darby, P. and Clough, P. 2013 Investigating the information-seeking behaviour of genealogists and family historians. Journal of Information Science February, 39, 73-84.

[12] Richard Butterworth and Veronica Davis Perkins. 2006. Using the information seeking and retrieval framework to analyse non-professional information use. In Proceedings of the 1st international conference on Information interaction in context (IIiX). ACM, New York, NY, USA, 162-168.

[13] Ingwersen, P. and Järvelin, K. 2005. The turn: integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.

[14] Fernando, S., Hall, M.M., Agirre, E., Soroa, A., Clough, P. & Stevenson, M. (2012) Comparing taxonomies for organizing collections of documents, Proceedings of COLING 2012: Technical Papers, 879-894.

Studying Extended Session Histories

Chaoyu Ye
Mixed Reality Lab
University of Nottingham, UK
psxcy1@nottingham.ac.uk

Martin Porcheron
Mixed Reality Lab
University of Nottingham, UK
me@mporcheron.com

Max L. Wilson
Mixed Reality Lab
University of Nottingham, UK
max.wilson@nottingham.ac.uk

ABSTRACT

While there is an increasing amount of interest in evaluating and supporting longer "search sessions", the majority of research has focused on analysing large volumes of logs and dividing sessions according to obvious gaps between entries. Although such approaches have produced interesting insights into some different types of longer sessions, this paper describes the early results of an investigation into sessions as experienced by the searcher. During interviews, participants reviewed their own search histories, presented their views of "sessions", and discussed their actual sessions. We present preliminary findings around a) how users understand sessions, b) how these sessions are characterised and c) how sessions relate to each other temporally.

Categories and Subject Descriptors

H5.2 [**Information interfaces and presentation**]: User Interfaces. - Graphical user interfaces.

Keywords

HCIR, Interactive, Information Retrieval, Sessions

1. INTRODUCTION

Information Retrieval (IR) specialists are becoming increasingly concerned with users who continue to search beyond a few queries or a few minutes[1]. Although Information Retrieval, and even Interactive IR, evaluations are well known, research is recognising situations where people continue to search after finding seemingly useful results [13]. Some might be in a larger session involving several related subtopics, while others may continue to search for entertaining videos until they struggle to find 'good' results [3, 1]. Consequently, researchers are interested in how to evaluate, measure, and ultimately better support searchers who continue to search for extended sessions.

Most research into extended search sessions, described in detail below, has focused on analysing search engine logs [1, 4, 8] by dividing the logs using obvious periods of inactivity and either qualitatively [1] or quantitatively [4, 8] characterising them. Some research has investigated human web behaviour and user goals qualitatively through interviews, however our research has focused on using such methods to better understand real extended search sessions. This paper begins by first summarising literature on sessions and then describes our research methods and preliminary findings about extended search sessions.

[1]The recent NII Shonan event and the forthcoming Dagstuhl are both, for example, focused on this topic.

2. UNDERSTANDING "SESSIONS"

Although investigations into web sessions can be dated back to around 20 years ago (e.g. [2]), the concept of a session still lacks clear definition. A number of researchers have generated diverse definitions of a session using different delimiters such as cutoff time, query context, or even the status of the browser windows (e.g. [7]). In 1995, Catledge and Pitkow used a "timeout", the time between two adjacent activities, to divide user's web activities into sessions and found that a 25.5 minute timeout was best [2]. Their research, however, was focused on general web activity rather than search sessions, but their 25.5 minutes timeout has been used by many others. He and Goker later aimed to find the optimal interval that would divide large sessions, whilst not affecting smaller sessions [4]. Their analysis found that optimal timeout values vary between 10 and 15 minutes.

In 2006, Spink et al [11] defined a session as the entire series of queries submitted by a user during one interaction with a search engine, and one session may consist of single or multiple topics. Their approach focused on topic changes rather than temporal breaks, yet it is perhaps unclear how they determined "one interaction" with a search engine.

A clear definition has also been cited as an important challenge in other research. While focusing on "revisitation" behaviour, Jhaveri and Räihä [6] and Tausher and Greenberg [12] found it challenging to differentiate between in-session revisitation and post-session revisitation, for which a clear detection of session boundaries would be useful.

When focusing on searching, rather than web sessions, some use the concept of a "query session". Nettleton et al defined a query session as at least one query made to a search engine, together with the results which were clicked on and other user behaviours as well [8]. They also evaluated the "session quality" based on the number of clicks, hold time and ranking of selected documents, and they used these measures to help determine the difference between sessions.

To summarise the different approaches used to define sessions, Jansen et al. provided a summary of the three most representative strategies [5], as shown in Table 1. As IP and cookies were utilised to identify a user, the most frequent strategies involve temporal cutoffs and topic change.

The methods summarised in Table 1 are primarily focused on temporal and topical boundaries, but other research has shown clear challenges to these strategies. Mackay et al, in

Table 1: Session Diving Strategies; Jansen et al [5]

Approach	Session Constraints
1	IP, cookie
2	IP, cookie, and temporal cutoff
3	IP, cookie, and content change

2008, examined tasks that frequently occur as multi-session tasks, where something thematically consistent occurs over multiple sessions [7]. Moreover, research into web, browser, and browser-tabs, has found that some users often keep web pages spread out over time, especially in the information gathering tasks, e.g. [10]. These situations indicate that the logged web behaviour may differ significantly from the actual behaviours and intentions of the searchers. This research focuses on the searcher's experience of web sessions, such that others may continue to develop strategies for more accurately dividing large scale logs into sessions.

3. EXPERIMENT DESIGN

To understand and characterise real extended search sessions, we employed similar interview methods to Sellen et al. [10]. Participants were engaged in a 90-120 minute interview about their own search behaviour. To ground the interviews in real data, participants focused on printouts of their own web history, and we used the card sorting technique [9] to probe their mental models of sessions. The procedure was approved by the school ethics board and pilot tested.

Participants began by providing their web history and they were advised to edit their history in advance should they wish to keep some logged activities private[2]. These logs were gathered by importing their search histories to Firefox (if not already there), and creating an XML export using "History Export 0.4"[3]. This log was then structured and preliminarily processed using a) automatic methods to find search URLs, and b) manual investigation to find possible sessions to discuss in the interview. After providing demographic information, participants spent around 20 minutes examining the structured printout of their history, using a pen to mark sessions. These sessions, unless duplicates of prior sessions, were written onto separate cards for later sorting until around 20 cards were produced. Each card had a number, a title, activity purpose, included history items from the history list and also whether it has been completed successfully or not; an example is shown in Figure 1.

The remainder of the interview involved first open, and then closed card sorting. Open card sorting allowed the participants to classify and group the sessions according to their own ideas, whilst closed card sorting allowed us to make sure the following dimensions were considered: purpose, for whom, with whom, location, duration, difficulty, importance, frequency, and priority. This exercise was to help explore the session feature in a more detailed way. For example, studying frequency helps to find out the most frequent sessions and elicit the pattern of user's web activity.

[2]Although this means we have likely missed common search sessions, like the lengthy adult sessions observed by Bailey et al [1], it was considered an important ethical provision.
[3]addons.mozilla.org/en-us/firefox/addon/history-export/

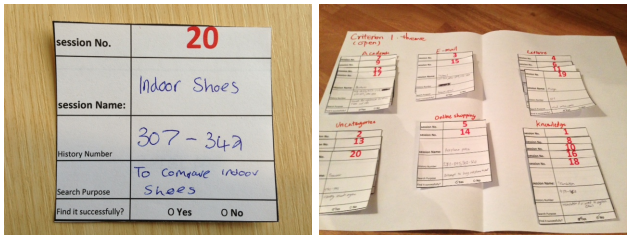


Figure 1: Session Card Information

In addition, the reasons for leading to non-success and difficulty can be investigated via the card sorting of difficulty, and the difference of user's web behaviour in different environments can also be examined by the sorting of location. The entire interview was audio recorded, and physical copies of the card sorts were kept for analysis.

This paper describes our preliminary analysis of the first phase of the study, which involved 11 interviews. Phase two, which is still under way, involves a slightly refined methodology to capture more information about topics that emerged from the initial analysis described below. A more comprehensive analysis of both phases will be published later.

4. PRELIMINARY FINDINGS

Based on our preliminary investigation, some potentially interesting results relating to perceived duration, time of day, and use of queries were found. We considered each of these below according to two aspects: activity goal and activity context. For activity goal, we used Sellen et al's [10] 6 categories: 'finding', 'information gathering', 'browsing', 'transaction', 'communication', and 'housekeeping'. This approach did not include any email, so this was added as a 7th category. For activity context, we applied Elseweiler et al's [3] comparison between work and non-work (leisure) activities, involving: 'work', 'serious-leisure', 'project-leisure', and 'casual-leisure'. At this early stage in the project, the primary author performed the classification individually based on corresponding examples given in the referenced work.

4.1 Defining Sessions

There were 216 sessions in total and 19.6 sessions per person have been studied thus far, as shown as Table 2. Amongst these, 94 were longer than 5 minutes, 99 featured search and only 9 sessions were unsuccessful.

Table 2: All Session Information

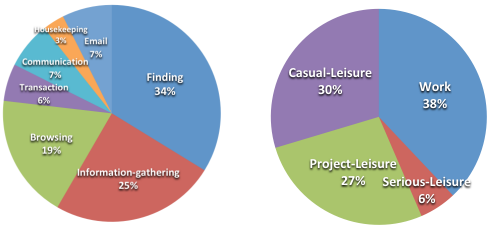
Participant	Session No.	Long Session No.	Unsuccess Session No.	Search Session No.	Query No.
1	18	9	1	13	45
2	30	14	0	11	34
3	20	12	1	12	101
4	20	8	1	9	22
5	16	10	0	6	17
6	26	6	0	16	27
7	30	5	1	0	0
8	17	7	1	12	74
9	10	6	0	6	18
10	10	8	4	4	57
11	19	9	0	10	23
Total	216	94	9	99	418
Avg.	19.6	8.5	0.8	9	38

All participants mentioned that activities with the same purpose and subject should be grouped into one session, as shown in Table 3. In addition, 8 of the 11 suggested that similar tasks happened in different time periods should be classified as a single session, rather than them being tem-

Table 3: Session Delimiters Summary

Participant	Topic	Type of Source	Differ time-> Differ Session	Emotion
1	+	+	-	-
2	+	-	+	-
3	+	-	-	-
4	+	-	-	-
5	+	-	-	-
6	+	-	+	-
7	+	-	-	+
8	+	-	+	-
9	+	-	-	-
10	+	-	-	-
11	+	-	-	-

porally connected. Some participants said that they always kept the browser windows open when doing long-term tasks. Finally, 1 participant advised that they care about the emotion involved within these web activities, even when they were doing the same task, such as "buying a pair shoes". In particular, this participant indicated that one topically consistent session should be divided between two disappointingly unproductive and excitingly productive phases.



(a) Acitivty Goal (b) Activity Context

Figure 2: Session Categories

Finally, besides the pre-defined dimensions, participants also came up with some unique sorting dimensions as shown in Table 4, and these may benefit in exploring the session's delimiters and features in new perspectives.

Table 4: Unique Dimensions

Unique Dimensions	
Google it or Go to Website directly	Content contributor
National	Certain topic or not
University related or not	Based on old knowledge or brand new
Amusement	Preference
Result Satisfaction	Eyes Ears Needed
Security	

4.2 Duration

As duration is one of the targeted dimensions, all participants were asked for their own definition of what constitutes a "long session". 45% of participants defined the session where the duration is more than 5 minutes, whereas 27% went with over 30 minutes, 18% more than 1 hour, and 1 participant chose over 2 hours.

Because participants first defined what they considered to be a long session, and then later sorted their sessions into length categories, we investigated the difference between sessions that met their definition of long, and ones they remembered as being long during the card sorts. Participants frequently grouped 'defined short' sessions as long and vice-versa. Consequently, we investigated both 'overestimated' and 'under-estimated' sessions in addition to 'defined long', 'long', 'actual long', 'defined short', 'short', and 'actual short' as given in Table 5.

Table 5: Duration Categories

Group	Detail
Defined Long	Sessions defined as Long by Participant
Long	Session whose actual duration is >= 5 mins
Actual Long	Session defined as Long and its actual duration is >= 5 mins
Over-estimated	Session defined as Long but its actual duration is less than 5 mins
Defined Short	Session defined as Short by participant
Short	Session whose actual duration is less than 5 mins
Actual Short	Session defined as short and its actual duration is less than 5 mins
Under-estimated	Session defined as Short but its actual duration is >= 5 mins

Table 6: Duration, by Acitivity Goal

	Defined Long	Over-esti	Defined Short	Under-esti
Finding	24	17 (70.8%)	36	3 (8.3%)
Info-gathering	35	15 (42.9%)	7	4 (57.1%)
Browsing	28	17 (60.7%)	5	0
Transaction	4	2 (50.0%)	5	2 (40.0%)
Communication	9	3 (33.3%)	5	0
Housekeeping	0	0	1	1 (100.0%)
Email	7	6 (85.7%)	7	0

Firstly, considering activity goals given in Table 6, the number of 'information-gathering' sessions defined as long was 5 times as that of those 'defined short', as was the same with 'browsing'. On the contrary, the number of 'finding' sessions defined as short was 1.5 times the number defined as long. Overall, nearly 70% of 'finding', 42% of 'information-gathering', 60.7% of 'browsing', 50% of 'transaction', and 85.5% of 'email' sessions defined as long were overestimated by users. Moreover, under-estimation occurred with 'finding', 'information-gathering', and 'housekeeping' although over-estimation was more frequent with 'finding', 'browsing', 'communication', and 'email' sessions.

Table 7: Duration, by Activity Context

	Defined Long	Over-est.	Defined Short	Under-est.
Work	38	22 (57.9%)	31	2 (6.5%)
Serious-Leisure	8	2 (25%)	1	0
Project-Leisure	22	15 (68.2%)	23	5 (21.7%)
Casual-Leisure	39	21 (53.8%)	11	3 (27.2%)

Table 7 above shows that the number of 'casual-leisure' sessions defined as long was as 3 times as that those 'defined short' and that 57.9% of 'work', 68.2% of 'project-leisure', and 53.8% of 'casual-leisure' sessions defined as long were over-estimated by users with lower levels of under-estimation occurring. This encouraged a further study on the feature of each kind of web activity to determine the main cause for an incorrectly perceived length.

4.3 Time of Day

Figure 3 shows that most the 'information-gathering', 'finding' and 'housekeeping' sessions seem to occur between 10:00 and 16:00 whilst more 'browsing', 'email', and 'communication' activities were done between 22:00 and 0:00, which was labelled "before bed time". Additionally, there is a peak around 14:00, in which more 'finding' and 'information-gathering' happened rather than other kinds of sessions. Finally, at 23:00, general 'browsing' is most prevalent.

Figure 4 shows that most of the 'serious-leisure' sessions occurred between 18:00 and 22:00. Most of the 'work' activities happened between 11:00 and 18:00, which seems to fit in within a typical working day. In the time 'before bed',

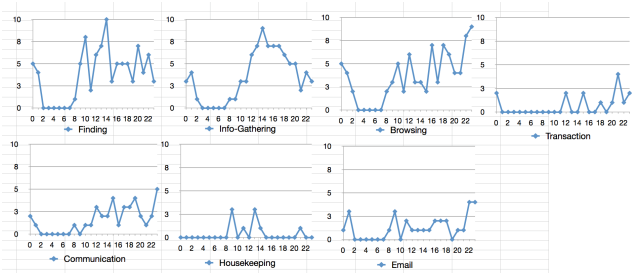


Figure 3: Time of Day, by Activity Goal

the most frequent activity is 'casual-leisure'.

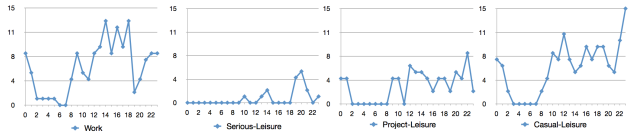


Figure 4: Time of Day, by Activity Context

Combined with the two comparisons above, there seems to be some overlap between 'information-gathering', 'finding', 'housekeeping' and 'work'. There was also some overlap between 'browsing' and 'casual-leisure'. Furthermore, these tend to suggest that there may be some patterns for user's web activity in their daily life.

4.4 Search Queries

In Figure 5 below, sessions with more search queries tend to be classified as 'defined long', 'long', and 'actual long' than those with fewer queries. An interesting observation is that what the user defined as a long session features a relatively low average number of search queries compared with 'long' and 'actual long' sessions. Equally, sessions defined as 'short' by the user actually feature relatively more queries compared to 'short' and 'actual short'. This may indicate that the user did not consider the number of queries performed when defining the duration of sessions and failed to realise the effect of this behaviour.

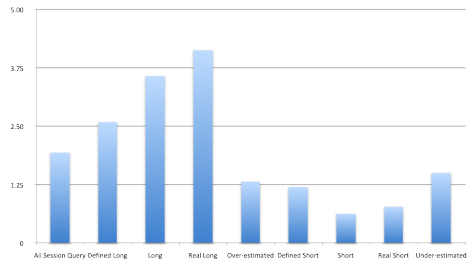


Figure 5: Average Number of Search Queries

5. CONCLUSIONS

Although this paper only describes a preliminary analysis of over 200 sessions from 11 participants, we have begun to see some potentially interesting early findings. Initially, participants varied greatly in their opinions about their own sessions, with some matching topical divisions, some temporal divisions, and some a combination of the two. The majority of participants judged "long sessions" as being longer than 5 minutes, but many had inaccurate recollections of the length of sessions. Long sessions were typically a mix of casual and serious leisure that often involved information gathering and browsing behaviour, while the majority of work related sessions were typically short. We also noticed that some of these activities may also be related to certain times of the day. All of the findings will be further explored after phase two of the study, but early insights suggest that real extended search sessions could be more accurately modelled based on additional factors such as: time of day, activity goal, activity context, and number of queries.

6. REFERENCES

[1] P. Bailey, L. Chen, S. Grosenick, L. Jiang, Y. Li, P. Reinholdtsen, C. Salada, H. Wang, and S. Wong. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems*, Kanagawa, Japan, October 2012.

[2] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[3] D. Elsweiler, M. L. Wilson, and B. K. Lunn. Understanding casual-leisure information behaviour. In A. Spink and J. Heinstrom, editors, *Library and Information Science*, pages 211–241. Emerald Group Publishing Limited, 2011.

[4] D. He and A. Göker. Detecting session boundaries from Web user logs. *Methodology*, pages 57–66, 2000.

[5] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on Web search engines. *JASIST*, 58(6):862–871, 2007.

[6] N. Jhaveri and K.-J. Räihä. The advantages of a cross-session web workspace. In *CHI2005 Ext. Abstracts*, page 1949. ACM Press, 2005.

[7] B. Mackay and C. Watters. Exploring Multi-session Web Tasks. *Time*, pages 1187–1196, 2008.

[8] D. Nettleton, L. Calderon-benavides, and R. Baeza-yates. Baezayates, analysis of web search engine query sessions. In *Proc. WebKDD 2006*, 2006.

[9] G. Rugg and P. McGeorge. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93, 1997.

[10] A. J. Sellen, R. Murphy, and K. L. Shaw. How knowledge workers use the web. In *Proc. CHI2002*, pages 227–234. ACM Press.

[11] A. Spink, M. Park, B. J. Jansen, and J. Pedersen. Multitasking during Web search sessions. *IP&M*, 42(1):264–275, 2006.

[12] L. Tauscher and S. Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *IJHCS*, 47(1):97–137, 1997.

[13] E. G. Toms, R. Villa, and L. McCay-Peet. How is a search system used in work task completion? *Journal of Information Science*, 39(1):15–25, 2013.

Comparative Study of Search Engine Result Visualisation: Ranked Lists Versus Graphs

Casper Petersen
Dept. of Computer Science
University of Copenhagen
cazz@diku.dk

Christina Lioma
Dept. of Computer Science
University of Copenhagen
c.lioma@diku.dk

Jakob Grue Simonsen
Dept. of Computer Science
University of Copenhagen
simonsen@diku.dk

ABSTRACT

Typically search engine results (SERs) are presented in a *ranked list* of decreasing estimated relevance to user queries. While familiar to users, ranked lists do not show inherent connections between SERs, e.g. whether SERs are hyperlinked or authored by the same source. Such potentially useful connections between SERs can be displayed as *graphs*. We present a preliminary comparative study of ranked lists vs graph visualisations of SERs. Experiments with TREC web search data and a small user study of 10 participants show that ranked lists result in more precise and also faster search sessions than graph visualisations.

Categories and Subject Descriptors

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

Keywords

Search Engine Result Visualization, Ranked List, Graph

1. INTRODUCTION

Typically search engine results (SERs) are presented in a ranked list of decreasing estimated relevance to user queries. Drawbacks of ranked lists include showing only a limited view of the information space, not showing how similar the retrieved documents are and/or how the retrieved documents relate to each other [4, 6]. Such potentially useful information could be displayed to users in the form of *SER graphs*; these could present at a glance an overview of clusters or isolated documents among the SERs, features not typically integrated into ranked lists. For instance, directed/undirected and weighted/unweighted graphs could be used to display the direction, causality and strength of various relations among SERs. Various graph properties (see [7]), such as the average path length, clustering coefficient or degree, could be also displayed, reflecting potentially useful or interesting features about how the retrieved data is connected.

We present a user study comparing ranked list vs graph-based SER visualisation interfaces. We use a web crawl of ca. 50 million documents in English with associated hyperlink information and 10 participants. We find that ranked lists result in overall more accurate and faster searches than graph displays, but that the latter result in slightly higher recall. We also find overall higher inter-rater agreement about SER relevance when using ranked lists instead of graphs.

2. MOTIVATION

While traditional IR systems successfully support known-item search [5], what should users do if they want to locate something from a domain where they have a general interest but no specific knowledge [8]? Such exploratory searching comprises a mixture of serendipity, learning, and investigation and is not supported by contemporary IR systems [5], prompting users to "develop coping strategies which involve [...] the submission of multiple queries and the interactive exploration of the retrieved document space, selectively following links and passively obtaining cues about where their next steps lie" [9]. A step towards exploratory search, which motivates this work, is to make explicit the hyper-linked structure of the ordered list used by e.g. Google and Yahoo. Investigation of such a representation does not exist according to our knowledge, but is comparable to Google's Knowledge Graph whose aim is to guide users to other relevant information from an initial selection.

3. PREVIOUS WORK

Earlier work on graph-based SER displays includes Beale et al.'s (1997) visualisation of sequences of queries and their respective SERs, as well as the work of Shneiderman & Aris (2006) on modelling semantic search aspects as networks (both overviewed in [10]). Treharne et al. (2009) present a critique of ranked list displays side by side a range of other types of visualisation, including not only graphs, but also cartesian, categorical, spring and set-based displays [6]. This comparison is analytical rather than empirical. Closest to ours is the work of Donaldson et al. (2008), who experimentally compare ranked lists to graph-based displays [2]. In their work, graphs model social web information, such as user tags and ratings, in order to facilitate contextualising social media for exploratory web search. They find that users seem to prefer a hybrid interface that combines ranked lists with graph displays. Finally, the hyperlinked graph representation discussed in the paper allows users to

SIGIR 2013 Dublin, Ireland
.

investigate the result space thereby discovering related and potential relevant information that might otherwise be bypassed. Such representation and comparison to a traditional ranked list does not exist according to our knowledge, but the idea underpinning the graph representation is comparable with Google's Knowledge Graph as the aim is to guide users to other relevant information from an initial selection.

4. INTERFACE DESIGN

This section presents the two different SER visualisations used in our study. Our goal is to study the effect of displaying exactly the same information to the user in two different ways, using *ranked list* and *graph* visualisations, respectively.

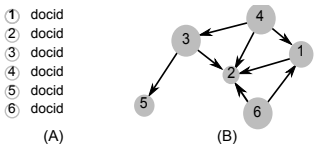


Figure 1: Ranked list (A) and graph (B) representation of the top-k documents from a query.

4.1 Ranked List (RL) Display

We use a standard ranked list SER display, where documents are presented in decreasing order of their estimated relevance to the user query. The list initially displays only the top-k retrieved document ids (docids) with their associated rank (see Figure 1 (A)). When clicked upon, each document expands to two mini windows, overlaid to the left and right of the list:

- The left window shows a document snippet containing the query terms. The snippet provides a brief summary of the document contents that relate to the query in order to aid the user to assess document relevance prior to viewing the whole document [4]. We describe exactly what the snippet shows and how it is extracted in Section 4.3.
- The right window shows a graph of the top-k ranked SERs (see Section 4.2). The position of the clicked document in the graph is clearly indicated, so users can quickly overview its connections, if any, to other top-k retrieved documents.

Previously visited documents in the list are colour-marked.

4.2 Graph (GR) Display

We display a SER graph $G = (V, E)$ as a directed graph whose vertices $v \in V$ correspond to the top-k retrieved documents, and edges $e \in E$ correspond to links (hyperlinks in our case of web documents) between two vertices. Each vertex is shown as a shaded circle that displays the rank of its associated document in the middle, see Figure 1 (B). The size of each vertex is scaled according to its out-degree, so that larger vertex size indicates more outlinks to the other top-k documents. Edge direction points towards the outlinked document. Previously visited documents are colour-marked.

When clicked upon, each vertex expands to two mini windows, overlaid to the left and right of the graph:

- The left window shows the same document snippet as in the RL display.
- The right window shows the ranked list of the top-k SERs. The position of the clicked document in the list is clearly marked.

We display the SER graph in a standard force-directed layout [1]. Our graph layout does not allow for other types of interaction with the graph apart from clicking on it. We reason that for the simple web search tasks we consider, layouts allowing further interaction may be confusing or time-consuming, and that they may be more suited to other search tasks, involving for instance decision making, navigation and exploration of large information spaces.

4.3 Document Snippets

Both the RL and GR interfaces include short query-based summaries of the top-k SERs (*snippets*). We construct them as follows: We extract from each document a window of $\pm$ 25 terms surrounding the query terms on either side. Let a query consist of 3 terms q_1, q_2, q_3. We extract snippets for all ordered but not necessarily contiguous sequences of query terms: $(q1, q2, q3)$, (q_1, q_2), (q_1, q_3), (q_2, q_3), (q_1), (q_2), (q_3). This way, we match all snippets containing query terms in the order they appear in the query (not as a bag of words), but we also allow other terms to occur in between query terms, for instance common modifiers.

Several snippets can be extracted per document, but only the snippet with the highest TF-IDF score is displayed to the user. The TF-IDF of each window is calculated as a normalised sum of the TF-IDF weights for each term:

$$S_{s(D)} = \frac{1}{|w|} \sum_{t=0}^{|w|} tf(t, D) \times \log\left(\frac{|C|}{|D \in C \ : \ t \in D|}\right)$$

where $|w|$ is the number of terms in the window extracted, $t \in w$ is a term in the window, tf is the term frequency of t in document D from which the snippet is extracted, C is the collection of documents, and $S_{s(D)}$ is the snippet score for document D. Finally, as research has shown that query term highlighting can be a useful feature for search interfaces [4], we highlight all occurrences of query terms in the snippet.

5. EVALUATION

We recruited 2 participants for a pilot study to calibrate the user interfaces; the results from the pilot study were not subsequently used. For the main study, we recruited 10 new participants (9 males, 1 female; average age: 33.05, all with a background in Computer Science) using convenience sampling. Each participant was introduced to the two interfaces. Their task was to find and mark as many relevant documents as possible per query using either interface. For each new query, the SERs could be shown in either interface. Each experiment lasted 30 minutes.

Participants did not submit their own queries. The queries were taken from the TREC Web tracks of 2009-2012 (200 queries in total). This choice allowed us to provide very fast response times to participants ($<$ two seconds, depending on disk speed), because search results and their associated graphs were pre-computed and cached. Alternatively, running new queries and plotting their SER graphs on the fly would result in notably slower response times that would risk dissatisfying participants. However, a drawback in using TREC queries is that participants did not necessarily have enough context to fully understand the underlying information needs and correctly assess document relevance.

	Ranked List	Graph
MAP@20	0.4195	0.3211
MRR	0.4698	0.3948
RECALL@20	0.0067	0.0069

Table 1: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) & Recall of the top 20 results.

To counter this, we allowed participants to skip queries they were not comfortable with. To avoid bias, skipping a query was allowed after query terms were displayed, but before the SERs were displayed.

We retrieved documents from the TREC ClueWeb09 cat. B dataset (ca. 50 million documents crawled from the web in 2009), using Indri, version 5.2. The experiments were carried out on a 14 inch monitor with a resolution of 1400 x 1050 pixels. We logged which SERs participants marked relevant, as well as the participants' click order and time spent per SER.

5.1 Findings

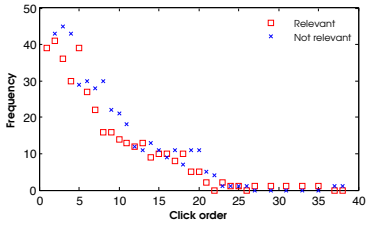
In total the 10 participants processed 162 queries (89 queries with the RL interface and 73 with the GR interface) with mean $\mu = 16.2$, and standard deviation $\sigma = 7.8$. Four queries (two from each interface) were bypassed (2.5% of all processed queries).

Table 1 shows retrieval effectiveness per interface, aggregated over all queries for the top $k = 20$ SERs. The ranked list is associated with higher, hence better scores than the graph display for MAP and MRR. MAP is +30.6% better with ranked lists that with graph displays, meaning that overall a higher amount of relevant SERs is found by the participants at higher ranks in the ranked list as opposed to the graph display. This finding is in agreement with the MRR scores, which indicate that the first SER to be assessed relevant is likely to occur around rank position 2.13 ($1/2.13 = 0.469 \approx 0.4698$) with ranked lists, but around rank position 2.55 ($1/2.55 = 0.392 \approx 0.3948$) with graph displays. Conversely, recall is slightly higher with graph displays. In general, higher recall in this case would indicate that participants are more likely to find a slightly larger amount of relevant documents when seeing them as a graph of their hyperlinks. However, the difference in recall between ranked lists and graphs is very small and can hardly be seen as a reliable indication.

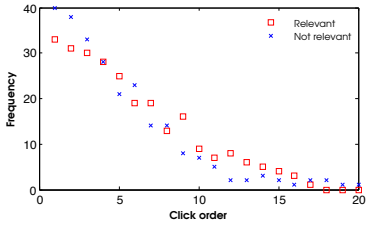
5.1.1 Click-order

On average participants clicked on 9.46 entries per query in the ranked list (842 clicks for 89 queries) but only on 6.7 entries per query in the graph display (490 clicks for 73 queries). The lower number of clicks in the latter case could be due to the extra time it might have taken participants to understand or navigate the graph. This lower number of clicks also agrees with the lower MAP scores presented above (if fewer entries were clicked, fewer SERs were assessed, hence fewer relevant documents were found in the top ranks).

Figures 2a and 2b plot the order of clicks for the ranked list and graph interfaces respectively on the x-axis, against the frequency of clicks on the y-axis. We see that in the ranked list, the first click of the participant is more often on a relevant document, but in the graph display, the first click is more often on a non-relevant document (as already indicated by the MRR scores shown above). We also see



(a)



(b)

Figure 2: Click-order and participant relevance assessments for the (a) ranked list interface and (b) graph interface

Interface	Min	Max	μ	σ
Ranked List	1.391	25.476	8.228	4.371
Graph	3.322	20.963	9.705	3.699

Table 2: Time (seconds) spent on each interface.

that for the graph display, the majority of participant clicks before the 5th click correspond to non-relevant documents. Even though the MRR scores of the graph display indicate that the first relevant document occurs around rank position 2.5, we see that participants on average click four other documents before clicking the relevant document at rank position 2.5. This indicates that in the graph display, participants click documents not necessarily according to their rank position (indicated in the centre of each vertex), but rather according to their graph layout or connectivity.

5.1.2 Time spent

Table 2 shows statistics about the time participants spent on each interface. Overall participants spent less time on the ranked list than on the graph display. This observation, combined with the retrieval effectiveness measures shown in Table 1, indicates that participants conducted overall slightly more precise and faster searches using the ranked lists than using graph displays. The time use also suggests that participants are used to standard ranked list interfaces, a type of conditioning not easy to control experimentally.

5.1.3 Inter-participant agreement

To investigate how consistent participants were in their assessments, we report the inter-rater agreement using Krippendorff's α [3]. Table 3 reports the agreement between the participants, and Table 4 reports the agreements between participants and the TREC preannotated relevance assessments per interface. In both cases, only queries annotated more than once by different participants are included (19 queries for the ranked list and 11 for the graph SER).

The average inter-rater agreements between participants vary considerably. For the graph interface, $\alpha = 0.04471$, which suggests lack of agreement between raters. On a query

basis, some queries (query 169 and 44) suggest a comparatively much higher agreement whereas others (e.g. query 104 and 184) show a comparatively higher level of disagreement. For the ranked list, inter-rater agreement is higher ($\alpha = 0.19813$). On a per query basis, quite remarkably, query 92 had a perfect agreement between raters, while queries 175 and 129 also exhibited a moderate to high level of agreement. However, most queries show only a low to moderate level of agreement or disagreement.

Overall, the lack of agreement may indicate the participants' confusion in assessing the relevance of SERs to pretyped queries. This may be aggravated by problems in rendering the HTML snippets into text. Some HTML documents were ill-formed, hence their snippets sometimes included HTML tags or other not always coherent text.

Inter-rater agreements between our participants and the TREC preannotated relevance assessments show an almost complete lack of agreement. For both interfaces there is a weak level of disagreement on average ($\alpha = -0.0750$ and $\alpha = -0.0721$ for the graph and ranked list respectively). On a per query basis there are only two queries (queries 169 & 110) exhibiting a moderate level of agreement. For most remaining queries our participants' assessments disagree with the TREC assessments.

Graph			Ranked list		
Query	Raters	α	Query	Raters	α
101	4	0.28696	110	3	0.41000
104	2	-0.21875	119	2	0.00000
132	2	-0.16071	120	2	0.49351
169	2	0.48000	129	2	0.86022
180	2	-0.10031	132	3	-0.08949
184	2	-0.25806	133	2	0.30108
3	2	0.00000	155	2	-0.02632
38	2	-0.07519	175	2	0.49351
44	2	0.49351	180	2	-0.37879
58	2	0.00000	51	2	0.00000
–	–	–	53	2	0.02151
–	–	–	74	2	-0.14706
–	–	–	80	2	0.14420
–	–	–	81	3	-0.12919
–	–	–	92	2	1.00000
–	–	–	95	2	0.15584
–	–	–	96	2	0.15584
–	–	–	97	2	0.30179
Average α: 0.04471			Average α: 0.19813		

Table 3: Inter-rater agreement (α) for queries assessed by >1 participant. *Query* is the TREC id of each query.

6. CONCLUSIONS

In a small user study, we compared ranked list versus graph-based search engine result (SER) visualisation. Our motivation was to conduct a preliminary experimental comparison of the two for the domain of web search, where document hyperlinks were used to display them as graphs. We found that overall more accurate and faster searches were done using ranked lists and that inter-user agreement was overall higher with ranked lists than with graph displays. Limitations of this study include: (1) using fixed TREC queries, instead of allowing users to submit their own queries on the fly; (2) having technical HTML to text rendering problems, resulting in sometimes incoherent document snippets; (3) using only 10 users exclusively from Computer Science, which makes for an overall small and rather biased user sample; (4) not using the wider context of the search session in the analysis (e.g. user task, behaviour, satisfaction). Future work includes addressing the above limitations and also testing whether and to what extent these results apply when scaling up to wall-sized displays with significantly larger screen real estate.

Graph			Ranked List		
Query	Raters	α	Query	Raters	α
101	4	0.09559	110	3	0.38654
104	2	-0.17861	119	2	-0.22370
132	2	0.06561	120	2	0.03146
169	2	0.33625	129	2	0.05600
180	2	-0.08949	132	3	0.01689
184	2	-0.08949	133	2	0.04398
3	2	-0.37209	155	2	-0.21067
38	2	-0.05006	165	2	-0.25532
44	2	-0.05861	175	2	-0.07886
54	2	-0.25532	180	2	-0.17861
58	2	-0.22917	51	2	-0.05006
–	–	–	53	2	-0.24694
–	–	–	74	2	-0.06033
–	–	–	80	2	-0.24694
–	–	–	81	3	-0.13634
–	–	–	92	2	-0.21181
–	–	–	95	2	0.04582
–	–	–	96	2	-0.12919
–	–	–	97	2	0.07813
Average α: -0.0750			Average α: -0.0721		

Table 4: Inter-rater agreement (α) between participants and TREC assessments for queries assessed by > 1 participant.

7. REFERENCES

[1] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[2] J. J. Donaldson, M. Conover, B. Markines, H. Roinestad, and F. Menczer. Visualizing social links in exploratory search. In *HT '08*, pages 213–218, New York, NY, USA, 2008. ACM.

[3] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

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

[5] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[6] K. Treharne and D. M. W. Powers. Search engine result visualisation: Challenges and opportunities. In *Information Visualisation*, pages 633–638, 2009.

[7] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Structural analysis in the social sciences. Cambridge University Press, 1994.

[8] R. W. White, B. Kules, S. M. Drucker, and M. Schraefel. Supporting exploratory search. *Communications of the ACM*, 49(4):36–39, 2006.

[9] R. W. White, G. Muresan, and G. Marchionini. Workshop on evaluating exploratory search systems. *SIGIR Forum*, 40(2):52–60, 2006.

[10] M. L. Wilson, B. Kules, B. Shneiderman, et al. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.

Evolving Search User Interfaces

Tatiana Gossen, Marcus Nitsche, Andreas Nürnberger
Data & Knowledge Engineering Group, Faculty of Computer Science
Otto von Guericke University Magdeburg, Germany
http://www.dke.ovgu.de/

ABSTRACT

When designing search user interfaces (SUIs), there is a need to target specific user groups. The cognitive abilities, fine motor skills, emotional maturity and knowledge of a sixty years old man, a fourteen years old teenager and a seven years old child differ strongly. These abilities influence the decisions made in the user interface (UI) design process of SUIs. Therefore, SUIs are usually designed and optimized for a certain user group. However, especially for young and elderly users, the design requirements change rapidly due to fast changes in users' abilities, so that a flexible modification of the SUI is needed. In this positional paper we introduce the concept of an *evolving search user interface* (ESUI). It adapts the UI dynamically based on the derived capabilities of the user interacting with it. We elaborate on user characteristics that change over time and discuss how each of them can influence the SUI design using an example of a girl growing from six to fourteen. We discuss the ways to detect current user characteristics. We also support our idea of an ESUI with a user study and present its first results.

Keywords

Search User Interface, Human Computer Interaction, Adaptivity, Context Support, Information Retrieval.

Categories and Subject Descriptors

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

General Terms

Design, Human Factors.

1. INTRODUCTION

Search user interfaces [8] are an integral part of our lives. Most common known SUIs come in the form of web search engines with an audience of hundreds of millions of people[1] all over the world. This is a very wide and heterogeneous target group with different backgrounds, knowledge, experience, etc. Therefore, researchers suggest providing a customized solution to cover the needs of individual users (e.g., [6]). Nowadays, solutions in personalisation and adaptation of backend algorithms, i.e. query adaptation, adaptive retrieval, adaptive result composition and presentation, have been proposed in order to support the search of an individual user [13, 14]. But the front end, i.e. the SUI, is usually designed and optimized for a certain user group and does not support many mechanisms for personalisation. Common search engines allow the personalisation of a SUI in a limited way: Users can choose a colour scheme or change the settings of the browser to influence some parameters like font size. Some search engines also detect the type of device the user is currently using – e.g. a desktop computer or a mobile phone – and present an adequate UI.

Current research concentrates on designing SUIs for specific user groups, e.g. for children [4, 6, 10] or elderly people [1, 2]. These SUIs are optimized and adapted to general user group characteristics. However, especially young and elderly users undergo fast changes in cognitive, fine motor and other abilities. Thus, design requirements change rapidly as well and a flexible modification of the SUI is needed. Therefore, we suggest to provide users with an *evolving search user interface (ESUI)* that adapts to individual user's characteristics and allows for changes not only in properties (e.g., colour) of UI elements but also influences the UI elements themselves and their positioning. Some UI elements are continuously adaptable (e.g. font size, button size, space required for UI elements), whereas others are only discretely adaptable (e.g. type of results visualization). Not only SUI properties, but also the complexity of search results is continuously adaptable and can be used as a personalisation mechanism for users of all age groups.

2. ESUI VISION

In this section we share our vision of an ESUI. In general, we suggest to use a mapping function and adapt the SUI using it, instead of building a SUI for a specific user group. Using a generic model of an adaptive system, as discussed in [14], we depict the model of an ESUI as following (see Fig. 1). We have a set of user characteristics (or skills) on one side. In the ideal case, the system detects the skills automatically, e.g. based on user's interaction with the information retrieval system (user's queries, selected results, etc.). On the other side, there is a set of options to adapt the SUI, e.g. using different UI elements for querying or visualisation of results. In between, an adaptation component contains a set of logic rules to map the user' skills to the specific UI elements of the ESUI.

[1] Google, for example, has over 170 million unique visitors per month, only in the U.S. `http://www.nielsen.com/us/en/newswire/2013/january-2013--top-u-s\--entertainment-sites-and-we-brands.html`

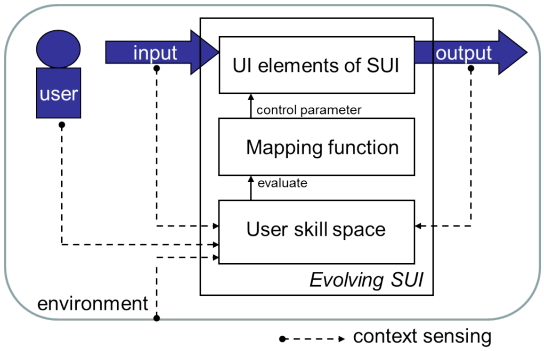


Figure 1: Model of an ESUI.

2.1 Mapping Function

The function between the user skill space and the options to adapt the UI elements of the SUI has to be found. We suggest using the knowledge about human development, e.g. from medical, cognitive, psychosocial science fields to specify the user skill space. The results of user studies about users' search behaviour and SUI design preferences can provide recommendations for UI elements. As far as the research provides information about the studied age group, we can use the age group as a connector between the skill space and the UI elements. Note that we use age groups in the sense of a more abstract category defining a set of specific capabilities while growing up. A lot of research is already done and can be used, e.g. [2, 4, 7]. In addition, if the set of adaptable UI elements is defined, we can evaluate the mapping function by letting users from different age groups put the UI elements of a SUI together (similar to the end user programming).

2.2 Evolving Skills

In order to allow a SUI to evolve together with a user we first have to determine those characteristics that vary from user to user and change during his life (or due to some circumstances like diseases). For example, discussion about the skills of young users is given in [7]. We suggest to consider cognitive skills, information processing rates, fine motor skills, different kinds of perception, knowledge base, emotional state, reading and writing skills.

In the following, brief summary of current research results in human development science is given. Human cognitive development occurs in a sequential order in which later knowledge, abilities and skills build upon the previously acquired ones [12]. *Cognitive abilities* of users in those stages differ, for example, before the last (formal operational) stage they are unable to think logically and to understand abstract concepts. Again, not only age but also some diseases or accelerated cognitive development cause that cognitive abilities, i.e. skills to gain, use and retain knowledge, differ from user to user. *Information processing* capabilities change during life. Children's information processing is slower than that of adults [11]. Therefore, children have a limited cognitive recall. It is widely agreed that elderly people have a decline in intellectual skills which affects the aggregation of new information [15]. *Fine motor skills* are influenced by information processing rates [9]. Therefore, young children's performance in pointing movements, e.g. using a mouse, are lower than that of adults. *Perception* of color can also change while aging. Color discrimination is more difficult for elderly people. Elderly people have also problems with hearing [3]. Children are immature in the *emotional domain* and, especially at the age of six to twelve, require additional emotional support and a resulting feeling of success [5]. Therefore, they require support to increase their confidence. In general, *reading and writing skills* of adults are better than those of children. *Knowledge* is gathered during life. Thus, elderly people posses a larger knowledge base than adults, and adults have usually more knowledge than children. We believe that the discussed characteristics can affect the design of SUIs. However, further research should be done in this direction.

2.3 Detection of User Abilities

An ESUI can provide a specific SUI for a specific user given the knowledge of his specific abilities. A simple case is an adaptable SUI, where a user manually adjusts the search user interface to his personal needs and tasks. An adaptable SUI may also provide several standard settings for a specific user selection to explore the options (e.g. young user, adult user, elderly user). More interesting and challenging is the case of an adaptive SUI, where a system automatically detects the abilities of a user and provides him with an appropriate SUI. Concepts for an automatic detection of user's abilities have been studied in the past. We can use the age of a registered and logged-in user. However, the age provides only an approximation of a user's capabilities. For an individual user an appropriate mapping to the age group has to be found, e.g. using psychological tests covered in form of games. Those games can be used to derive the quality of user's fine-motor skills as well. Furthermore, we can use the user history from log files, in specific, issued queries (their topic and specific spelling errors) and accessed documents. *However, research is required to determine how to adapt a SUI in the way users would accept the changes.*

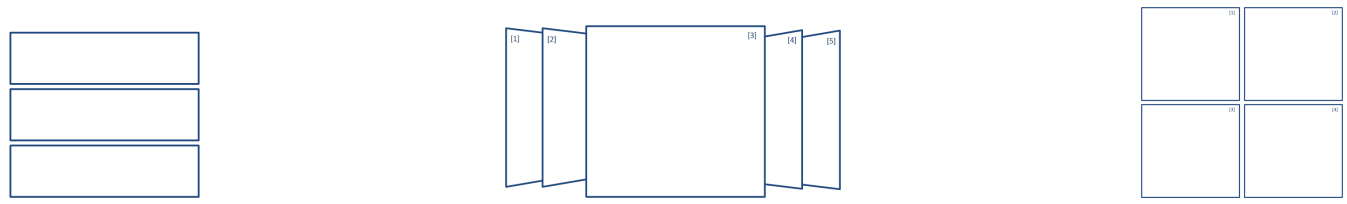
3. DESIGN IDEAS

When designing an ESUI, we first have to define the components of a SUI that should be adapted. We consider three main components. The first component is an *input*, i.e. UI elements which allow a user to transform his information need into a machine understandable format. This component is traditionally represented by an input field and a search button. Other variants are a menu with different categories or voice input. The second component is an *output* of an information retrieval (IR) system. The output consists of UI elements that provide an overview of retrieved search results. There can be different kinds of output, e.g. a vertical list of snippets (Fig. 2a), tiles (Fig. 2c) or coverflow (Fig. 2b). The third is a *management* component. Management covers UI elements that support users in information processing and retaining. Examples of management UI elements are bookmark management components or other history mechanisms like breadcrumbs. Historically, management UI elements are not part of an SUI. But recent research [6] shows that users are highly motivated to use elements of management. Besides these main components, there also exist general properties of UI elements that might affect all the three categories, e.g. font size or color. *We propose to adapt these three main components of a SUI and its general UI properties to the user's skills.*

3.1 Use Cases

In order to demonstrate the proposed ESUI, we consider a young girl called Jenny who is growing older. We show how input and output of a SUI can be adapted to changes of Jenny's abilities.

Use Case 1: Jenny is six years old. She started to learn reading, but she has difficulties with writing. Jenny's active vocabulary is limited to 5,000 words. She cannot yet think in abstract categories and is not able to process much information. Due to her limited writing abilities, Jenny is not able to use an input field and write a query. She is learning to read, therefore, she can use a menu



(a) (b) (c)

Figure 2: Different kinds of output of an information retrieval system: a) *vertical list of snippets* offers a fast overview of several results at once b) *coverflow view of results* offers an attractive animation by browsing, uses a familiar book metaphor, central element is clear separated from the rest c) *tiles of search results* offer a fast overview of several results at once, a user has small jumps by reading within results, however the ordering of results is not so clear as by a list.

(a) (b) (c)

Figure 3: Different kinds of input of an information retrieval system: a) an ESUI enables a six-year-old Jenny to draw her query b) an ESUI supports nine-year-old Jenny by voice input and through several pre-defined categories c) an ESUI enables fourteen-year-old Jenny to use keyword-based input supported by an adaptive query cloud.

with different categories which are supported by images. In order to search for any information Jenny can *draw* her query (Fig. 3a). Jenny's fine motor skills are not fully developed yet. She has difficulties using interactions like scrolling. She also cannot process much information at once. Therefore, the coverflow (Fig. 2b) result visualisation fits her abilities (best). Coverflow allows her to concentrate on one item at a time, thus, her cognitive load is reduced. Jenny can interact with it using simple point-and-click interactions. An integrated text-to-speech reader supports Jenny by reading the results to her.

Use Case 2: Jenny is nine years old. Jenny can read and write short stories with just a few spelling errors. Jenny has some difficulties with typing using a keyboard. She "hunts and pecks" on the keyboard for correct keys. This increases the amount of spelling errors and also slows down the process. Jenny is frustrated because the system does not understand her well. Thus, a standard keyword input field does not fit Jenny's abilities well. Jenny still cannot think in abstract categories and process a lot of information. But her language skills improved and her vocabulary size is increased. Therefore, she can use *voice input* to search for information. A menu with different categories in addition to voice input can inspire Jenny to search for some new information. However, these categories should match her cognitive abilities (Fig. 3b). Jenny can already manage different interaction techniques and is able to process more information than the six-year-old Jenny. Therefore, a list of snippets (Fig. 2a) is an adequate output visualization. It requires not that much cognitive recall as tiles, but allows to process more results items at a time than coverflow does.

Use Case 3: Jenny is 14 years old. Jenny's writing skills are further developed with use of correct grammar, punctuation and spelling. She learns to think logically about abstract concepts. Her vocabulary size is about 20,000 words. She chats a lot with her friends which results in fast typing skills using a keyboard. Therefore, Jenny is able to use a keyword-oriented input search supported by spelling correction and suggestion mechanisms. A SUI can still support Jenny by finding the "right" keywords, for example using a *query cloud*[2] (Fig. 3b). Jenny can already manage different interaction techniques and is able to process more information than the nine-year-old Jenny. Therefore, coverflow and a vertical list visualisation would probably restrain her performance, whereas tiles (Fig. 2c) allow Jenny a better overview of results.

4. USER STUDY

In order to demonstrate the idea of an ESUI, we conducted a user study to compare users' preferences in the visualization of different UI elements of a SUI. In specific, our hypothesis was that *users from different age groups would prefer to use different UI elements and different general UI properties*. We built a SUI that can be personalized, i.e. users can choose input, output and tune general UI properties. In this paper we present our first results, i.e. users' preferences in results visualization. Our SUI allows users to choose between a vertical list of snippets, tiles (Fig. 4b) and coverflow (Fig. 4a). In our experiment we demonstrated these three output types. The subjects interacted with the search system to get a better feeling and were encouraged to solve a simple search task using the prefered SUI setup. 44 subjects participated in the study, 27 children and 17 adults. The children were between eight and ten years old (8.9 on average), 19 girls and 8 boys from third (18 subjects) and fourth (9 subjects) grade. The adults were between 22 and 53 years old (29.2 on avarage), five women and 12 men. Nine of them were students in computer science and four worked in the IT sector. The results for the output are presented in Fig. 5. The majority of the children prefered the coverflow results visualization, whereas the adults had a week tendency towards tiles. These results can be explained by the fact that on average children cannot

[2]Similar to the *quinturakids.com* search engine, accessed on 02.05.2013

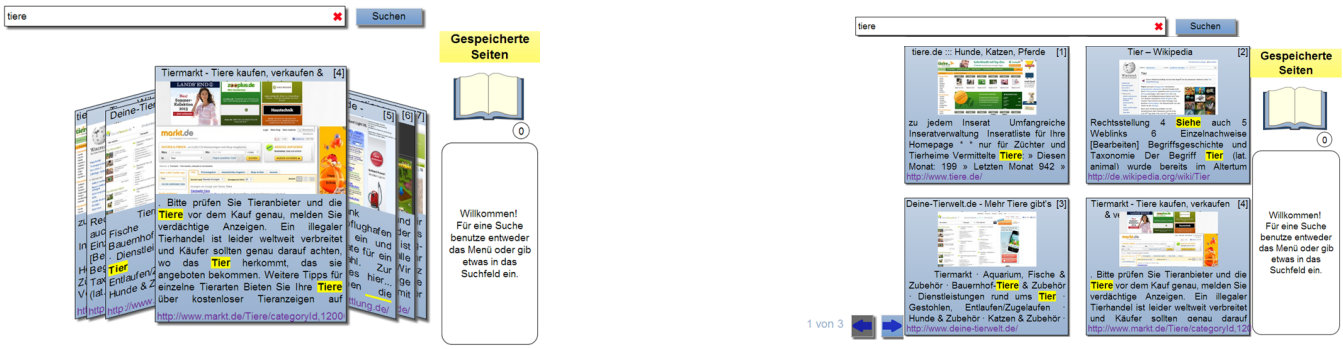


Figure 4: Different kinds of result visualization: a) ESUI with coverflow result visualization b) ESUI with tiles result visualization.

process much information, but adults do. Thus, it is easier for children to use coverflow. Coverflow offers an animation by browsing that is attractive for children. Many adults told us that they prefer tiles as, since many results can be compared at once, tiles offer a good overview of results.

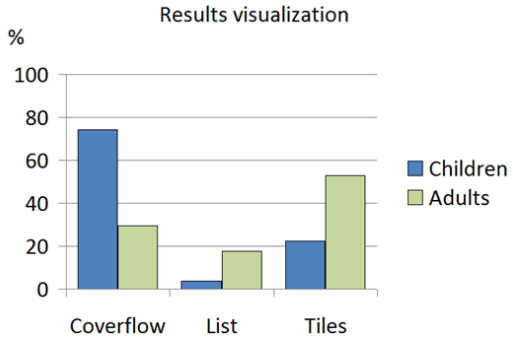


Figure 5: Study results: what type of visualization do children and adults prefer.

5. CONCLUSION

In this positional paper we introduced the concept of an evolving search user interface that adapts itself to abilities of a particular user. Instead of building a SUI for a specific user group, we use a mapping function between user skills and UI elements of a search system in order to adapt it dynamically, allowing the user to perform his search process in a more efficient way. We considered different abilities of a user, e.g. his cognitive skills, knowledge, reading and writing skills, that change during life. Furthermore, we proposed to adapt three main components of a SUI, i.e. input, output and management, and its general UI properties to the user skills. A key component of an ESUI is a mapping function between user skill space and UI elements of a SUI, that has to be found. We elaborate on ways to learn this function. In order for an ESUI to be adaptive, ways to detect user abilities are required. We pointed in several directions how the detection can be done.

6. ACKNOWLEDGEMENTS

The work presented here was partly supported by the German Ministry of Education and Science (BMBF) within the ViERforES II project, contract no. 01IM10002B.

7. REFERENCES

[1] A. Aula. User study on older adults'use of the web and search engines. *Universal Access in the Information Society*, 4(1):67–81, 2005.

[2] A. Aula and M. Käki. Less is more in web search interfaces for older adults. *First Monday*, 10(7-4), 2005.

[3] J. E. Birren and K. W. Schaie. *Handbook of the psychology of aging*, volume 2. Gulf Professional Publishing, 2001.

[4] C. Eickhoff, L. Azzopardi, D. Hiemstra, F. de Jong, A. de Vries, D. Dowie, S. Duarte, R. Glassey, K. Gyllstrom, F. Kruisinga, et al. Emse: Initial evaluation of a child-friendly medical search system. In *IIiX Symposium*, 2012.

[5] E. Erikson. *Children and society*. WW Norton & Company, 1963.

[6] T. Gossen, M. Nitsche, and A. Nürnberger. Knowledge journey: A web search interface for young users. In *Proc. of the Sixth Symposium on HCIR*, 2012.

[7] T. Gossen and A. Nürnberger. Specifics of information retrieval for young users: A survey. *Information Processing & Management*, 49(4):739–756, 2013.

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

[9] J. Hourcade, B. Bederson, A. Druin, and F. Guimbretière. Differences in pointing task performance between preschool children and adults using mice. *ACM Transactions on Computer-Human Interaction*, 11(4):357–386, 2004.

[10] M. Jansen, W. Bos, P. van der Vet, T. Huibers, and D. Hiemstra. TeddIR: tangible information retrieval for children. In *Proc. of the 9th Int. Conf. on Interaction Design and Children*, pages 282–285. ACM, 2010.

[11] R. Kail. Developmental change in speed of processing during childhood and adolescence. *Psychological bulletin*, 109(3):490, 1991.

[12] J. Ormrod and K. Davis. *Human learning*. Merrill, 1999.

[13] B. Steichen, H. Ashman, and V. Wade. A comparative survey of personalised information retrieval and adaptive hypermedia techniques. *Information Processing & Management*, 2012.

[14] S. Stober and A. Nürnberger. Adaptive music retrieval–a state of the art. *Multimedia Tools and Applications*, pages 1–28, 2012.

[15] I. Stuart-Hamilton. *Intellectual changes in late life*. John Wiley & Sons, 1996.

A Pluggable Work-bench for Creating Interactive IR Interfaces

Mark M. Hall
Sheffield University
S1 4DP, Sheffield, UK
m.mhall@sheffield.ac.uk

Spyros Katsaris
Sheffield University
S1 4DP, Sheffield, UK
evolve.sheffieldis@gmail.com

Elaine Toms
Sheffield University
S1 4DP, Sheffield, UK
e.toms@sheffield.ac.uk

ABSTRACT

Information Retrieval (IR) has benefited from standard evaluation practices and re-usable software components, that enable comparability between systems and experiments. However, Interactive IR (IIR) has had only very limited benefit from these developments, in part because experiments are still built using bespoke components and interfaces. In this paper we propose a flexible workbench for constructing IIR interfaces that will standardise aspects of the IIR experiment process to improve the comparability and reproducibility of IIR experiments.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces

Keywords

evaluation, framework, standardisation

1. MOTIVATION

Information Retrieval (IR) has benefited from standard evaluation practices and re-usable software components. The Cranfield-style evaluation methodology enabled evaluation programmes such as TREC, INEX, or CLEF. At the same time provision of re-usable software components such as Lucene[1], Terrier[2], Heritrix[3], or Nutch[4] have enabled IR researchers to focus on the development of those components directly related to their research. However, Interactive IR (IIR) as had only very limited benefit from these developments.

Typically IIR research is still conducted using a single system in a laboratory setting in which a researcher observed and interacted with a participant [5], usually using a bespoke IIR interface. Developing and running such experiments is a time-consuming, resource exhaustive and labour intensive process [6]. As a result of this bespoke approach, the comparability of IIR experiments and their results suffers. Where studies of the same activities show divergent results, it is difficult to determine whether the differences are due to the specific aspect of IIR under investigation, or simply due to different participant samples or small differences in how the non-investigated user-interface (UI) components were implemented. The bespoke nature also makes it harder to replicate studies, as publications frequently do not contain sufficient detail to exactly replicate the experiment.

In [3] we have proposed a flexible, standardised IIR evaluation framework that aims to address the issues created by variations in the experimental processes and by how context information is acquired from the participants. However, the framework makes no provisions towards providing standardised IIR components that would improve the comparability of the experiment itself, the ease of setting up the experiment, and the ease of reproducibility.

A number of attempts at developing a configurable, reusable IIR evaluation system have been made in the past. In 2004, Toms, Freund and Li designed and implemented the WiIRE (Web-based Interactive Information Retrieval) system [6], which devised an experimental workflow process that took the participant through a variety of questionnaires and the search interface. Used in TREC 11 Interactive Track, it was built using Microsoft Office desktop technologies, severely limiting its capabilities. The system was re-created for the web and successfully used in INEX2007 [7], but lacked flexibility in setup and data extraction. More recently, SCAMP (Search ConfigurAtor for experiMenting with PuppyIR) [4] was developed to assess IR systems, but does not include the range of IIR research designs that are typically done. A heavy-weight solution is PIIRExS[5] [1], which supports the researcher through the whole process from setting up the experiment to analysis, providing greater support but also a steeper learning curve. These approaches highlight the difficulty of balancing the two main constraints that limit a system's wide-spread use:

- sufficient flexibility to support the wide range of IIR interfaces and experiments;
- sufficiently simple to implement that it does not increase the resource commitment required to set up the experiment.

[1] https://lucene.apache.org/
[2] http://terrier.org/
[3] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
[4] http://nutch.apache.org/
[5] `http://sourceforge.net/projects/piirexs`

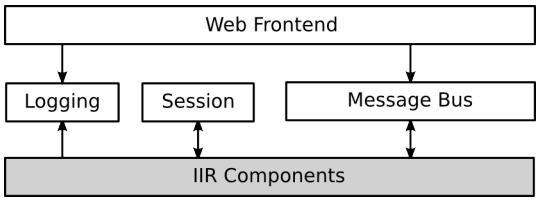


Figure 1: The evaluation workbench consists of the four core modules, into which the IIR components used in the experiment are plugged.

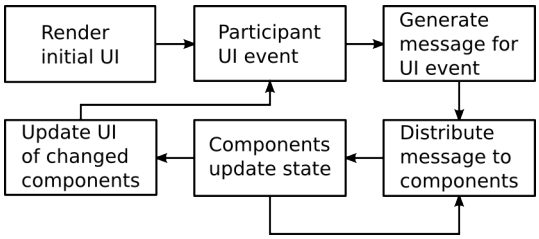


Figure 2: The workbench's main workflow starts with the generation of the initial UI and then waits for the participant to generate a UI event. The event is processed, the affected component's state and UI are updated and the workbench goes back to waiting for the next UI event. A powerful aspect of the workflow is that components when they receive a message, can generate their own messages.

2. DESIGN

To achieve the goal of developing a system that fulfils these requirements, we propose a system design that is based around a very lean core into which the researcher can plug the IIR components they wish to include in their experiment. We have implemented this design in our web-based evaluation framework (fig. 1), which complements the larger IIR experiment support system presented in [3]. To achieve maximum flexibility, the system was designed using a message-passing architecture that consists of the following four components:

- **Web Frontend** is handles the interface between the participant's browser and the evaluation workbench and is implemented using a combination of client-side and server-side functionality.
- **Message Bus** handles the inter-component communication and forms the core of the system. It is responsible for passing messages from the **Web Frontend** to the IIR components configured to be listening for those messages and also for passing messages directly between the components.
- **Session** handles loading and saving the components' current state for a specific participant, hiding the complexities of web-application state from the individual components.
- **Logging** provides a standardised logging interface that allows the components to easily attach logging information to the UI event generated by the participant.

```
[SearchResults]
handler = application.components.SearchResults
name = search_results
layout = grid-9 vgrid-expand
connect = search_box:query
```

Figure 3: Configuration for a *Standard Results List* component, showing how the component's layout (9 grid-cells wide and vertically expanding) and connections to other components (to the "search_box" component via the `query` message) are specified.

When the researcher sets up the workbench for their experiment, they can freely configure which components to use, how to lay them out, and which components to connect to which other components. Based on this configuration the **Web Frontend** generates the initial user-interface that is shown to the participants. Then, when the participant interacts with a UI element (fig. 2), the resulting UI event is handled by the **Web Frontend**, which generates a message based on the UI event. This message is passed to the **Message Bus**, which uses the configuration provided by the researcher to determine which components to deliver the message to. The components that are listening for that message update their own **Session** state based on the message and then mark themselves as *changed*. After message processing has been completed for all components, the **Web Frontend** then updates the UI for each of the *changed* components.

An example of the configuration used to set-up the experiment is shown in figure 3 (from the experiment in figure 4), specifying the configuration of the "search_results" component. It specifies that the component should be displayed 9 grid-cells wide (the application layout uses a 12-by-12 cell grid layout) and should expand vertically to use as much space as is available. The component is configured to be connected to the "search_box" component via the "query" message. It is this ability to freely plug components together that, we believe, makes the framework sufficiently flexible to support the wide range of IIR experiments, while remaining simple to set-up and use.

3. STANDARD COMPONENTS

The core system provides only the framework into which the IIR components can be plugged. This allows the researcher to build any custom IIR UI they wish to test, while at the same time being able to take advantage of the standardised session and log handling functionality. As IIR UIs frequently include required elements that are not the focus of the study the researcher wishes to undertake, an optional set of default components for core IR UI elements is provided to reduce set-up time. This has the additional advantage that as their behaviour is consistent across experiments, the comparability of experiments using the framework is improved.

3.1 Search Box

The *Search Box* component ([8], p. 49, "Formulate Query Interface" [2], p. 76) provides a standard search box. When the participant enters text and clicks on the "Search" button,

it generates a `query` message, which is usually connected to a *Standard Results List*.

3.2 Standard Results List

The *Standard Results List* component ([8], p. 50, "Examine Results Interface" [2], p. 77) provides a default 10 item listing of search results. The *Standard Results List* includes support for displaying snippets ([8], p. 51) and what Wilson calls "Usable Information" ([8], p. 51) for each result document. Unlike the other standard components, which can be used out-of-the-box, the *Standard Results List* has to be extended by the researcher in order to be able to access the search-engine used to power the UI.

3.3 Pagination

The *Pagination* component ([8] p. 70) displays a configurable number of pages around the current search-results page. In response to user interaction it sends a `start` message with the rank of the first document to paginate to.

3.4 Category Browsing

The *Category Browsing* component ([8], p. 54) provides a hierarchical category structure that the participant can use to explore a collection. Clicking on a category sends a `query` message with the category's identifier.

3.5 Saved Documents

The *Saved Documents* component provides an area where the participant can save things that they have found interesting, to support them in their current task. Documents are added through a `save_document` message. The *Saved Documents* component supports an optional tagging feature enabling the participant to tag the document with values specified by the researcher. This can be used to let the participant specify why they have chosen that document or how much it helps them in their current task.

3.6 Task

The *Task* component provides a static display of the task information to show to the user. Two versions of this component are provided, one that displays a static text set in the configuration, and one that can fetch a task description from the database, based on a parameter passed to it.

4. APPLICATION

The evaluation work-bench has so far been used to build two IIR experiments, very different in their nature, clearly demonstrating the work-bench's flexibility.

The first experiment (fig. 4) re-uses the standard *Task*, *Search Box*, *Pagination*, and *Saved Documents* components, and extends the *Standard Results List* to work with the specific search backend. This set-up re-creates what is essentially a relatively standard search UI configuration, that is being used to investigate query session behaviour.

The second experiment (fig. 5) demonstrates a much richer interface, with more modifications to the components and an experiment-specific component. It re-uses the *Task* and *Category Browsing* components, extends the default *Search Box*, *Pagination*, *Standard Results List*, and *Saved Documents* components, and adds a new *Item View* component. The message-passing nature of the system made it possible to quickly integrate the new component, so that when the participant clicks on a meta-data facet in the *Item View*, a `query` message is sent to the *Standard Results List* to find items with the same bit of meta-data. The interface was used to investigate un-directed exploration behaviour in a large digital cultural heritage collection.

5. WHERE TO GO NEXT?

The stated aim of this paper was to present a novel, pluggable, extensible, and configurable IIR interface work-bench, that supports our wider aim of improving IIR experiment comparability. The work-bench is sufficiently flexible to support the wide range of web-based IIR experiments that are undertaken, while being sufficiently simple and light-weight to encourage wide-spread use of the workbench.

To enable this wide-spread use, the system has been released under an open-source license[6]. We are also moving to engage with the wider research community to determine to what degree the work-bench satisfies their needs for an evaluation system and what needs to be done to achieve the wide-spread use needed to improve IIR experiment comparability.

6. ACKNOWLEDGEMENTS

The research leading to these results was supported by the Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191.

7. REFERENCES

[1] R. Bierig, M. Cole, J. Gwizdka, N. J. Belkin, J. Liu, C. Liu, J. Zhang, and X. Zhang. An experiment and analysis system framework for the evaluation of contextual relationships. In *CIRSE 2010*, page 5, 2010.

[2] C. Chua. A user interface guide for web search systems. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, OzCHI '12, pages 76–84, New York, NY, USA, 2012. ACM.

[3] M. M. Hall and E. G. Toms. Building a common framework for iir evaluation. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative - CLEF 2013*, 2013.

[4] G. Renaud and L. Azzopardi. Scamp: a tool for conducting interactive information retrieval experiments. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 286–289. ACM, 2012.

[5] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.

[6] E. G. Toms, L. Freund, and C. Li. Wiire: the web interactive information retrieval experimentation system prototype. *Information Processing & Management*, 40(4):655–675, 2004.

[7] E. G. Toms, H. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. Macnutt. Task effects on interactive search: The query factor. In *Focused access to XML documents*, pages 359–372. Springer, 2008.

[8] M. L. Wilson. *Search User Inteface Design*, volume 20. Morgan & Claypool Publishers, 2011.

[6] `https://bitbucket.org/mhall/pyire`

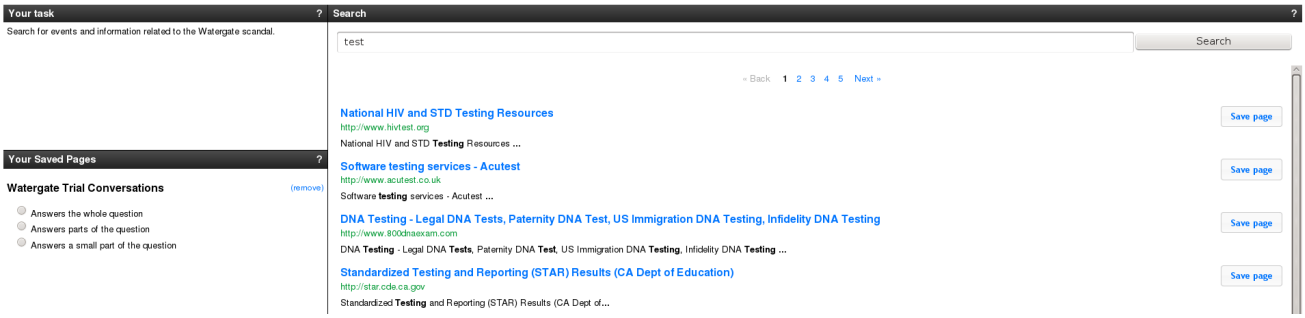


Figure 4: Screenshot showing an experiment with a very basic configuration consisting of *Task*, *Search Box*, *Pagination*, *Standard Results List*, and *Saved Documents* components. This is being used to investigate query behaviour for tasks that require query reformulations.

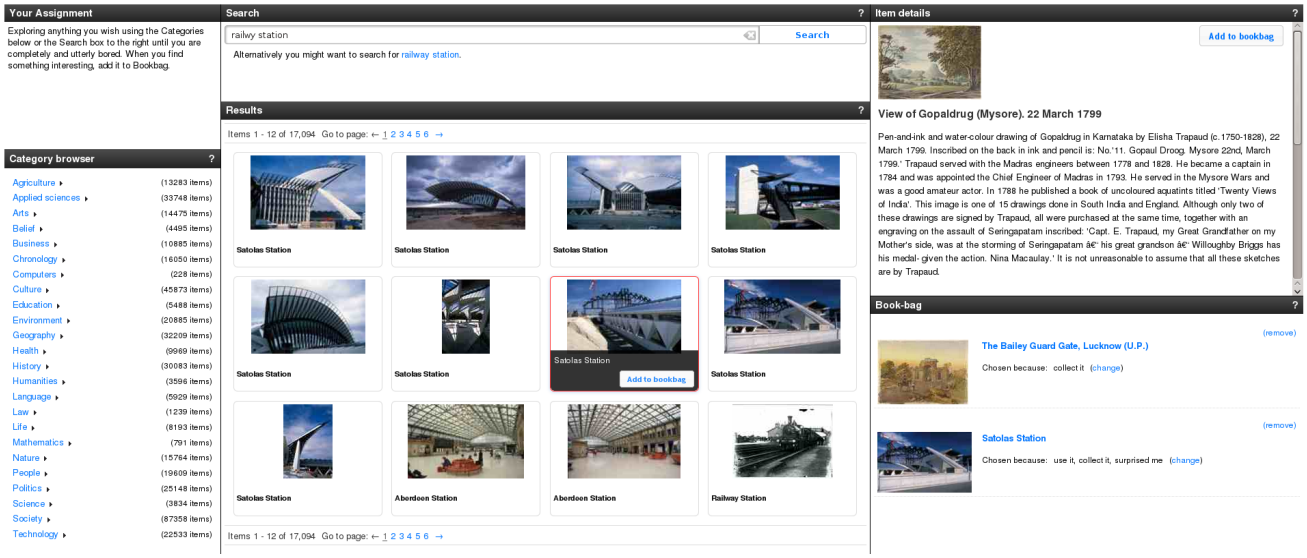


Figure 5: Screenshot showing an experiment that makes heavy use of the customisation options offered by the workbench. This configuration was used to investigate un-directed exploration in a digital cultural heritage collection.

A Proposal for User-Focused Evaluation and Prediction of Information Seeking Process

Chirag Shah
School of Communication & Information (SC&I)
Rutgers University
4 Huntington St, New Brunswick, NJ 08901, USA
chirags@rutgers.edu

ABSTRACT
One of the ways IR systems help searchers is by predicting or assuming what could be useful for their information needs based on analyzing information objects (documents, queries) and finding other related objects that may be relevant. Such approaches often ignore the underlying search process of information seeking, thus forgoing opportunities for making process-based recommendations. To overcome this limitation, we are proposing a new approach that analyzes a searcher's current processes to forecast his likelihood of achieving a certain level of success in the future. Specifically, we propose a machine-learning based method to dynamically evaluate and predict search performance several time-steps ahead at each given time point of the search process during an exploratory search task. Our prediction method uses a collection of features extracted solely from the search process such as dwell time, query entropy and relevance judgment in order to evaluate whether it will lead to low or high performance in the future. Experiments that simulate the effects of switching search paths show a significant number of subpar search processes improving after the recommended switch. In effect, the work reported here provides a new framework for evaluating search processes and predicting search performance. Importantly, this approach is based on user processes, and independent of any IR system allowing for wider applicability that ranges from searching to recommendations.

Categories and Subject Descriptors
H.3: INFORMATION STORAGE AND RETRIEVAL **H.3.3: Information Search and Retrieval**: *Search process*; H.3: INFORMATION STORAGE AND RETRIEVAL **H.3.4: Systems and Software**: *Performance evaluation (efficiency and effectiveness)*

General Terms
Measurement, Performance, Experimentation

Keywords
Exploratory search, Evaluation, Performance prediction

Presented at EuroHCIR2013.

1 INTRODUCTION
IR evaluations are often concerned with explaining factors relating to user or system performance after the search and retrieval are conducted [20]. Most recommender systems, however, operate with an objective to suggest objects that could be useful to a user based on his/her or others' past actions [2][19]. We commenced our investigation by broadly asking how we could take valuable lessons from both IR evaluations and recommender systems to not only evaluate an ongoing search process, but also predict how well it will unfold and suggest a better path to the searcher if it is likely to underperform. The motivation behind this investigation was based on the following assumptions and realizations grounded in the literature.

1. The underlying rational processes involved in information search are reflected in the actions users make while searching. These actions include entering search queries, skimming the results, as well as selecting and collecting useful information [8][14][15].
2. A searcher's performance is a function of these actions performed during a search episode [7][22].

With these assumptions, we propose to quantify a search process using various user actions, and use it for user performance (henceforth, 'search performance' or 'performance') prediction as well as search process recommendations.

2 BACKGROUND
Past research on predictive models that relates to the approach we describe in this paper can be grouped into two main categories: (1) behavioral studies and (2) IR approaches. In both cases; however, the focus has been on end products instead of in the process required to produce them.

As far as the behavioral studies go, research has been conducted to explore users models that help anticipating specific aspects of the search process. One goal in this context has been the determination of whether a search process will be completed in a single or multiple sessions. For example, Agichtein *et al.* [3] investigated different patterns that can be identified in tasks that require multiple sessions. As a result, the authors devised an algorithm capable of predicting whether users will continue or abandon the task. Similar work is described in Diriye *et al.* [6], which focuses on predicting and understanding of why and when users abandon Web searches. To address this problem, the authors studied features such as queries and interactions with result pages. Based on this approach, the authors were able to determine reasons for search abandonment such as accidental causes (e.g. Web browser crashing), satisfaction levels, and query suggestions, among others.

There have been also attempts to understand past users' behaviors in order to predict future ones in similar conditions. For example, Adar *et al.* [1] visually explored behavioral aspects using large-scale datasets containing queries and other information objects produced by users. The authors were able to identify different behavioral patterns that seem to appear consistently in different datasets. While not directly related to performance prediction, this work focused on attributes of the search process instead of in final products derived from it.

Research like the ones described above often relies on historic data from large populations and the use of trend and seasonal components, which are used to model long-term direction and periodicity patterns of time-series [17]. For example, some have explored seasonal aspects in Web search (e.g. weekly, monthly, or annual behaviors) that provides useful information to predict and suggest queries [5].

From an IR perspective, Radinski *et al.* [18] explored models to predict users' behaviors in a population in order to improve results from IR systems. The authors also developed a learning algorithm capable of selecting an appropriate predictive model depending on the situation and time. As described by the authors, applications of this approach could go from click predictions to query-URL predictions. In contrast to this approach, our method presented in this paper considers both the population trends and an individual user behavior.

In a similar track, several works have been conducted on query performance prediction, focusing on developing techniques that help IR system to anticipate whether a query will be effective or not to provide results that satisfy users' needs [4][10][11]. For example, Gao *et al.* [10] found that features derived from search results and interactions features offer better prediction results than a prediction baseline defined in terms of query features. Results from this study have direct implications to individual users by aiding the auto evaluation process of IR systems.

In information search, users may be unaware of their individual performance when solving an information search task. For instance, Shah & Marchionini [23] showed how lack of awareness about different objects involved in searching (queries, visited pages, bookmarks) could result in mistaken perception about search performance during an exploratory search task. Even if an IR system is highly effective, users may run into multiple query formulation and evaluation of several pages before finding what they need. This process, which can be related to search strategies, implies effort and time that is usually underestimated by the users themselves. In this sense, instead of predicting end products (i.e., overall performance), the approach we introduce in this paper is oriented toward predictions at different times in order to increase the level of awareness of users about their own search process. Similar to weather forecast, this information could help users to be aware of possible trends based on past and current behavior.

For a more recent discussion on IR evaluations and their shortcomings, see [12]. To the best of our knowledge, search process performance prediction at different times from a user perspective has not been explored. Similar approaches can be found in weather and stock market studies. For example, using machine learning approaches such as Support Vector Machine (SVM), some models have been implemented to predict the trends of two different daily stock price indices using NASDAQ and Korean Stock prices [13][16]. In a similar fashion, our approach is oriented to forecast users' search performance *N*-steps ahead with the aim to aid their search process awareness and performance trends.

Unlike previous works in IR, we are not proposing to use time series analyses or seasonal components of historic data. Instead, we investigate predictive models based on machine learning (ML) techniques; namely: SVM, logistic regression, and Naïve Bayes which are trained over a set of features such as time, number of queries, and page dwell time. In contrast to most IR evaluations, our method focuses on user-processes. Also, unlike most recommender systems, our approach could output alternative strategies instead of similar/relevant products to help the searcher. In essence, the work reported here takes several lessons from tradition IR evaluations, recommender system designs, and weather/stock forecasting to come up with a new approach for evaluating and predicting search performance.

In the next section we provide a detailed description of our method, feature selection, and the measures we used in order to create ML-based predictive models.

3 METHOD

In order to analyze the search processes followed by different users/teams, we assume that the underlying dynamics of the search processes are expressed by a collection of activities that take place from the beginning to the end of the search processes.

The first part of our method is a feature extraction step in which we extract a wide array of features relating to webpages, queries and snippets saved from the search processes for each unit of time *t.* This step is performed in order to evaluate how well we could use those features to capture the underlying dynamics which would lead to recognizing whether a search process is going to lead to high or low performance in the future time steps at $t+n$ $(n=1,2,....,N)$, where N is the furthest time step.

The decision to include or exclude a feature was based on literature (e.g., [7]) as well as our past experience [22] with representing and evaluating search objects and processes. Each feature is extracted for each user or team, *u,* up to time *t* from the search processes and they are explained in detail as follows.

- *Total coverage (u,t)*: The total number of distinct Webpages visited by a user (*u*) up to time *t*. This feature captures the Webpage based activity performed by a user and provides a measure to see how much distinct information has been found by the user up to this time.
- *Useful coverage (u,t)*: The total number of distinct webpages in which a user spent at least 30 seconds, up to time *t*. This measure evaluates out of the total pages he/she has visited how many of them were useful in finding relevant information leading to satisfaction with their context in completing the exploratory task [9][22][25].
- *Number of queries (u,t)*: Total number of unique queries executed by a user up to time *t*. This feature implicitly relates to how much effort and cognitive thinking a user has put in to this task.
- *Number of saved snippets (u,t)*: Total number of snippets saved by user *u* up to time *t*. This measures the amount of information that the user thought that might be relevant in the future to complete the task and needed to be remembered. In other words, this feature is an indication of explicit relevance judgments made by the user.
- *Length of Query (u,q,t)*: Length of each query(*q*) executed by a user *u* based on the character count of the query up to time *t*. This feature captures how the user imposed the

queries and how long they were at different times of the search process.

- *Number of tokens in each query (u,q,t)*: This is the count of tokens/words in each query(q) executed by user u up to time t. This query based measure takes into account how specific a user was in defining the query. By inspecting the datasets, we realized that queries with a less number of tokens tend to get general results. On the other hand, composed queries with multiple terms are related to more specific searchers. We also observed that typically the users started with general queries with few words at the beginning of the search process but then went into more detailed queries to find more specific information later. For all these reasons, we found it to be useful to capture the number of token used in a query.
- *Query entropy (u,q,t)*: This measures the information content in a given query (q), by finding the expected value of information contained in a query. We used the widely recognized notion of Shannon entropy [24] in Information Theory to calculate the information content of a query. We calculated the number of unique characters appearing in each of the queries, which represent the observed counts of the random variable. This was used as the input to Shannon entropy calculation and we used to the maximum-likelihood method to calculate the entropy. Query entropy feature has been used in the past to predict *goodness* of a query for making query expansion decision [21].

The method used to assess the search performance of a user is described below. We define a measure called *Efficiency (u,t),* for each user u up to time t *in order to* predict whether a given search process is going to yield in high/low performance in the future We first define *Effectiveness* of user u up to time t as the ratio of useful coverage and total coverage (both defined earlier). A similar measure was used in [7] and [22].

$$Effectiveness(u,t) = \frac{Useful\ coverage(u,t)}{Total\ coverage(u,t)} \quad (1)$$

We then calculated *Efficiency* as defined in Equation 2.

$$Efficiency(u,t) = \frac{Effectiveness(u,t)}{NumberofQueries(u,t)} \quad (2)$$

In other words, *Efficiency* is defined as the *Effectiveness* obtained per query, or how effective a query is in terms of achieving a certain level of useful coverage.

The performance for each user u at each time t was classified in to the two classes; high performance and low performance based on the following criteria:

$$Class = \begin{cases} high & ;if \quad Efficiency(u,t) \geq \overline{Efficiency(u,t)} \\ low & ;else \end{cases} \quad (3)$$

Using various user studies data available to us, we constructed feature matrices which consist of all aforementioned features for each minute of time t for all the users in each dataset, and converted in to a long vector of features which we fed as the input to the classification models used.[1] The class labels were generated as high/low performance at minute $t+n$ based on the above mentioned criteria and threshold and used as the output class labels to be used in the n-step ahead prediction model. If a class label at n-step ahead was correctly predicted based on the features extracted up to time t from the classification model it was considered as correctly classified and if not as misclassified.

[1] In the interest of space and scope of work here, details of these experiments have been omitted, but will be available for discussion at the workshop.

4 EXPERIMENTS

In order to evaluate whether users who are predicted to perform at low performance in the future based on the current search process, could benefit from this analysis to improve their search process, we conducted some simple simulation analysis.

We considered the individual user search processes as a collection of *search paths*, where each *search path* is defined as the search process from the time a user issued a query up to the time user issued another quite different query. This was found out using generalized Levenshtein (edit) distance, which is a commonly used distance metric for measuring the distance between two character sequences. If the Levenshtein (edit) distance between two subsequent queries were greater than 2 (assuming less than 2 was when there were changes in the queries due to simple spelling mistakes or refining of the query), we considered the search process from the former query to the next query as a single *search path*.

Following this method, we found the first *search path* of each user and based on the features extracted up to the end of the first *search path*, and based on the classification model learnt from that corresponding n-step ahead prediction we predicted whether the user is going to have low/high performance at the end of the session. If the user was going to have low performance, then out of the users who predicted to have high performance, we looked at which high performing user has the lowest Levenshtein (edit) distance between the queries issued by low performing user within the first search path and considered it as a pair of users, whom we are going to use in the simulation. Then, for each low performing user and high performing user that was matched, we switched the search process of low performing user at the end of the first *search path* with the high performing user's *search path* up to $t=T$ minutes, where T is the total number of minutes for a session. Then we evaluated by switching the search process early during the overall process whether it would benefit each low performing user to improve their performance. We found that we were able to move most of the underperforming search processes to higher performance by early detection and switching, while keeping the higher performing processes unharmed.

These simulations provide verification that by realizing early during the search process whether a user is going to perform well or not, one could recommend better search processes/strategies for that user which would lead to uplifting the search performance of a previously destined to low performing user.

5 CONCLUSION

When it comes to prediction, information retrieval and filtering systems are primarily focused on objects while assessing what and if something could help the users. These approaches are often system-dependent even though the process of information seeking is usually user-specific. Personalization and recommendations are frequently exercised as methods to address user-specific IR and filtering, but still limited to comparing and recommending objects, not focusing on underlying IR processes that are carried out by the searchers. We presented a new approach to address these shortcomings. We began by asking

whether we could model a user's search process based on the actions he/she is performing during an exploratory search task and forecast how well that process will do in the future. This was based on a realization that an information seeker's search goal/task can be mapped out as a series of actions, and that a sequence of actions or choices the searcher makes, and especially the search path he/she takes, affects how well he/she will do. Thus, in contrast to approaches that measure the *goodness* of search products (e.g., documents, queries) as a way to evaluate the overall search effectiveness, we measured the likelihood of an existing search process to produce *good* results.

Here we presented simulations to demonstrate what could happen if one can make process-based predictions, but one could develop an actual recommender system using the proposed method. Another potential application of such prediction-based method would be to use such approach in IR systems to provide the awareness to users how their future performance will be based on the current/past search process. The system could identify that a user will have low performance if, he continues this manner at an early stage of the process, and what could be done to provide suggestions to improve overall performance.

Given that the proposed technique is independent of any specific kind of system, and solely focused on user-based processes, it will presumably be easy to apply it to a variety of IR systems and situations irrespective of retrieval, ranking, or recommendation algorithms. Finally, while we have used datasets borrowed from previous user studies, one could easily apply the proposed method to Web logs, TREC data, and other forms of datasets with various user actions recorded over time.

6 ACKNOWLEDGEMENTS

The work reported here is supported by The Institute of Museum and Library Services (IMLS) Cyber Synergy project as well as IMLS grant # RE-04-12-0105-12. The author is also grateful to his PhD students Chathra Hendahewa and Roberto Gonzalez-Ibanez for their valuable contributions to this work.

7 REFERENCES

[1] Adar, E., Weld, D. S., Bershad, B. N., & Gribble, S. D. (2007). Why we search: visualizing and predicting user behavior. In *Proceedings of World Wide Web (WWW) Conference 2007*.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749.

[3] Agichtein, E., White, R.W., Dumais, S.T., & Bennett. P.N. (2012). Search interrupted: Understanding and predicting search task continuation. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR) 2012*.

[4] Cronen-Townsend, S., Zhou, Y., & Croft, B. (2002). Predicting query performance. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR) 2002*.

[5] Dignum, S., Kruschwitz, U., Fasli, M., Yunhyong, K., Dawei, S., Beresi, U.C., & De Roeck, A. (2010). Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 2010*, vol.1, no., pp.425-430, Aug. 31 2010-Sept. 3 2010

[6] Diriye, A., White, R.W., Buscher, G., & Dumais, S.T. (2012). Leaving so soon? Understanding and predicting web search abandonment rationales. In *Proceedings of CIKM 2012*.

[7] González-Ibáñez, R., Shah, C., & White, R. W. (2012). Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems. In *Proceedings of the 75th Annual Meeting of the Association for Information Science and Technology (ASIS&T)*. Baltimore, MD, USA.

[8] Gwizdka, J. (2008). Cognitive load on web search tasks. *Workshop on Cognition and the Web, Information Processing, Comprehension, and Learning*. Granada, Spain. Available from http://eprints.rclis.org/14162/1/GwizdkaJ_WCW2008_short_paper_finalp.pdf

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

[10] Gao, Q., White, R., Dumais, S.T., Wang, S., & Anderson, B. (2010). Predicting query performance using query, result and interaction features. In *Proceedings of RIAO 2010*.

[11] He, B., & Ounis, I. (2006). Query performance prediction, Information Systems, Volume 31, Issue 7, November 2006, Pages 585-594, ISSN 0306-4379, 10.1016/j.is.2005.11.003.

[12] Järvelin, K. (2012). IR research: systems, interaction, evaluation and theories. *ACM SIGIR Forum, 45*(2), 17. doi:10.1145/2093346.2093348

[13] Kyoung-jae, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, Volume 55, Issues 1–2, September 2003, Pages 307-319, ISSN 0925-2312, 10.1016/S0925-2312(03)00372-2.

[14] Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. *American Society for Information Science*, *47*(17). Available from http://dl.acm.org/citation.cfm?id=1920331.1920356

[15] Liu, J., Gwizdka, J., Liu, C., & Belkin, N. J. (2010). Predicting task difficulty for different task types. In *Proceedings of the Association for Information Science*, *47*(16). Available from http://dl.acm.org/citation.cfm?id=1920331.1920355

[16] Ming-Chi, L. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, Volume 36, Issue 8, October 2009, Pages 10896-10904, ISSN 0957-4174, 10.1016/j.eswa.2009.02.038

[17] Ord, J., Hyndman, R., Koehler, A., & Snyder, R. (2008). Forecasting with Exponential Smoothing (The State Space Approach). Springer, 2008.

[18] Radinski, K., Svore, K., Dumais, S. T., Teevan, J., Horvitz, E., & Bocharov, A. (2012). *Modeling and predicting behavioral dynamic*s on the Web. In *Proceedings of WWW 2012*.

[19] Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM*, *40*(3), 56–58.

[20] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 138–146).

[21] Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 2-9). Sheffield, UK.

[22] Shah, C., & Gonzalez-Ibanez, R. (2011). Evaluating the Synergic Effect of Collaboration in Information Seeking. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 913–922). Beijing, China.

[23] Shah, C., & Marchionini, G. (2010). Awareness in Collaborative Information Seeking. *Journal of American Society of Information Science and Technology (JASIST)*, *61*(10), 1970–1986.

[24] Shannon, C. E. and Weaver, W. *Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press, 1963.

[25] White, R. W., & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*. Geneva, Switzerland.

Directly Evaluating the Cognitive Impact of Search User Interfaces: a Two-Pronged Approach with fNIRS

Horia A. Maior[1,2], Matthew Pike[1], Max L. Wilson[1], Sarah Sharples[3]
[1]Mixed Reality Lab, [2]Horizon DTC, [3]Human Factors - School of Engineering
University of Nottingham, UK
{psxhama,psxmp8,max.wilson,sarah.sharples}@nottingham.ac.uk

ABSTRACT

Recent research has pointed towards further understanding the cognitive processes involved in interactive information retrieval, with most papers using secondary measures of cognition to do so. Our own research is focused on using direct measures of cognitive workload, using brain sensing techniques with fNIRS. Amongst various brain sensing technologies, fNIRS is most conducive to ecologically valid user studies, as it is less affected by body movement and can be worn while using a computer at a desk. This paper describes our two pronged approach focusing on a) moving fNIRS research beyond simple psychological tests towards actual interactive IR tasks and b) evaluating real search user interfaces.

Categories and Subject Descriptors

H5.2 [**Information interfaces and presentation**]: Evaluation/methodology, Theory and methods

Keywords

Functional near-infrared spectroscopy(fNIRS), Brain-computer interface(BCI), Human cognition, Information processing system, Multiple resource model, Limited resource model

1. INTRODUCTION

The cognitive aspects of Information Retrieval (IR) have repeatedly received focus over time, from Ingwersen's Cognitive Model [11], to recent analyses of cognitive workload during search tasks [2, 10]. The recurring interest is in what users think about at different task stages, and how much mental workload is involved. The benefits of knowing more about the searcher's cognitive state would come from providing better support for their needs, with Wilson et al suggesting that better designed Search User Interfaces (SUIs) could reduce unnecessary workload on the user [23].

Although some prior work (e.g. [2]) have used indirect techniques to analyse workload during search tasks, the decreasing cost of brain sensing hardware has meant that more recent research is using more objective techniques. Pike et al [17] and Gwizdka et al [10] used EEG technology, while Moshfeghi et al used fMRI to measure workload when making relevance judgements [15]. Each of these technologies have known limitations for studying actual interactive IR behaviour, with EEG being highly affected by even tiny body movement, and fMRI requiring users to lay in tunnel void of any metal objects. Recent Human-Computer Interaction research has listed the benefits of fNIRS brain sensing techniques, which are less affected by body movement, and can be more easily used in ecologically valid study conditions.

Functional Near Infrared Spectroscopy (fNIRS) is an emerging neuroimaging technique that is non-invasive, portable, inexpensive and suitable for periods of extended monitoring. fNIRS measures the hemodynamic response - the delivery of blood to active neuronal tissues. fNIRS is designed to be placed directly upon a participants scalp, typically targeting the prefrontal cortex. This paper describes our two-pronged approach to using fNIRS to study the cognitive workload created by SUIs, focused on a) task analysis and b) SUI analysis.

2. RELATED WORK

Understanding the cognitive aspects of interactive searching (as well as interaction in general) has been a long-standing goal for researchers in the field of Interactive IR. In the 1970s Bates suggested that searchers employ both search tactics and idea tactics [7]. In an attempt to explain an individual's path during IR, Bates' "Berrypicking" model [8] argued that search will vary as the user recognises information and has new ideas and questions.

In the main cognitive evolution of information seeking research, Ingwersen proposed a cognitive model of IR [11], where the searcher's understanding of the document collection, system, and task that would determine which path a search would take. The model again put the user's cognition as the central point of interest. More recently, Joho [12] argued that the cognitive effects typically observed in Psychology could provide a potential building block of theoretical development for evaluating interactive IR. Back et al [2], for example, examined the cognitive demands on users during the relevance judgement phase, suggesting that the amount of workload involved was the reason behind searchers rarely providing relevance judgements in previous work. Using a secondary measure, the Stroop task, Gwizdka [10] mapped varying levels of workload at multiple stages of search.

More recently, researchers have focused on objectively measuring interactive IR phases, in line with Back et al's work, Moshfeghi et al measured workload during relevance assessments by asking people to make judgements while lying in an fMRI machine. As making relevance judgements can be performed without directly interacting with a computer, this made use of an fMRI machine more realistic. Using more commercialised tools, Anderson [1] used an EEG sensor to compare visualization techniques in terms of the burden they

place on a viewer's cognitive resources. Similarly, Pike et al [17] developed a prototype tool named CUES that was capable of collecting a variety of data including EEG whilst interacting with a website. Pike et al used this to monitor aspects such as frustration and concentration, but their work demonstrated the variability of EEG data across the several minutes involved in an interactive IR task.

Using fNIRS, as introduced above, Peck [16] performed a similar study of different visualisation techniques, while a system called Brainput [18] was able to identify and correlate brain activity patterns among users during multitasking studies, and intervene when it sensed workload exceeding a certain level. Our work intends to build upon these HCI studies, to study interactive IR tasks and SUIs in more ecologically valid user study situations.

3. RESEARCH PATHS

Pike et al [17] highlighted the challenges of using brain sensing technologies to evaluate IIR tasks: that tasks have different stages, that behaviour quickly diverges after the first interaction (and thus is hard to compare), and that brain measurements vary dramatically over time. In order to address these challenges, we have initiated two clear research paths, both utilising fNIRS technology: 1) evaluating the cognitive aspects of Interactive IR tasks and 2) methods to evaluate the design of SUIs. The aim of the first path, is to move beyond using fNIRS to measure workload in simplistic psychology memory tasks (like Peck et al [16]), towards being able to break down real search tasks into primary components. This implies three considerations:

- Collected data would be meaningless if is not related to existing knowledge. Therefore, to interpret sensed fNIRS data we use proposed theories and models.
- It is known that fNIRS can sense cognition information [19, 16] related to so called working memory (if placed on the forehead). Assuming this is correct, we are using models of working memory.
- The proposed models will help us interpret the sensed data with fNIRS and have a better understanding of the cognitive impact of various complex tasks (such as a IR).

Such a technique would allow researchers to analyse data by stage, and find effective points of comparison during several minutes of continuous measurements. The second path is focused on identifying which aspects of working memory are affected by different features of SUIs, such that researchers can objectively evaluate the effect of different SUI design decisions. A combination of both paths works towards being able to proactively evaluate how SUIs support searchers.

4. PATH 1: WORKLOAD MODELS

To understand the cognitive aspects of IIR, it is essential to learn about user's capabilities and limitations in terms of their cognition: how people perceive, think, remember, and process information. This path of research focuses on existing models from Cognitive Psychology and Human Factors, models that conceptualize and highlight aspects that typically describe or influence elements of human cognition.

One important part of cognition during interactive searching involves human memory systems. There are two different types of memory [21]: working memory (sometimes called short-term memory) and long-term memory. Wickens describes working memory as the temporary holding of information that is "active", while long-term memory involving the unlimited, passive storage of information that is not currently in working memory.

Working memory. Working memory, proposed by Baddeley and Hitch (1974) [6], refers to a specific system in the brain which "provides temporary storage and manipulation of information..." [3]. Working memory [6, 4, 5] processes information in two forms: verbal and spatial, and has four main components (Figure 1):

- A **central executive** managing attention, acting as supervisory system and controlling the information from and to its "slave systems".
- A **visuo-spatial sketch pad** holding information in an analogue **spatial** form (e.g. Colours, shapes, maps, etc.), specialised on learning by means of visuospatial imagery.
- A **phonological loop** holding **verbal** information in an acoustical form (e.g. Numbers, words, etc.); specialised on learning and remembering information using repetition.
- A **episodic buffer** dedicated to linking verbal and spatial information in chronological order. It is also assumed to have links to long-term memory.

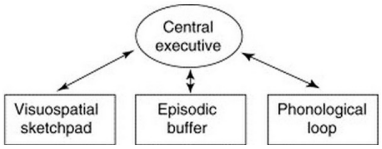


Figure 1: Baddeley's Working Memory Model

Information processing system. As humans, we are exposed to large amounts of information via our sensory systems. One of our strengths is in selecting information from our environment, perceiving it, processing it, and creating a response. Therefore we can use this understanding of brain activity to identify which elements of an interactive IR environment need to be considered when measuring brain activity, and how we can reduce rather than increase a user's mental workload via interface and system design.

Wicken's Information Processing Model [21] aims to illustrate how elements of the human information processing system such as attention, perception, memory, decision making and response selection interconnect. We are interested in observing how and when these elements interconnect during IR. He describes three different 'stages' (see STAGES dimension in Figure 2) at which information is transformed: a perception stage, a processing or cognition stage, and a response stage, the first two being processes involved in cognition. The first stage involves perceiving information that is gathered by our senses and provide meaning and interpretation of what is being sensed. The second stage represents the step where we manipulate and "think about" the perceived information. This part of the information processing

system takes place in working memory and consists of a wide variety of the mental activities. In relation to IR, it is interesting to observe how elements of cognition, such as rehearsal of information, planning the search strategy and deciding on the search keywords interconnect.

Multiple Resource Model. One model of mental workload that has been widely accepted in Human Factors is Wickens Multiple Resource Model [20] (Figure 2). The elements of this model overlap with the needs and considerations of evaluating complex tasks (such as IR). He describes the aspects of human cognition and the multiple resource theory in four dimensions:

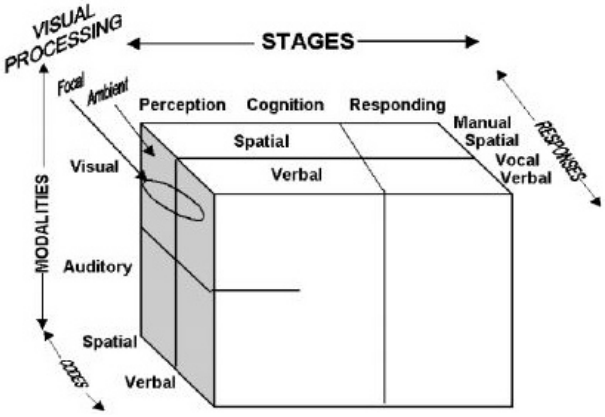


Figure 2: The 4-D multiple resource model [20]

- The STAGES dimension refers to the three main stages of information processing system (Wickens, 2004 [21]).
- The MODALITIES dimension indicating that auditory and visual perception have different sources.
- The CODES dimension refers to the types of memory encodings which can be spatial or verbal.
- The VISUAL PROCESSING dimension refers to a nested dimension within visual resources distinguishing between focal vision (reading text) and ambient vision (orientation and movement).

Our aim is to understand how these elements link together and compose more complex components/tasks. Additionally we want to consider how complex tasks (such as a search task) can be divided into primary components according to the models described. This will help identify possible problems in SUI design as well as indicating a possible solution to the problem (suggested implications by Wickens [21]):

- Minimize working memory load of the SUI system and consider working memory limits in instructions;
- Provide more visual echoes (cues) of different types during IR (verbal vs spatial);
- Exploit chunking (Miller, 1956 [14]) in various ways: physical size, meaningful size, superiority of letters over numbers, etc;
- Minimize confusability;
- Avoid unnecessary zeros in codes to be remembered;
- Encourage regular use of information to increase frequency and redundancy;
- Encourage verbalization or reproduction of information that needs to be reproduced in the future;
- Carefully design information to be remembered;

Resource vs Demands. One other model that is of interest is the limited resource model [22] describing the relationship between the demands of a task, the resources allocated to the task and the impact on performance.

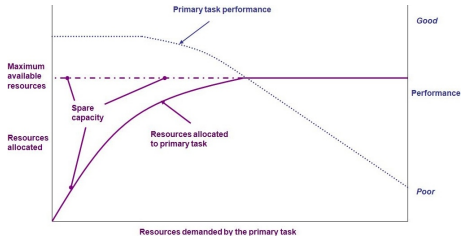


Figure 3: Resources available vs task demands → impact on performance [22]

The graph from Figure 3 is used to represent the limited resource model. The X-axes represent the resources demanded by the primary task and as we move to the right of the axes, the resources demanded by the primary task increase. The axes on the left indicate the resources being used, but also the maximum available resources point (if we think of working memory that is limited in capacity). The right axes indicate the performance of the primary task (the dotted line on the graph). The key element of this model is the concept of a limited set of resources which, if exceeded, has a negative impact on performance. However, it does not distinguish between resource modality, therefore we propose to use both the limited and multiple resources models to inform our work.

5. PATH 2: SUI EVALUATION

Relating quantitative data from brain sensing devices into feedback about SUI designs is one of our ultimate goals in conducting this research. SUIs are inherently information rich and thus affect both visual (results page layout) and verbal (text based results) memory. Detecting a change in either verbal or spatial working memory would help determine if a workload difference was caused by SUI design (spatial) or the amount of information the design provides (verbal). Our first in-progress study has stimulated each memory type in different tasks - Verbal memory was tested by performing an n-back [13] number memory task, whereas spatial memory was tested using an n-back visual block matrix task. Other studies have also looked at each type of memory and confirmed fNIRS ability to detect changes in heamodynamic responses accordingly [9].

In addition to developing an understanding of the extent to which we can monitor different memory, our initial study also sought to measure the effect of artefacts on the fNIRS data. Controlling the environment and human derived sources of noise is a potentially difficult factor to control without effecting the ecological validity of a study.

Solovey et al [19] showed that fNIRS is relatively resilient to motion derived artefacts when compared to EEG [17] for example, but still required some consideration by researchers conducting studies. In our own experience, we found that asking participants to remain still as much as possible was fairly successful. We are additionally looking at possible methods for correcting motion derived artefacts using an external gyroscope connected to the participant.

Designing tasks for experiments that measure cognitive effect via a brain sensor require careful consideration in order to ensure that results can be attributed to a cause. Thankfully this problem space has been well explored in the field of Psychology and we are able to adapt the approaches described in the literature to suit our task type requirements. A primary example of this adaptation is demonstrated by Peck et al [16], where 2 data visualisations techniques were compared using a methodology based loosely on the n-back task - a widely used psychology task that is designed to increase load on working memory.

Additionally, we are interested in exploring standard search studies (without following a psychological study layout) and seeing whether interesting states can be detected. Solovey et al [18] performed a similar function by utilising a machine learning algorithm that had classified "states of interest" prior to performing a task.

Using a similar approach, we could evaluate a SUI to determine whether a particular change in layout has a positive or negative impact on visual memory. Alternatively, to test the relevance of a results page (which would be dependant on the textual results), we could analyse the effects on verbal memory between 2 varied results pages, we could then reflect these changes to the Wickens Multiple Resource Model [20]. We are also working towards enabling the interpretation of data within the context of complex multimodal tasks to further extending our knowledge of the processes involved during IR and how they interact and effect one another.

6. SUMMARY

This paper has aimed to summarise our two-pronged approach towards actually evaluating the design of search user interfaces, in realistic ecologically valid study conditions, using fNIRS technology. The approach first involves braking down interactive IR tasks into how they effect the different elements of working memory, and second understanding how SUIs are processed by different parts of working memory. Our two paths of research will build towards a stage where we can combine them and objectively evaluate cognitive workload involved in interactive IR. We believe that this research will provide a novel new direction that SUI's and indeed HCI in a broader sense can benefit from. The association of physical recordings in ecological valid settings, to an existing theoretical model, provides a new measure from which future SUI development and evaluation could benefit.

7. REFERENCES

[1] E. W. Anderson, K. Potter, L. Matzen, J. Shepherd, G. Preston, and C. Silva. A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum*, 30(3):791–800, 2011.

[2] J. Back and C. Oppenheim. A model of cognitive load for IR: implications for user relevance feedback interaction. *Information Research*, 6(2):6–2, 2001.

[3] A. Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[4] A. Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

[5] A. D. Baddeley. Is working memory still working? *European psychologist*, 7(2):85–97, 2002.

[6] A. D. Baddeley and G. Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[7] M. J. Bates. Idea tactics. *JASIST*, 30(5):280–289, 1979.

[8] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424, 1989.

[9] X. Cui, S. Bray, D. M. Bryant, G. H. Glover, and A. L. Reiss. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *Neuroimage*, 54(4):2808–2821, 2011.

[10] J. Gwizdka. Distribution of cognitive load in web search. *JASIST*, 61(11):2167–2187, 2010.

[11] P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of documentation*, 52(1):3–50, 1996.

[12] H. Joho. Cognitive effects in information seeking and retrieval. In *Proc. CIRSE2009*, 2009.

[13] W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[14] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63:81–97, 1956.

[15] Y. Moshfeghi, L. R. Pinto, F. E. Pollick, and J. M. Jose. Understanding Relevance: An fMRI Study. In *Proc. ECIR2013*, pages 14–25. Springer, 2013.

[16] E. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *Proc. CHI2013*. ACM, 2013.

[17] M. Pike, M. L. Wilson, A. Divoli, and A. Medelyan. CUES: Cognitive Usability Evaluation System. In *EuroHCIR2012*, pages 51–54, 2012.

[18] E. Solovey, P. Schermerhorn, M. Scheutz, A. Sassaroli, S. Fantini, and R. Jacob. Brainput: enhancing interactive systems with streaming fNIRs brain input. In *Proc. CHI2012*, pages 2193–2202. ACM, 2012.

[19] E. T. Solovey, A. Girouard, K. Chauncey, L. M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. J. Jacob. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST2009*, pages 157–166. ACM, 2009.

[20] C. D. Wickens. Multiple resources and mental workload. *The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455, 2008.

[21] C. D. Wickens, S. E. Gordon, and Y. Liu. *An introduction to human factors engineering*. Pearson Prentice Hall Upper Saddle River, 2004.

[22] J. R. Wilson and E. N. Corlett. *Evaluation of human work*. CRC Press, 2005.

[23] M. L. Wilson. Evaluating the cognitive impact of search user interface design decisions. *EuroHCIR 2011*, pages 27–30, 2011.

Dynamics in Search User Interfaces

Marcus Nitsche, Florian Uhde, Stefan Haun and Andreas Nürnberger
Otto von Guericke University, Magdeburg, Germany
{marcus.nitsche, stefan.haun, andreas.nuernberger}@ovgu.de,
florian.uhde@st.ovgu.de

ABSTRACT

Searching the WWW has become an important task in today's information society. Nevertheless, users will mostly find static search user interfaces (SUIs) with results being only calculated and shown after the user triggers a button. This procedure is against the idea of flow and dynamic development of a natural search process. The main difficulty of good SUI design is to solve the conflict between good usability and presentation of relevant information. Serving a UI for every task and every user group is especially hard because of varying requirements. *Dynamic search user interface elements* allow the user to manage desired information fluently. They offer the possibility to add individual meta information, like tags, to the search process and enrich it thereby.

Keywords

Search User Interface, User Experience, Exploratory Search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces.

General Terms

Design, Human Factors, Management.

1. MOTIVATION

Since the launch of the WWW, users accumulated a vast amount of information. With broadband technologies becoming a part of everyday life[1] the WWW offers a great opportunity in terms of learning and education. University courses, for instance, are available online and nearly every topic is handled somewhere in the great amount of blogs, Q&A pages, fora, web pages or databases. Yet there is no map, no guide leading through this vast amount of information. Users need to search for information, to locate the bits fitting to their specific information need, indexing the amount of knowledge available online. Therefore, a proficient tool to analyse the structure of the web and to provide guidance to specific sources of information is needed. This task is accomplished by modern search engines like *Google*[2], *Bing*[3], *Yahoo*[4] and other local or topic centred search engines. By the increase of computational power in smart phones and wider access to online resources the demand for these search tools has risen and the quality of the search terms has changed. Instead of single-query-searches, users tend to request complex answers[5], trying to learn about topics in deep. While the need for information and the expectations of users increased, matching the broader knowledge base contained in the Internet in the last few years. About 300 Mio. websites were added in 2011[6]. Search engines mainly remain the same. This leads to the fact that a *"significant design challenge for web search engine developers is to develop functionality that accommodates the wide variety of skills and information needs of a diverse user population"* [1]. Therefore, this paper proposes the concept of using *dynamic elements* in SUIs, that focus on fluent work flow characteristics, a high grade of interactivity and an adequate answer-time-behaviour.

2. INFORMATION GATHERING

Looking at users' habits in search, they no longer perform simple lookup searches. There is an increasing need to answer complex information needs. Therefore, we mainly consider information gathering processes, searches where users are not familiar with the domain. Users need to refine search queries, branch out into other queries to gain additional understanding and collect results to merge them into a single topic. This kind of search process is called *exploratory search* and is contrary to a *known-item search* task as stated in [2]. Exploratory search processes *"depend on selection, navigation, and trial-and-error tactics, which in turn facilitate increasing expectations to use the Web as a source for learning and exploratory discovery"* [3]. Search tasks are fragmented, consisting of single queries and search requests. The search requests may yield additional data or parts of the final information which in the end form the information requested by the user. While performing such a complex search task, a pattern called *berry picking* [4] can be observed. While reading through a source of data, looking for qualified information the user discovers new *traces* leading to other sources, which have to be handled one after the next. By re-

[1] `http://www.internetworldstats.com/images/world2012pr.gif`, 02.05.2013

[2] `http://www.google.com`, 02.05.2013
[3] `http://www.bing.com`, 02.05.2013
[4] `http://www.yahoo.com`, 02.05.2013
[5] see the 2009 HitWise study for more details: `http://image.exct.net/lib/fefc1774726706/d/1/SearchEngines_Jan09.pdf`, 10.07.2013
[6] `http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/`, 02.05.2013

fining the search and gaining deeper information the user satisfies the initial need for it. These different traces span a map in the end, representing the whole search and its processing. When someone is learning about something this map is refined and expanded. The learner may track back to a certain node and deepen the understanding about it by adding new queries, and therefore new branches. Or he may discard a whole part of the map because it turned out that the contained information was not relevant to him. When the user is satisfied with the gained information this map is encapsulated and represents the whole development of this complex information. According to this concept the result is not a single object. It is a set of sources, representing the learning process for a specific user.

Looking at the current process of information gathering in the Internet there are only two *places*. The Internet itself, containing the pool of existing information, in an unstructured form and a mental model about the information (space) that is constructed. This system may work perfect when dealing with short, exact search queries like *postal code New York City*, but when it comes to complex information needs, where the user needs to access a lot of information and generate more detailed search queries while looming through pages this system reaches it boundaries. The user might retrieve only partial facts. For example, if the user needs explanation of a term used in its initial query. The user is now in need of another place, where he can store information, reorder it and put it into the context of other information pieces.

3. STATE OF THE ART

Looking at *Google*, the most used search engine today [5], the user interface of a modern search engine is mostly static. *Google's* features include some dynamic elements like real time search. For example *"[..] Google Suggest which interactively displays suggestions in a drop-down list as the searcher types in each character of his/her query. The suggestions are based on similar queries submitted by other users."* [1] Dynamic previews of results will be offered when clicking on the double arrow beside a result. But the core of the interface has not changed a lot since its launch in 1997[7]. While adopting fast to new information sources like Facebook and Twitter, Google discarded the adoption of new HCI methods in favour of a clean, slim interface. With increasing touch support on the devices, a richer user interface can be designed to provide the user with immediate feedback and allows haptic interaction with the search process. Some mobile clients take advance of the additional information available, like the iOS search client, which switches to voice queries when the phone is lifted to the head, but there is no full extension of *Google's* search services. While *Google* is an adequate tool for short queries and queries calling for a direct answer, features for deep research on complex topics are missing.

One way to integrate *dynamic elements* into existing SUI infrastructure is to build an overlay. Thereby, dynamic UI utilize existing, well known search engines and provide a benefit by enriching them. This approach is shown in the *Boolify*[8] search engine, which provides a dynamic drag and drop interface on top of *Google's* search engine. This engine is relatively new and was build to promote the understanding of boolean queries. Users *build* a query by dragging jigsaw like parts onto a search surface. These parts contain words (general or exact) and linkers like AND and OR. Additional parts have been added to provide search on a specific page or for synonyms. By adding and linking those parts the user constructs a boolean query which will be submitted to the *Google* search engine. *Boolify* was built for children and elderly. Tests in a third grade technology class showed that children without any knowledge of boolean queries were able to construct complex queries just by pulling them together piece by piece[9]. A similar approach was implemented at *SortFix*[10]. This tool offers the user the *"ability to drag and drop search terms in between several* buckets*"* [6] to in- and exclude them in the query. With a *Standy Bucket* users are *"able to keep track of all [their] inspirations and alternative search words off to the side, ready to be dragged and dropped into your search box if needed."* [6] Another possible use of dynamic interface elements is the weighting of search terms based on their font size as used at *SearchCloud.net*[11]. The ranked keywords are shown in a Tag Cloud like manner and additionally the site shows, based on the ranking, *"the calculated relevance score for each [result]"* [6]. Not only the query building process can be enchanted by dynamic elements, also the presentation of the result can benefit from it. Dynamic side loading can provide the user a lens like view to parts of the result where keywords occur. *Microsoft's WaveLens "[...] fetches a longer sample for the page containing your keywords, without you having to download it."* [8] *Microsoft Research* shows that in a study using WaveLens, presenting the participants with a normal interface and two versions of *WaveLens'* UI (instant zoom and dynamic zoom), *"participants were not only slower with the normal view than the other two, but they were more than twice as likely to give up"* [9]. Another way of result presentation was shown at *SearchMe*[12]: *"Fragmentation into multiple sites, domains and identities becomes a huge distraction. User don't know which site to visit for which purpose, and the lack of consistent, intuitive inter-site search and navigation makes it hard to find content [..]"* [6]. All these dynamic features can be used as a mask over traditional SUIs to extend them. By hiding the dynamic part, dynamic elements can be added to an existing search engine and let the user make a choice which part should be shown and used. The proposed concept is similar to Byström & Hansen's approach in [19].

Issues. Comparing the state of the art with the process of information gathering some issues appear, which may be resolved or at least damped by using of dynamic elements. While collecting information pieces for solving complex questions the user discovers new sources, containing more information. These sources may not form a linear search process every time. Sometimes there will be a split and the user needs to decide which trace to follow first. This issue is also noted in [10]. Today's search engines offer only little support for this. The user needs to save web pages to favourites or organize them himself for later reading. Searching different terms one by one allows users to follow new pages like traces through the Internet. By connecting these traces and setting them into relation the user can retrieve the whole information needed to cover his query. Most modern search engines discard this feature, it is again something the user needs to do by himself. This leads to another more general problem, the *enclosing of search queries*. *Google* for example handles every search term as a new operation. Data is stored, but contains only general information about the user, queries are not related to each other and therefore miss-

[7] `http://www.google.com/about/company/history/`, 02.05.2013

[8] `http://www.boolify.org/`, 02.05.2013

[9] `http://ed-tech-axis.blogspot.de/2009/03/boolified.htm`, 02.05.2013

[10] *SortFix.com*, offline since 11/2011, Firefox plugin: `https://addons.mozilla.org/en-us/firefox/addon/sortfix-Extension`, 02.05.2013

[11] `http://searchcloud.net/`, 02.05.2013

[12] `http://www.searchme.com`, offline since 2009

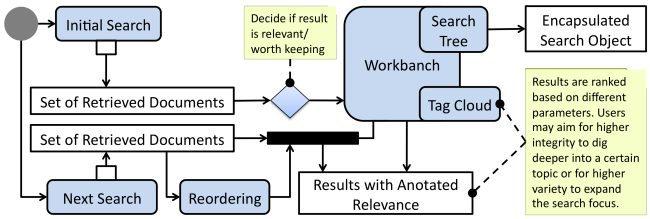


Figure 1: Data flow while refining during search.

ing its broader context. But when learning about a complex topic refining the search query is more important to the user. In the iteration of search processes, to narrow down the mass of information and to tap new sources, the searcher needs to rewrite and modify the query, to link it to other related search tasks. Building a connection between parts of information and evaluating it against each other is a core principle of learning. This leaves the user targeting a broader, intense search, in the need to build a custom solution to extract knowledge and manage it. This is strictly against the guideline for online interfaces which suggests to *"[..] not require users to remember information from place to place on a Web site"* [11] as this is a distraction from the main process of searching and destroys the interaction flow triggered by the search process.

4. COMPOSING A DYNAMIC SUI

The proposed approach shows a design based on today's search engines, enriched with dynamic UI elements to provide a plus for the user. The design includes principles to form web based learning applications [12] to focus on the completion of complex search tasks. By adding dynamic elements internal states can be visualized for the user to give a better overview about the current position in the search process. Furthermore it will allow the serialization of search processes and to step in at every point of the process later on. As stated in Beyond Box Search *"different interfaces (or at least different forms of interaction) should be available to match different search goals"* and *"[t]he interface should facilitate the selection of appropriate context for the search"* [13]. Both of this quality measurements should be regarded when conceptualizing a SUI. The first point will be covered by a modular UI, the user may move, hide and scale elements to fit his current need. The second point is strongly bounded to the use of dynamic items in the UI design. By giving immediate feedback to the user it is easier to classify the current results. The context of the whole search process will be persistent over multiple search queries and provide a method of accumulation parts of the search process into a single object.

Four *features* are proposed and explained in this paper, showing a use-case for dynamic search interfaces and giving a suggestion how this can be accomplished. Together these features build up a mid instance to accumulate into a bigger context for a search process. This clipboard (Fig. 1) reshapes the search process and provide the place to store information between search queries. Instead of trying to accumulate knowledge and information directly the user is able to construct a solution of the search query in this buffer and save it as a complete collection of the information retrieval process.

Reordering. Giving users the opportunity to reorder and therefore to rate a search result is an important step towards dynamics in SUIs. Every result is handled as a single item and can be picked by the user and dropped in another place. The other items reorder fluently, giving user feedback while the user moves on. The SUI holds an array of parameters, which is used to evaluate every item. Possible criteria are *Accuracy*, *Clarity*, *Currency* and *Source Novelty*. These and more criteria are mentioned and explained in [14]. When a user reorders items to fit his preferences the search engine may use the information provided by this ranking to weight the existing parameters to yield better results in the future. The engine will be able to present results ranked according to the user's preference. This can be done for all users and also search process wide, as some search tasks require documents and papers while others may focus on web pages or media. This addition to classical user interfaces can make great use of the up-trend for touch based devices, in 2012 89% of mobile phones and smart-books support touch [15]. Designing the SUI responsive to touch and gesture is maybe one of the most natural solutions for human computer interaction and adds an amount of possible actions based on gestures.

Workbench. The workbench targets the issue of loosing information while switching between different searches. It adds a third place to the proposed search process, located outside of the search scope but still related to it. The user may drop queries here to keep them throughout the whole search process. When entering a query, indicators show how relevant items on the bench are. This allows the user to classify new results in terms of integrity towards already selected snippets. The workbench acts as a buffer between search queries, adding a broader context to every entry. Like a frame, it contains information exclusively attached to the current search process, leading to the possibility of customization and user centred search environments. When the user switches between queries he can immediately determine how well the new results fit into already selected items. This allows identifying false positive as well as exploratory search [16] results. Users may just enter queries that lead to a peripheral topic and check the indicators whether the result is relevant to his initial information.

Tag Cloud. The tag cloud is another feature to guide the user in the search process. As shown in [17] a tag cloud supported retrieval system can increase the find rate of adjacent data nodes by nearly 15%. When adding an item to the workbench its most relevant tags are extracted and visualized in the tag cloud. It is able to show how often a tag occurs and how different tags are related to each other. When entering a new search query the tag cloud displays the relevant tags and reorders the cloud to revolve around the current tags. By combining distance and size of the entered tag with their direct neighbours the user can directly spot how homogeneous its current query is in terms of the whole process. The tag cloud can also use the existing tags to show the user other closely related tags and suggest query refinement based on tag proximity. Colours can indicate the state a tag is currently in. A possible color scheme for western culture can be based on the three colors used in traffic lights. The concept of three-coloured traffic lights also work for color-blind people, since they do have a given position. Therefore, we also use second coding paradigm: form. A green triangle is proposed for tags resulting from the current query, which are contained in the overall tag cloud spanned by the workbench. An orange circle indicates a warning for tags, either in the current query result or the bench, which are not related to the rest of the cloud. A red square is avoided for the reason that uncontained tags may not be bad, they can lead to a new direction or add a reasonable value to the whole search process. The tags are scaled depending on their frequency. When the user selects any item from the bench or the search result the corresponding tags are centred. The other tags are located based on their coherence with the selected tags; closer means the tag is in a direct relation to the selected item. A user can quickly

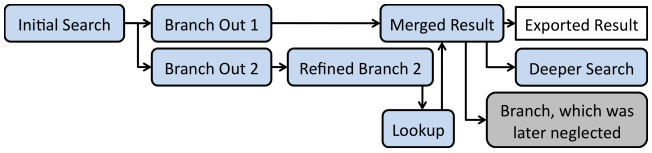


Figure 2: Search map, representing the search process.

check the integrity of his search process by looking at the tag cloud. A slim, packed cloud means the results are all related to each other, an open, wide cloud indicates a broad result field, covering many aspects. False positives may be filtered out, when enough items exist, as they stick out the rest of the cloud.

Search Map Support. The search map (Fig. 2) acts as a representation of the whole search process, by storing every query and following up querying and visualize it in a chronological order. The user may select single nodes in the map to get into the state of search process at this moment and refine it. The map provides a kind of top view to the path of the search and shows where the user branched out into new queries. It allows the user to cut off nodes and whole branches if they are not needed any more to fulfil the need for information. As it contains every action and some data in the current search process, the search map might be serialized and stored to retrieve the search process later on. With this map at hand a user can save whole search tasks just like he saves favourite web pages. He can step back into the process at any time and reconstruct the whole learning process or correct parts of the search which has proven to be not correct. This kind of *Story Telling* helps to visualize the given data, *"[...] lead to findings, which prompt actions [...] [and] can indicate the need to forage for new data."* [18] The search map [7] features two ways of expanding. The user may follow a result to expand it vertically. The result is added as a new node and resides in the map until it is processed further. When the user selects an existing node he steps back to the vertical position of this node and can now branch out horizontally. This deals with an issue of berrypicking [4], where the new sources has to be processed one by one. While not abolishing this the search map provides a visual representation to simulate parallelism. The map also allows scoping of the analysis by creating a horizontal or vertical bound. Only tags and items inside this bound will be considered, the rest is greyed out. This allows the user to dig deep into a certain topic (small vertical bounds) or create a better understanding of a certain term and add more results to a certain query (horizontal boundary). This can help the user to concentrate on smaller pieces of a big search process and to narrow down problems one by one.

5. CONCLUSION

This paper has shown certain design flaws of today's search engines and some proposed dynamic design principles to counter them. The application of the envisioned elements can extend a search engine towards a software capable of complex research tasks. With the current up-trend of online learning this unlock a new way of using them. The surplus resides not only in the dynamic and vivid interface, it prepares a whole new tier of online search solutions. The process of learning can be preserved and shared with others. One can come back at any time, jump right into the saved search process and reconstruct the development of certain knowledge. With this tool chain at hand learning becomes a social and an integrative part of the WWW. The next step in deploying dynamic elements into search user interfaces would be prototyping them. Design snippets need to be tested for usability and acceptance in the real world. Starting as overlays and additional feature of existing search engines may develop and emerge into independent solutions.

Acknowledgement

Part of the work is funded by the German Ministry of Education and Science (BMBF) within the ViERforES II project (01IM10002B).

6. REFERENCES

[1] Sandvig, J. C., Deepinder B.: User Perceptions of Search Enhancements in Web Search. In: J. of Comp. Inform. Syst. 52, no. 2, 2011.

[2] White, R. W., Marchionini, G.: A Study of Real-Time Query Expansion Effectiveness. In: SIGIR Forum 39, 2006.

[3] Marchionini, G.: Exploratory Search: From Finding to Understanding. In: Comm. of the ACM 49, 4.2006.

[4] Bates, Marcia J.: The design of browsing and berrypicking techniques for the online search interface. Univers. of Calif. at L.A., 1989.

[5] Purcell, K., Brenner, J., Rainie, L.: Search Engine Use 2012. In: Pew Internet & American Life Project, 2013.

[6] Bates, M. E.: Make Mine Interactive. Vol. 31, Issue 10, p. 63, 12/2008.

[7] Heer, J., Viégas, F. B., Wattenberg, M.: Voyagers and voyeurs: Supporting asynchronous collaborative visualization. In: Commun. of the ACM, 52, No. 1, pp. 87–97, ACM, New York, NY, USA, 01/2009.

[8] MS Research: Cutting Edge. New Scientist 181, no. 2434, 2004.

[9] Paek, T., Dumais, S., Logan, R.: WaveLens: A new view onto Internet search results. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04), pp. 727–734, 2004.

[10] Morville, P., Callender, J.: Search Patterns - Design for Discovery. In: O'Reilly, 2010.

[11] U.S. Department of Health and Human Services, Research-Based Web Design and Usability Guidelines. Washington, D.C.: GPO, n.d.

[12] Jayasimman, L., Nisha Jebaseeli, A., Prakashraj, E.G., Charles, J.: Dynamic User Interface Based on Cognitive Approach in Web Based Learning. In: Int. J. of CS Iss. (IJCSI), 2011.

[13] Buck, S., Nicholas, J.: Beyond the search box. Reference & User Services 51(3), pp. 235-245, 2012.

[14] Beresi, U. C., Kim, Y., Song, D., Ruthven, I.: Why did you pick that? Visualising relevance criteria in exploratory search. In: Int. J. on Dig. Lib. 11 (2), pp. 59–74, 2010.

[15] Lee, D: The State of the Touch-Screen Panel Market in 2011. In: Walker Mobile, LLC, SID Information Display Magazine, 3.2011.

[16] White, R. W., Kules, B., Drucker, S. M., schraefel, m. c.: Supporting Exploratory Search. In: Comm. of the ACM 49, 4.2006.

[17] Trattner, C.: QUERYCLOUD: Automatically linking related documents via search query (Tags) Clouds. In: Proc. of the IADIS Int. Conf. on WWW/Internet, 2010.

[18] Mackinlay, J. D.: Technical Perspective: Finding and Telling Stories with Data. In: Comm. of the ACM 52, 2009.

[19] Byström, K., Hansen, P.: Conceptual framework for tasks in information studies: Book Reviews. In: J. Am. Soc. Inf. Sci. Technol., Vol. 56, 10, pp. 1050–1061, John Wiley & Sons, Inc., New York, NY, USA, 2005.

SearchPanel: A browser extension for managing search activity

Simon Tretter
University of Amsterdam
Amsterdam, The Netherlands
s.tretter@gmail.com

Gene Golovchinsky
FX Palo Alto Laboratory, Inc.
3174 Porter Drive
Palo Alto, CA
gene@fxpal.com

Pernilla Qvarfordt
FX Palo Alto Laboratory, Inc.
3174 Porter Drive
Palo Alto, CA
pernilla@fxpal.com

ABSTRACT

People often use more than one query when searching for information; they also revisit search results to re-find information. These tasks are not well-supported by search interfaces and web browsers. We designed and built a Chrome browser extension that helps people manage their ongoing information seeking. The extension combines document and process metadata into an interactive representation of the retrieved documents that can be used for sense-making, for navigation, and for re-finding documents.

1. INTRODUCTION

Broder *et al.* [3] proposed a taxonomy of web search that included transactional and navigational searches in addition to the more traditional (from an IR perspective) informational searches. To this taxonomy we might add re-finding [17] [5], the task of locating a previously-found document. From a theoretical perspective, it is not clear whether refinding is a different kind of search activity or an orthogonal dimensions. Regardless, while major web search engines offer simple and efficient interfaces for navigational and transactional searches, relatively little support is available for more complex informational search or re-finding.

These seemingly neglected activities are not unimportant, however: Teevan *et al.* [17] reported that 39% of queries are re-finding queries; furthermore, 20-30% of searches represent open-ended informational needs [13]. Related, Qvarfordt *et al.* [11] found query overlap rates of 50-60% in exploratory search, and suggested that awareness of this overlap may be useful in supporting more efficient searching behavior. Thus we decided to explore ways in which searchers' interactions with search engines could be enhanced to support these more complex information-seeking tasks.

We created a web browser extension that enriches common web search engine interfaces and addresses important deficits with respect to open-ended (exploratory) search and re-finding. Our extension visualizes search results to help users find the right document or documents by visualizing metadata of the retrieved pages.

Following Golovchinsky *et al.* [7] we distinguish *document metadata* from *process metadata*. Document metadata – dates of publication, titles, hosting web sites, etc. – are basic characteristics of documents that are independent of the means by which these documents were retrieved. Process metadata, on the other hand, characterize aspects of documents in relation to the searcher's activity: how many times was a document retrieved, whether it was viewed before, etc. This kind of information can help searchers to remember, understand and plan their search processes.

The browser plugin enhances the searcher's ability to use process metadata to understand their search results and to plan subsequent activity by displaying surrogates for the current set of retrieved documents. We represent prior retrieval state, whether a document was opened, and whether it was bookmarked in an integrated overview that appears at the side of the browser window. We also make it possible for searchers to examine multiple documents without returning to the search results or using multiple tabs.

The remainder of this paper is organized as follows: we review the relevant related work, describe the browser extension, and conclude with a discussion of the design space.

2. RELATED WORK

There are two broad categories of related work: the management of search history and the representation of search results. Refinding has received increasing attention recently. While the browser implements some history mechanisms, these are typically not well-suited to users' needs [15]. Elsweiler and Ruthven [5] described different patterns of refinding; Teevan [16] proposed a mechanism for merging previously-found and newly-retrieved documents. More explicit management of search history has also been investigated in the literature; see [7] for a succinct summary.

Information overload due to large numbers of results is a common problem in information seeking [2]. This problem can be addressed in a variety of ways. MetaSpider [4] uses a 2D map to display and classify retrieved documents. Grokker [8] uses nested circular and rectangular shapes to present results and also shows them in a hierarachical grouped way. Sparkler [12] uses a star plot for the result presentation, where every star represents a document.

One potential issue with the systems above is that the overall organization of the interface itself may induce usability problems. Complex interfaces allow more individual settings to be specified by a user, but simple interfaces allow a broader spectrum of users to use them. This tradeoff is not trivial to handle, and as we see nowadays, most Web search interfaces tend to be quite simple.

Supporting the searcher's decision making process can be crucial for effective search performance for complex information needs. This support can take the form of enhanced surrogates for documents. One type of information often used for this purpose is document metadata (author, date,

images of the document, etc.). Even *et al.* [6] has shown that the decision making process can be highly improved by adding process metadata (in our case information that is related to the search process) to the user interface. Research has shown that presenting simple tasks in a slightly different way may help the user to understand how the search is performing and what can be done to gain better results [18]. One common example of incorporating process metadata in web browsers is the practice of changing the color of a traversed link anchor.

Spoerri [14] showed that users can benefit from different or additional visualizations of web search results. However, none of the techniques above have been integrated by major search engines into their main interfaces. In some cases, extension developers have enhanced the user experience of web search. Examples include: SearchPreview[9] that fetches screen shots of the result pages and shows them directly next to the each search result. Bettersearch[1] is a Firefox extension that performs a similar task, but also enriches the result page with more features and links. For example, this extention allows users to open a result in a new tab, or adds links to a search result to quickly show the web page on the "Wayback Machine"[1]. WebSearch Pro [10] is also a Firefox extension that adds the ability to look up a text by highlighting it on a page. Another feature is drag&drop zones to search for things directly from any website.

3. BROWSER EXTENSION

To compensate for the deficiencies of SERPs we created a browser extension called SearchPanel. This extension combines document and process metadata in a visual representation of search results to help people manage their information seeking. We chose the browser extension approach rather than creating a proxy for several reasons. While both offer the potential of parsing and augmenting SERP and document pages, a browser extension has some advantages. It scales better with respect to storing user history data. It ensures a higher level of data privacy, since data that might potentially reveal user interests (e.g., query keywords, selected URLs, etc.) can be logged as hashed values. Finally, it has access to bookmarks and local browsing history.

3.1 Design space

When performing search tasks, searchers may need different kinds of information to support their information seeking. We represent the design space as consisting of three categories of activities: search activity, navigation activity, and organization activity.

Historically, web UI support for the search process, or search activity, has been focused on query formulation and understanding the current query. Web browsers offer limited support for comparing current results set with earlier activity by marking the visited status of documents.

When engaged with a search task, users need to shift their attention between the SERP and the retrieved pages. In some cases, the searcher does not find the desired information in a retrieved document, but rather in links to other documents containing relevant information. This navigation activity can be an important part of the information seeking process.

[1]The Wayback Machine is a service that provides access to archived and historical versions of web sites.

Table 1: Design space: Activities and supporting features related to document and process metadata. "Doc" refers to document metadata and "Proc" to process metadata.

Activity	Feature	Doc	Proc
Search	perform search	yes	no
	switch engine	no	yes
	results list	yes	no
	visit status	no	yes
	visualize no. of visits	no	yes
Navigation	access results	-	-
	mark current result path	no	yes
	identify results: preview snippet	yes	no
	identify results: favicon	yes	no
Organization	bookmarking	no	yes
	organize bookmarks	no	yes

When searchers find useful web pages, they may wish to save those documents for future access. More specialized search engines sometimes support this capability directly, but it is most often supported only by the browser's bookmarking capability.

We can consider these search and sense-making activities in light of the kinds of information required to satisfy them. In particular, Table 1 shows when document and process metadata might be pertinent for the different categories of search activities. A representation of the number of visits to a retrieved result (process metadata) could be used by a searcher to decide how to interact with that result. In a refinding sub task, for example, searchers might want to ignore newly-found documents or pages that were not opened.

The purpose of the search panel is to complement the SERP and to be available when exploring search results; we wanted the design to be simple and unobtrusive but still convey useful information. Some features (e.g., organizating bookmarks) listed in Table 1 are too complex to be integrated into the extension. Others, such as favicons, while seemingly trivial, may still provide useful information for navigating search results.

3.2 Implementation

SearchPanel displays automatically on the right side of the browser window when it is enabled (Figure 1). The right side of the content page has been chosen because this location is frequently free of document content. In cases of overlap, its vertical position can be adjusted manually to accommodate page content that may be occluded.

SearchPanel displays immediately after a search has been performed on a supported web search engine (currently, they are Google, Google Scholar, Yahoo, Bing and Microsoft Academic Search). SearchPanel remains visible even if the searcher follows links from retrieved documents. In addition, searchers can return directly to the original query, or re-run it on a different search engine.

A short tutorial page is displayed at installation, and can also be reached through the option menu. This page also allows logging (see 3.2.4) to be disabled, and can be used to delete the recorded history.

3.2.1 Document metadata

SearchPanel displays several kinds of document metadata. Documents are represented by bars arranged in order corre-

Figure 1: SearchPanel control annotated to show important aspects. (1) search engine selector; (2) bar representing a newly-found page; (3) favicon representing the site from which the page was retrieved; (4) bar representing page that has been visited; (5) highlighted bar based on cursor position; (6) bookmark indicator; (7) currently-selected page.

sponding to the retrieved list; clicking on a bar is equivalent to clicking on a link on the SERP. Almost all websites have icons (favicons) to help re-identify the web page quickly; these icons are shown to the right of the bar (see Figure 1, item (3)). A tooltip with the title of the document is added to each bar as well. We considered identifying other metadata such as document MIME type, but that would incur the overhead of a separate HTTP request for each document. At least initially, we chose not to pursue this strategy.

3.2.2 *Process metadata*

Process metadata is also incorporated into SearchPanel. First, the icon of the search engine that ran the search is highlighted in the top bar (item (1)). Other icons represent available comparable search engines. Clicking on one of these icons re-runs the query with the selected search engine. Search engines are grouped into two categories (web search and academic research) and only the relevant ones are shown. The current selection (highlighted with a black border) links back to the search result page if the user navigates to one of the retrieved documents.

Each bar can have one of three different colors, depending on the link history. If a link has never been retrieved before, the state of the link is "new" and the color will be teal. Results that have been retrieved by prior queries but have not been clicked on are colored blue. Visited links are colored violet. The local browser history is examined to retrieve the link status. This allows us to incorporate page views that occurred before SearchPanel was installed.

Each bar's length reflects the frequency of retrieval of the corresponding page. The more frequently a page has been retrieved, the shorter the bar gets (item (3)). The retrieval history is stored locally in the browser for privacy reasons and can be deleted through SearchPanel's option page.

In SearchPanel, the bookmarking function serves two purposes (item (6) in Figure 1). First, searchers can click on

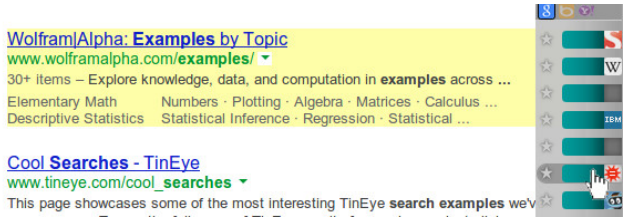


Figure 2: Highlighting of snippet on the SERP when mousing over SearchPanel.

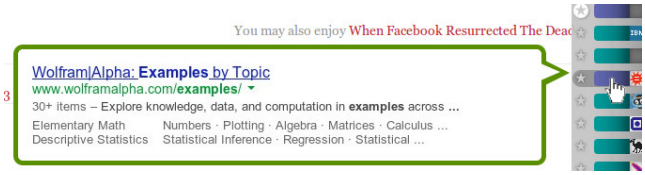


Figure 3: Snippets of other pages are shown on a document page when mousing over SearchPanel.

the star to bookmark the corresponding page. Second, previously bookmarked documents in the SERP will show a yellow star next to them. This allows to re-find a web page quicker, as the user does not need to navigate to a document to know if they have previously bookmarked it.

3.2.3 *Navigational support*

The selection indicator (see item (7) in Figure 1) indicates the currently-selected result page. If a link on a result page is clicked, the page indicator will stay on the last retrieved document page to indicate that navigation started with it. Hovering over the result highlights the associated bar (item (5)), and also highlights the corresponding snippet in the SERP (Figure 2); the SERP is scrolled as necessary to bring highlighted snippet into view. Conversely, when the mouse is over a snippet on the SERP, the related bar jiggles left-right to reinforce the connection between the two.

When the user navigates off the SERP to a search result, SearchPanel remains active. Clicking on bars navigates among the retrieved documents, bypassing the intermediate step of reloading the search results. When the mouse is over a bar in SearchPanel, the SERP snippet of that result will be shown. This can be seen in Figure 3, where a preview of the Wolfram Alpha snippet is shown. If the snippet is not available, a tooltip with the document title is shown instead. Both of these features should make it easier and more efficient to navigate the search results without necessarily creating a large number of tabs in the process.

3.2.4 *Logging*

The extension was created to study people's information seeking behaviors. The goal of the project is to understand how people use the web when looking for information to improve their search experience. Therefore logging of user activity was necessary. To encapsulate it from the basic functionality it was designed as plugin that could be connected or disconnected from SearchPanel. It collects information related to the use of SearchPanel for the purposes of statistical analysis of patterns of behavior.

To maximize searchers' privacy, no personally-identifying information is saved. Queries and found URLs are recorded as MD5-hashed values only. This allows us to identify re-

curring queries and documents, without being able to read the content of the query or to observe which pages people view. Specifically, the following information is recorded:

- The IP address and the time the event was logged
- When a search result was clicked and where this happened (SearchPanel or SERP)
- Hash strings that represent the queries and found web pages.
- Time spent with the mouse on different interface parts (SearchPanel vs SERP)
- Various actions related to the extension (adding bookmarks by clicking the start, moving it, etc.).

4. NEXT STEPS

After an in-house pilot deployment, SearchPanel has been made available through the Google Chrome store. The goal of the deployment is to understand whether the extension helps people with their search tasks, and to assess the relative utility of document vs. process metadata. We also expect to collect a dataset that characterizes people's browsing and searching behaviors in terms of patterns of retrieval and re-retrieval, search result navigation, etc.

5. CONCLUSIONS

Web search engines are used for many different kinds of search tasks. While navigational and transactional uses of search engines are well-supported by current interfaces and algorithms, searchers are left to their own devices for more open-ended information seeking and re-finding. We created a Google Chrome browser extension to help people manage their search activity. We explored the design space of document and process metadata related to the wide range of activities searchers may engage in during information seeking. The extension keeps track of retrieval, page visits, and bookmarking, and integrates traces of these activities with document metadata to give people a more complete impression of their search activity. An upcoming deployment will explore the effect that this extension has on how people interact with search results.

6. REFERENCES

[1] ABAKUS. Bettersearch a firefox addon for enhancing search engines. `http://mybettersearch.com/`, 2010. [Online; accessed 06/06/2013].

[2] Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[3] Broder, A. A taxonomy of web search. *SIGIR Forum 36*, 2 (Sept. 2002), 3–10.

[4] Chen, H., Fan, H., Chau, M., and Zeng, D. Metaspider: Meta-searching and categorization on the web. *Journal of the American Society for Information Science and Technology 52*, 13 (2001), 1134–1147.

[5] Elsweiler, D., and Ruthven, I. Towards task-based personal information management evaluations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 23–30.

[6] Even, A., Shankaranarayanan, G., and Watts, S. Enhancing decision making with process metadata: Theoretical framework, research tool, and exploratory examination. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on* (2006), vol. 8, IEEE, pp. 209a–209a.

[7] Golovchinsky, G., Diriye, A., and Dunnigan, T. The future is in the past: designing for exploratory search. In *Proceedings of the 4th Information Interaction in Context Symposium* (New York, NY, USA, 2012), IIIX '12, ACM, pp. 52–61.

[8] Hong-li, Q. A novel visual search engines: Grokker. *Journal of Library and Information Sciences in Agriculture 8* (2008), 047.

[9] KG, P. U. . C. Searchpreview, the browser extension previously known as googlepreview. `http://searchpreview.de/`, 2013. [Online; accessed 06/06/2013].

[10] Martijn. Web seach pro, search the web the way you like... `http://websearchpro.captaincaveman.nl`, 2012. [Online; accessed 06/06/2013].

[11] Qvarfordt, P., Golovchinsky, G., Dunnigan, T., and Agapie, E. Looking ahead: Query preview in exploratory search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in Information Retrieval* (New York, NY, USA, 2013), SIGIR '13, ACM.

[12] Roberts, J., Boukhelifa, N., and Rodgers, P. Multiform glyph based web search result visualization. In *Information Visualisation, 2002. Proceedings. Sixth International Conference on* (2002), IEEE, pp. 549–554.

[13] Rose, D. E., and Levinson, D. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web* (2004), ACM, pp. 13–19.

[14] Spoerri, A. How visual query tools can support users searching the internet. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on* (2004), IEEE, pp. 329–334.

[15] Tauscher, L., and Greenberg, S. How people revisit web pages: empirical findings and implications for the design of history systems. *Int. J. Hum.-Comput. Stud. 47*, 1 (July 1997), 97–137.

[16] Teevan, J. The re:search engine: simultaneous support for finding and re-finding. In *Proceedings of the 20th annual ACM symposium on User interface software and technology* (New York, NY, USA, 2007), UIST '07, ACM, pp. 23–32.

[17] Teevan, J., Adar, E., Jones, R., and Potts, M. A. S. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 151–158.

[18] Wang, T. D., Deshpande, A., and Shneiderman, B. A temporal pattern search algorithm for personal history event visualization. *Knowledge and Data Engineering, IEEE Transactions on 24*, 5 (2012), 799–812.

A System for Perspective-Aware Search

M. Atif Qureshi[*†◇], Arjumand Younus[*†◇], Colm O'Riordan[*], Gabriella Pasi[◇], Nasir Touheed[†]
[*]Computational Intelligence Research Group, Information Technology, National University of Ireland, Galway, Ireland
[◇]Information Retrieval Lab, Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
[†]Web Science Research Group, Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan
muhammad.qureshi,arjumand.younus@nuigalway.ie, colm.oriordan@nuigalway.ie, pasi@disco.unimib.it, ntouheed@iba.edu.pk

ABSTRACT

Traditional search engines fail to capture the notion of "perspective" in their search results and at times present the results skewed towards a particular topic. Under most of these cases even query reformulation fails to retrieve desired search results and the underlying reason for such failure is often the bias within the document collection itself (e.g., news articles). A perspective-aware search interface enabling users to look into search results for some "perspective" terms may be of great use for certain information needs. In this paper we describe such a system.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Human Factors, Performance

Keywords

Perspective, Wikipedia, Bias

1. INTRODUCTION AND RELATED WORK

It is often the case that when using a search engine for information seeking users have an underlying intent [1]. Traditional search interfaces fail to capture the user intent for certain topics and at times return results that may be skewed towards a certain perspective. Here, perspective as defined by the Oxford Dictionary refers to a "point of view"[1] within the search results that may or may not be something what user is looking for. We explain further through the following motivating examples:

- Consider the case of a user who wishes to find more about a certain event (say, a bomb attack in a certain region). The search results returned contain a majority of news reports blaming Islam relating it with terrorism in most of the cases. This prompts the user to explicitly evaluate how much Islam is related to terrorism in the returned search results.
- Consider the case of a user who wishes to find out about roles and rights of women in Islam but the search engine returns articles that contain a high amount of terms highlighting oppression against women instead of women rights and roles. In this case the user is prompted to check the correlation between women and oppression within the search results that have been returned.

Note that the perspective given by most search results (*Islam* in our motivating example (1) and *oppression* in our motivating example (2)) may or may not be aligned with the user's query intent. In case of search results not being aligned with his/her query intent he/she may be interested in observing the amount of perspective tendencies in various news reports.

This paper proposes the concept of a "perspective-aware" search interface that enables the user to explicitly analyse search results for information from a particular perspective with respect to an issued query. To the best of our knowledge, previous research within Human-Computer Interaction and Information Retrieval has failed to capture the notion of "perspective" within the information retrieval process. Early research related to Interactive Information Retrieval by Belkin [2] and Ingwersen [6] suggests the integration of cognitive aspects within the information retrieval process: in line with this suggestion we argue for incorporating the essential cognitive element of "perspectives"[2] within the search engine interface.

Recently the information retrieval community has turned attention to diversification of search results which aims to tackle the issue of query ambiguity on the user side [8]. However, even when formulating a non-ambiguous query users may have an intent that influences the perspective from which the query terms can be interpreted in a text; in case of

[1]This may also be seen as topic drifts within a document.

[2]According to Wikipedia the definition of perspective states the following: "Perspective in theory of cognition is the choice of a context or a reference (or the result of this choice) from which to sense, categorize, measure or codify experience, cohesively forming a coherent belief, typically for comparing with another."

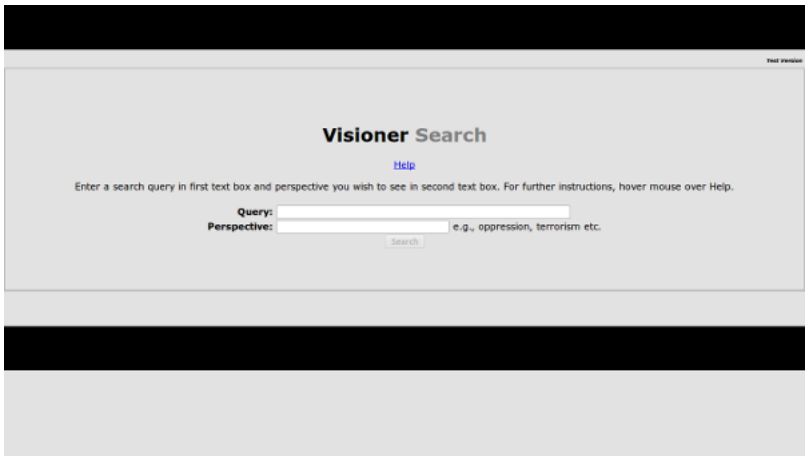


Figure 1: Entry Point of Perspective-Aware Search Interface

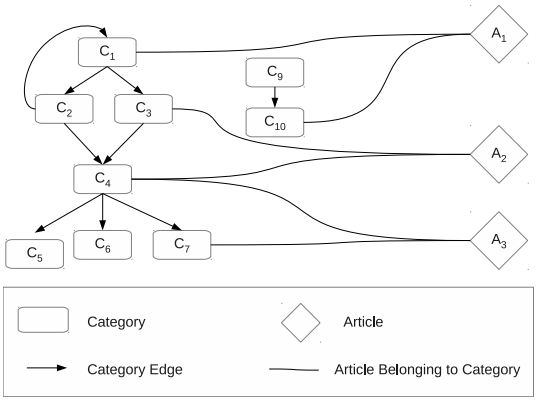


Figure 2: Wikipedia Category Graph Structure along with Wikipedia Articles

perspective mismatch between the user intent and the documents returned in first positions by a search engine, users may find the retrieved results annoying or subjective to a non-agreed perspective [7]. One may argue that a query reformulation technique could be employed to tackle this problem [5]; e.g. considering the motivating example (2), the user could issue a reformulated query such as "roles and rights of women in islam". However, for some topics query reformulation may fail to retrieve the desired search results, and the underlying reason for such failure is often the bias within the document collection itself (e.g., news articles) [10]. Under such a scenario it would be interesting to provide a search interface that would enable the users to look into the search results for some "perspective" terms and we describe such a system in this paper.

2. PERSPECTIVE-AWARE SEARCH INTERFACE AND IMPLEMENTATION DETAILS

This section presents the essential details of the proposed perspective-aware search interface along with the underlying implementation details. We keep the interface as simple as possible on account of research suggesting users' reluctance in switching from a simple search form [3]. Figure 1 shows the entry point of the interface which resembles the standard type-keywords-in-entry-form interface with the augmentation of an additional input text box for entry of perspective terms.

The underlying perspective detection algorithm makes use of the encyclopedic structure in Wikipedia; more specifically the knowledge encoded in Wikipedia's graph structure is utilized for the discovery of various perspectives in documents returned by the search engine. Wikipedia is organized into categories in a taxonomy-like[3] structure (see Figure 2). Each Wikipedia category can have an arbitrary number of subcategories as well as being mentioned inside an arbitrary number of supercategories (e.g., category C_4 in Figure 1 is a subcategory of C_2 and C_3, and a supercategory of C_5, C_6 and C_7.) Furthermore, in Wikipedia each article can belong to an arbitrary number of categories, where each category is a kind of semantic tag for that article [11]. As an example, in Figure 2, article A_1 belongs to categories C_1 and C_{10}, article A_2 belongs to categories C_3 and C_4, while article A_3 belongs to categories C_4 and C_7. It can be seen that the articles and the Wikipedia Category Graph are interlinked and our system makes use of these interlinks for the detection of a certain perspective within a document retrieved by the search engine.

2.1 Underlying Algorithm

The underlying perspective detection algorithm within our system requires the perspective term/phrase to match the title of a Wikipedia article. This may seem to impose a cognitive load on the user at search time. However, this is not the case: as shown in Figure 3 the entered text automatically turns green when a certain user-specified perspective term matches the title of a Wikipedia article, and symmetrically the entered text automatically turns red in case of a mismatch.

Once the perspective term is entered correctly the system fetches the Wikipedia article corresponding to the perspective term referred to as *Seed Perspective Article (PA_{seed})* along with the categories to which it belongs and we use

[3]We say taxonomy-like because it is not strictly hierarchical due to the presence of cycles in the Wikipedia category graph.

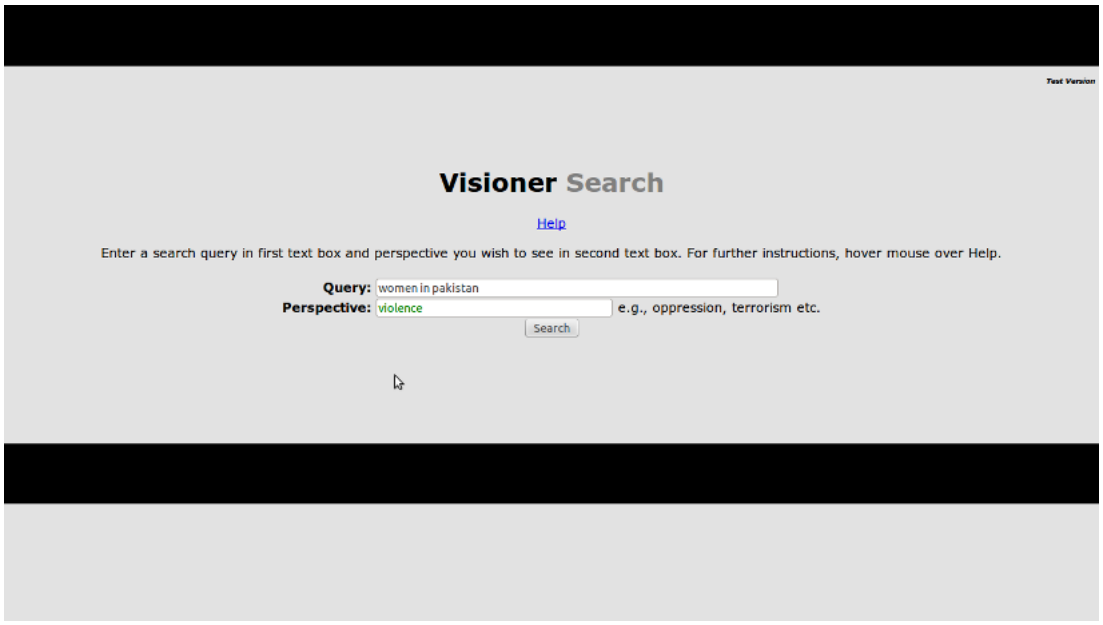


Figure 3: Automatic Text Color Changing to Test Match of Perspective Term with Wikipedia Article Title

PC_0[4] to refer to these categories. After fetching of Wikipedia categories in PC_0, the system retrieves sub-categories of PC_0 until depth 2 i.e., PC_1 and PC_2[5] and collectively these categories related to PA_{seed} are referred to as PC (where PC is union of PC_0, PC_1 and PC_2.). Next, the set of all articles within the Wikipedia category set PC is retrieved and we refer to this set as *Expanded Perspective Article Set* ($PA_{expanded}$). The system then retrieves all categories associated with the set $PA_{expanded}$ which we refer to as WC; note that PC is a subset of WC. Finally, the intersection between PC and WC is retrieved which is a set of categories representative of the domain of the perspective term originally input by the user, we refer to this set of representative categories as RC.

After building the Wikipedia category sets as defined above[6] i.e., PC, RC and WC we match variable-length n-grams within a document with articles in the set $PA_{expanded}$, and we check for cardinality of RC and WC. The cardinality scores along with n-gram frequencies are used to compute a perspective score for each document.

2.2 Search Results Presentation

The perspective scores computed in section 2.1 are displayed within the search results, and based on the perspective score a document receives , we define four levels of perspective adherence as follows: a) High, b) Medium, c) Low, and d) Neutral. Moreover, in case of documents with high, medium and low scores we also report the top-scoring perspective terms that were extracted using the Wikipedia graph structure as explained previously. A sample search corresponding to search query "india pakistan relations" and "terrorism" is shown in Figure 4. As evident from the top search result, there is a high perspective of terrorism within the returned document and perspective terms that our algorithm fetches are as follows: a) the war on terrorism, b) ayman al zawahiri, and c) osama bin laden.

[4]These are basically perspective categories at depth zero.
[5]These are basically perspective categories at depth one and two.
[6]The set building phase is performed through a custom Wikipedia API that has pre-indexed Wikipedia data and hence, it is computationally fast. For details `http://www3.it.nuigalway.ie/cirg/prj/WikiMadeEasy.html`

3. DISCUSSION

There have been many efforts in the information retrieval research to present to users information regarding the relationship between the query and the answer set and the query and document collection. Capturing this information during the retrieval process provides the user with much valuable information (e.g. whether a term is overly specific, or whether a term is ambiguous etc.). Various attempts have been made to tackle this problem, ranging from the definition of snippets to the definition of approaches to cluster search results (Clusty.com), to the presentation of diversified search results in the first position of the ranked list offered to the users. Recently there has been a resurgence of interest in defining visualization techniques of search results that offer an effective and more informative alternative to usual and scarcely informative ranked lists. Pioneer visualization systems are represented by Tilebar [4], and Infocyrstal [9], and these attempts have been aimed to provide the user with more information than that provided by the traditional ranked list.

This additional information can help the user in their search task (e.g. allowing them to navigate the collection more easily or providing evidence to allow the user to reformulate their query more efficiently).

Our proposed system, although related in that we also attempt to give the user an insight into the answer set and its relation to the query, differs in a fundamental manner. Our system, we posit, allows the user to gain insight into the answer set and its relation to the query, but moreover, allows to the user to gain an insight into a *perspective* inherent in the answer set. Our system uses an external and collectively created knowledge resource (which is less likely to be biased

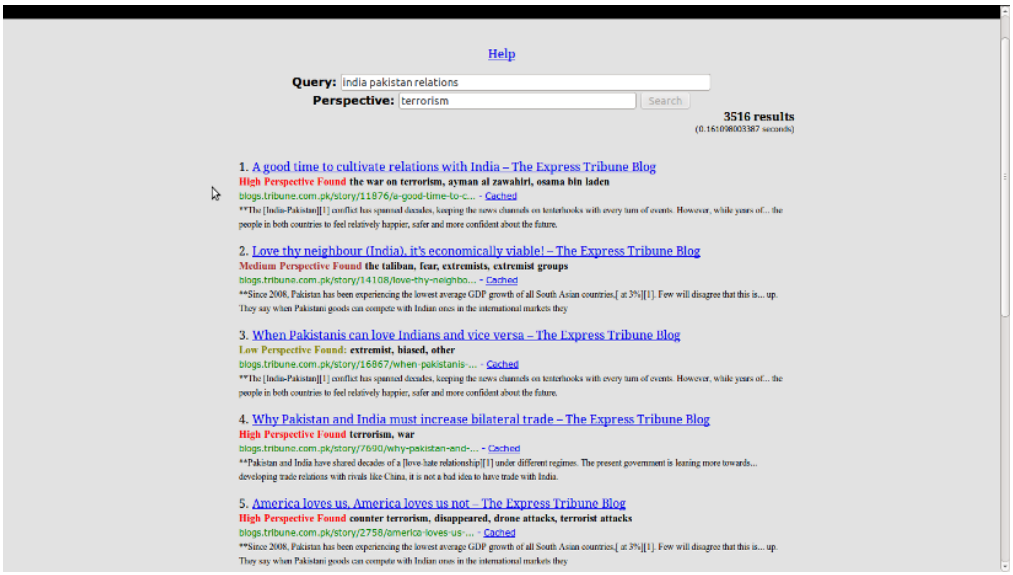


Figure 4: Search Results within Perspective-Aware Search

in a given direction) to obtain extra terms to represent the perspective of interest to the user. This knowledge (perspective term and related terms) does not modify the query (as would an additional query term), but is instead used to highlight the presence of a perspective in the answer set.

In this paper we have proposed a novel approach for capturing the relationship between a user's query and the returned answer set. We do not rely on evidence in the document collection or the query stream, but rather instead extract terms from an external source of evidence to help users quickly see the presence of a particular perspective in the document collection and answer set.

4. FUTURE WORK

Having built the system and undertaken preliminary user evaluations[7], we aim at undertaking a complete and systematic review of the approach. This will comprise a number of separate user evaluation tasks. The initial experiments will involve comparing our search approach with and without the perspective-aware component over a number of tasks to see if the additional context and information provided by our perspective aware system aids the users in a range of information-seeking tasks. Our second planned experiments will be focussed on persons seeking information from newspaper articles, a domain wherein a degree of bias often exists. We wish to explore the users' experience with regards to any perceived bias in the considered corpora.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, 2009.

[2] N. Belkin. Cognitive models and information transfer. *Social Science Information Studies*, 4(2âĂŞ3):111 – 129, 1984.

[3] M. A. Hearst. 'natural' search user interfaces. *Commun. ACM*, 54(11):60–67, Nov. 2011.

[4] M. A. Hearst and J. O. Pedersen. Visualizing information retrieval results: a demonstration of the tilebar interface. In *Conference Companion on Human Factors in Computing Systems*, pages 394–395, 1996.

[5] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77–86, 2009.

[6] P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50, 1996.

[7] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, May 2008.

[8] R. L. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 595–604, 2011.

[9] A. Spoerri. Infocrystal: A visual tool for information retrieval & management. In *Proceedings of the second international conference on Information and knowledge management*, pages 11–20, 1993.

[10] A. Younus, M. A. Qureshi, S. K. Kingrani, M. Saeed, N. Touheed, C. O'Riordan, and P. Gabriella. Investigating bias in traditional media through social media. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 643–644, 2012.

[11] T. Zesch and I. Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[7]The preliminary user evaluations have not been shared in this paper.