euroHCIR2011

4th July 2011 – Newcastle, UK

Proceedings of the 1st European Workshop on Human-Computer Interaction with Information Retrieval

A workshop at BCS-HCI2011

Executive Summary

EuroHCIR2011 was the first in a series of new workshops aimed to stimulate the European Human Computer Interaction and Information Retrieval (HCIR) community in a similar manner to series of successful workshops held in the USA. The workshop, which won industry sponsorship from LexisNexis, was highly successful, accepting 11 short papers and drawing participants from a dozen countries across Europe. In addition to the 8 insightful presentations and 3 poster presentations, Ann Blandford, from University College London's Interaction Centre, gave an inspiring keynote about their work on Exploratory Search and Serendipity.

Organised by

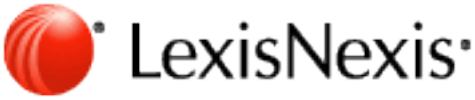
Max L. Wilson
Future Interaction Technologies Lab
Swansea University, UK,
m.l.wilson@swansea.ac.uk

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

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

James Kalbach
LexisNexis, UK
James.kalbach@lexisnexis.co.uk

Sponsored by



Session 1

Session 2

Posters

Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users

Marc Bron
ISLA, University of Amsterdam
m.m.bron@uva.nl

Jasmijn van Gorp
TViT, Utrecht University
j.vangorp@uu.nl

Frank Nack
ISLA, University of Amsterdam
nack@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

ABSTRACT

As archives are opening up and publishing their content online, the general public can now directly access archive collections. To support access, archives typically provide the public with their internal search tools that were originally intended for professional archivists. We conduct a small-scale user study where non-professionals perform exploratory search tasks with a search tool originally developed for media professionals and archivists in an audio visual archive. We evaluate the tool using objective and subjective measures and find that non-professionals find the search interface difficult to use in terms of both. Analysis of search behavior shows that non-professionals often visiting the description page of individual items in a result list are more successful on search tasks than those who visit fewer pages. A more direct presentation of entities present in the metadata fields of items in a result list can be beneficial for non-professional users on exploratory search tasks.

Categories and Subject Descriptors

H.5.2 [**User interfaces**]: Evaluation/methodology

General Terms

Measurment, Performance, Design, Experimentation

Keywords

Exploratory search, Usability evaluation

1. INTRODUCTION

Traditionally, archives have been the domain of archivists and librarians, who retrieve relevant items for a user's request through their knowledge of the content in, and organization of, the archive. Increasingly, archives are opening up and publishing their content online, making their collections directly accessible for the general public. There are two major problems that these non-professional users face. First, most users are unfamiliar or only partially familiar with the archive content and its representation in the repository. The internal representation is designed from the expert point of view, i.e., the type of information included in the metadata, which does not necessarily match the expectation of the general public. This leads to an increase in exploratory types of search [5], as users are unable to translate their information need into terms that correspond with the representation of the content in the archive. The second problem is that archives provide users with professional search tools to search through their collections. Such tools were originally developed to support professional users in searching through the metadata descriptions in a collection. Given their knowledge of the collection, professionals primarily exhibit directed search behavior [3], but it is unclear to what extent professional search tools support non-professional users in exploratory search.

The focus of most work on improving exploratory search is towards professionals [1]. In this paper we present a small-scale user study where non-professional users perform exploratory search tasks in an audio-visual archive using a search tool originally developed for media professionals and archivists. We investigate the following hypotheses: (i) a search interface designed for professional users does not provide satisfactory support for non-professional users on exploratory search tasks; and (ii) users with high performance on exploratory search tasks have different search behavior than users with lower performance.

In order to investigate the first hypothesis we evaluate the search tool performance objectively in terms of the number of correct answers found for the search tasks and subjectively through a usability questionnaire. To answer the second hypothesis, we perform an analysis of the click data logged during search.

2. EXPERIMENTAL DESIGN

The environment. The setting for our experiment was the Netherlands Institute for Sound and Vision (S&V), the Dutch national audiovisual broadcast archive. In the experiment we used the archive's collection consisting of around 1.5 M (television) programs with metadata descriptions provided by professional annotators.

We also utilized the search interface of S&V.[1] The interface is available in a simple and an advanced version. The simple version is similar to search engines known from the web. It has a single search box and submitting a query results in a ranked list of 10 programs. Clicking on one of the programs, the interface shows a page with the complete metadata description of the program. Table 1 shows the metadata fields available for a program. Instead of

[1] http://zoeken.beeldengeluid.nl

the usual snippets presented with each item in a result list, the interface shows the title, date, owner and keywords for each item on the result page. Only the keywords and title field provide information about the actual content of the program while the other fields provide information primarily used for the organization of programs in the archive collection. The description and summary fields contain the most information about the content of programs but are only available by visiting the program description page.

We used the advanced version of the interface in the experiment which next to the search box offers two other components: search boxes operating on specific fields and filters for certain categories of terms. Fielded searches operate on specific fields in the program metadata. The filters become available after a list of programs has been returned in response to a query. The filters display the top five most frequent terms in the returned documents for a metadata field. The metadata fields displayed in the filter component of the interface are highlighted in bold in Table 1. Once a checkbox next to one of the terms has been ticked, programs not containing that term in that field are removed from the result list.

Table 1: All metadata fields available for programs. We differentiate between fields that describe program content and fields that do not. Bold indicates fields used by the filter component.

content descriptors		organizational descriptors	
field	explanation	field	explanation
description	program highlights	**medium**	storage medium
person	people in program	**genre**	gameshow; news
keyword	terms provided by annotator	**rights**	parties allowed to broadcast
summary	summary of the program format	owner	owner of the broadcast rights
organization	organization in program	**date**	broadcast date
location	locations in program	origin	program origin
title	program title		

Subjects. In total, 22 first year university students from media studies participated in the experiment. The students (16 female, 6 male) were between 19 and 22 years of age. As a reward for participation the students gained free entrance to the museum of the archive.

Experiment setup. In each of the five studios available at S&V either one or two subjects performed the experiment at a time in a single studio. In case two subjects were present, each of them worked on machines facing opposite sides of the studio. We instructed subjects not to communicate during the experiment. During the experiment one instructor was always present in a studio. Before starting, the subjects learned the goals of the experiment, got a short tutorial on the search interface and performed a test query. During this phase the subjects were allowed to ask questions.

In the experiment each subject had to complete three search tasks in 45 minutes. If after 15 minutes a task was not finished, the instructor asked the subject to move on to the next task. Search tasks are related to matters that could potentially occur within courses of the student's curriculum. Each search task required the subjects to find five answers before moving on to the next task. A correct answer was a page with the complete metadata description of a program that fulfilled the information need expressed by the search task. Subjects could indicate that a page was an answer through a submit button added to the interface for the experiment.

We used the following three search tasks in the experiment: (i) For the course "media and ethnicity" you need to investigate the role of ethnicity in television-comedy. Find five programs with different comedians with a non-western background. (ii) For the course "television geography" you need to investigate the representation of places in drama series. Find five drama series where location plays an important role. (iii) For the course "media and gender" you need to give a presentation about the television career of five different female hosts of game shows broadcasted during the 1950s, 1960s or 1970s. Find five programs that you can use in your presentation.

Subjects received the search tasks in random order to avoid any bias. Also, subjects were encouraged to perform the search in any means that suited them best. During the experiment we logged all search actions, e.g., clicks, performed by each subject. After a subject had finished all three search tasks, he or she was asked to fill out a questionnaire about the experiences with the search interface.

Methodology for evaluation and analysis. We performed two types of evaluation of the search interface: a usability questionnaire and the number of correct answers submitted for the search tasks. The questionnaire consists of three sets of questions. The first set involves aspects of the experienced search behaviour with the interface. The second set contains questions about how useful users find the filter component, fielded search component, and metadata fields presented in the interface. The third set asks subjects to indicate the usefulness of a series of term clouds. The primary goal is not to evaluate the term clouds or their visualization but to find preferences for information from certain metadata fields. We generated a term cloud for a specific field as follows. First, we got the top 1000 program descriptions for the query "comedian." We counted the terms for a field for each of the documents. The cloud then represented a graphical display of the top 50 most frequent terms in the fields of those documents, where the size of a term was relative to its frequency, i.e, the higher the frequency the bigger the term. In the questionnaire subjects indicate agreement on a 5 point Likert scale ranging from one (not at all) to five (extremely). The second type of evaluation was based on the evaluation methodology applied at TREC [2]. We pooled the results of all subjects and let two assessors make judgements about the relevance of the submitted answers to a search task. An answer is only considered relevant if both assessors agree. Performance is measured in terms of the number of correct answers (#correct) submitted to the system.

For the analysis of the search behavior of subjects we looked at (i) the number of times a search query is submitted using any combination of components (#queries); (ii) the number of times a program description page is visited (#pages); and (iii) the number of times a specific component is used, i.e., the general searchbox, filters and fields. A large value for #queries indicates a look up type search behavior. It is characterized by a pattern of submitting a query, checking if the answer can be found in the result list and if it is not, to formulate a new query. The new query is not necessarily based on information gained from the retrieved results but rather inspired by the subject's personal knowledge [4]. A large value for #pages indicates a learning style search behavior. In this search strategy a subject visits the program description of each search result to get a better understanding of the organization and content of the archive. New queries are then also based on information gained from the previous text analysis [4]. We check the usage frequency of specific components to see if performance differences between subjects are due to alternative uses of interface components.

3. RESULTS

Search interface evaluation. Figure 1 shows the distribution of the amount of *correct* answers submitted for a search task, together with the distribution of the amount of answers (correct or incorrect) submitted. Out of the possible total of 330 answers, 173 are actually submitted. Subjects submit the maximum number of five

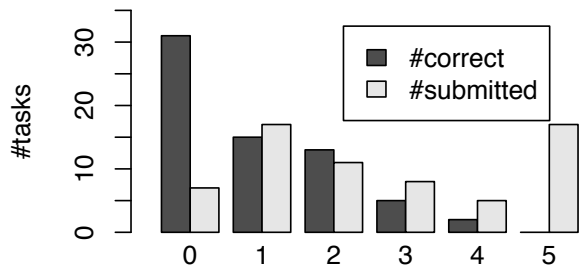


Figure 1: Distribution of amount correct/submitted answers.

answers for 18 of the tasks. This suggests that subjects have difficulties in finding answers within the given time limit. Subjects find no *correct* answers for 31 of the tasks, five subjects find no *correct* answer for any of the tasks, and none of the subjects reaches the maximum of five *correct* answers for a task. In total 64 out of 173 answers are *correct*. This low precision indicates that subjects find it difficult to judge if an answer is correct based on the metadata provided by the program description. Table 2 shows questions about the satisfaction of subjects with the interfaces. Subjects indicate their level of agreement from one (not at all) to five (extremely). For all questions the majority of subjects find the amount of support offered by the interface on the exploratory search tasks marginal. This finding supports our first hypothesis that the search interface intended for professional users does not provide satisfactory support to non-professional users on exploratory search tasks.

Search behavior analysis. Although all subjects are non-experts with respect to search with this particular interface, some perform better than others. We investigate whether there is a difference in the search behavior of subjects that have high performance on the search tasks and users that have lower performance. We divide subjects into two groups depending on the average number of correct answers found aggregated over the three tasks, i.e., 2.9 out of the possible maximum of 15. The group with higher performance (group G) consists of 11 subjects with 3 or more correct answers, whereas the group with lower performance (group B) consists of 11 subjects with 2 or less correct answers.

Table 3 shows the averages of the search behavior indicators for each of the two groups. We first look at the usage frequency of the filter, field, and search box components by subjects in group G vs. group B. There is no significant difference between the groups, indicating that there is no direct correlation between performance on the search tasks and use of specific search components. Next we look at search behavior as an explanation for the difference in performance between the groups. Our indicator for lookup searches, i.e., #queries, shows a small difference in the number of submitted queries. That subjects in both groups submit a comparable number of queries suggests that the difference in performance is not due to one group doing more lookups than the other. The indicator for learning type search, i.e., #pages, shows that there is a significant difference in the number of program description pages visited between subjects of the two groups, i.e., subjects in group G tend to visit program description pages more often than subjects of group B. We also find that the average time subjects in group G spend on a program description page is 27 seconds, while subjects from group B spend on average 39 seconds. These observations support our hypothesis that there are differences in search behavior between subjects that have high performance on exploratory search tasks and subjects with lower performance.

Usefulness of program descriptions. One explanation for this difference in performance is that through their search behavior subjects from group G learn more about the content and organization of the archive and are able to assimilate this information faster from the program descriptions than subjects from group B. As subjects process more program descriptions they learn more about the available programs and terminology in the domain. This results in a richer set of potential search terms to formulate their information need. To investigate whether subjects found information in the program descriptions useful in suggesting new search terms, we analyse the second set of questions from the questionnaire. The top half of Table 4 shows subjects' responses to questions about the usefulness of metadata fields present on the search result page. Considering responses from all subjects the genre and keyword fields are found most useful and the title and date fields as well, although to a lesser degree. The fields intended for professionals, i.e., origin, owner, rights, and medium are found not useful by the majority of subjects. Between group B and G there are no significant differences in subject's judgement of the usefulness of the fields.

Table 2: Questionnaire results about the satisfaction of subjects with the search interface. Agreement is indicated on a 5 point Likert scale ranging from one (not at all) to five (extremely).

question	mode	avg
To what degree are you satisfied with the search experience offered by the interface?	2	2.3
To what degree did the interface support you by suggesting new search terms?	2	2.4
To what degree are you satisfied with the suggestions for new search terms by the interface?	2	2.3

Table 3: Analysis of search behavior of subjects. Significance is tested using a standard two-tailed t-test. The symbol $^{\blacktriangle}$ indicates a significant increase at the $\alpha < 0.01$ significance level.

	filter	field	searchbox	#queries	#pages
B avg	21.3	29.5	44.8	35.2	21.2
G avg	15.2	44.0	42.0	34.3	$35.7^{\blacktriangle}$

Table 4: Questions about the usefulness of metadata fields on program description pages and the mode and average (avg) of the subjects responses: for all subjects, the good (G) and bad (B) performing group. We use a Wilcoxon signed rank test for the ordinal scale. The symbol $^{\triangle}$ ($^{\blacktriangle}$) indicates a significant increase at the $\alpha < 0.05$ (0.01) level.

		all	B		G	
question	field	mode	mode	avg	mode	avg
Degree to which fields on the result page were useful in suggesting new terms	date	3	2	2.2	3	3.0
	owner	1	1	1.6	1	2.0
	rights	1	1	1.3	1	1.4
	genre	4	1	2.8	4	3.9
	keyword	4	1,5	3.1	4	3.5
	origin	1	1,2	1.7	1	2.0
	title	3,4	2	2.2	4	3.0
	medium	1	1	1.5	1,2	1.6
Degree to which fields in program descriptions were useful in suggesting new terms	summary	4	1,4	2.8	5	3.8
	description	4	4	3.3	$4^{\triangle}$	4.1
	person	4	1,3,4	2.8	$4^{\blacktriangle}$	3.8
	location	1,3,4	1,3	2.0	$4^{\blacktriangle}$	3.0
	organization	1	1	1.8	1,2	2.0

The bottom part of Table 4 shows subject's responses to questions about the usefulness of metadata fields only present on the program description page and not already shown on the search result page. Based on all responses, the summary, description, person and location metadata fields are considered most useful by the majority of the subjects. These findings further support our argument that program descriptions provide useful information for subjects to complete their search tasks.

When we contrast responses of the two groups we find that group G subjects consider the description, person, and location metadata fields significantly more useful than subjects from group B. This suggests that group B subjects have more difficulties in distilling useful information from these fields (recall also the longer time spent on a page). This does not say that these users cannot understand the provided information. All that is indicated is that the chosen modality, i.e., text, might not be the right one. A graphical representation, for example as term clouds, might be better.

Fields as term clouds. In response to the observations just made, we also investigated how users would judge visual representations of search results, i.e., in the form of term clouds directly on the search result page. Here the goal is not to evaluate the visualization of the clouds or the method by which they are created. Of interest to us is whether subjects would find a direct presentation of information normally "hidden" on the program description page useful.

Recall from §2 that we generate term clouds for each field on the basis of the terms in the top 1000 documents returned for a query. From Table 5 we observe that subjects do not consider the description and summary clouds useful, while previously these fields were judged most useful among the fields in the program description. Both clouds contain general terms from the television domain, e.g., program and series, which do not provide subjects with useful search terms. Although this could be due to the use of frequencies to select terms, these fields are inherently difficult to visualize without losing the relations between the terms. The genre, keyword, location and, to some degree, person clouds are all considered useful, but they support the user in different ways. The genre field supports the subject in understanding how content in the archive is organized, i.e., it provides an overview of the genres used for categorization. The keyword cloud provides the user with alternative search terms for his original query, for example, satire or parody instead of cabaret. The location and person clouds offer an indication of which locations and persons are present in the archive and how prominent they are. For these fields visualization is easier, i.e., genre, keywords or entities by themselves are meaningful without having to represent relations between them. Subjects consider the title field only marginally useful. For this field the usefulness is dependent on the knowledge of the subject as titles are not necessarily descriptive. The subjects also consider the organization field marginally useful, probably due to the nature of our search tasks, i.e., two tasks focus on finding persons and in one locations play an important role. We assume though that in general this type of information need occurs when the general public starts exploring the archive. Together, the above findings suggest that subjects find a direct presentation of short and meaningful terms, i.e., categories, keywords, and entities, on the search results page useful.

Table 5: Questions about the usefulness of term clouds based on specific metadata fields. Agreement is indicated on a 5 point Likert scale ranging from one (not at all) to five (extremely).

cloud	mode	avg	cloud	mode	avg
title	2	2.8	description	1	2.5
person	2,3	2.9	genre	4	3.4
location	4	3.3	summary	1	2.3
organization	2	2.2	keyword	4	3.8

4. CONCLUSION

We presented results from a user study where non-professional users perform exploratory search tasks with a search tool originally developed for media professionals and archivists in an audio visual archive. We hypothesized that such search tools provide unsatisfactory support to non-professional users on exploratory search tasks. By means of a TREC style evaluation we find that subjects achieve low recall in the number of correct answers found. In a questionnaire regarding the user satisfaction with the search support offered by the tool, subjects indicate this to be marginal. Both findings support our hypothesis that a professional search tool is unsuitable for non-professional users performing exploratory search tasks.

Through an analysis of the data logged during the experiment, we find evidence to support our second hypothesis that subjects perform different search strategies. Subjects that visit more program description pages are more successful on the exploratory search tasks. We also find that subjects consider certain metadata fields on the program description pages more useful than others. Subjects indicate that visualization of certain fields as term clouds directly in the search interface would be useful in completing the search tasks. Subjects especially consider presentations of short and meaningful text units, e.g., categories, keywords, and entities, useful.

In future work we plan to perform an experiment in which we present non-professional users with two interfaces: the current search interface and one with a direct visualization of categories, keywords and entities on the search result page.

Acknowledgements. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

REFERENCES

[1] J.-w. Ahn, P. Brusilovsky, J. Grady, D. He, and R. Florian. Semantic annotation based exploratory search for information analysts. *Inf. Proc. & Management*, 46(4):383 – 402, 2010.

[2] D. K. Harman. The TREC test collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and evaluation in information retrieval*. MIT, 2005.

[3] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive. *J. Am. Soc. Inf. Sci. and Techn.*, 61:1180–1197, 2010.

[4] G. Marchionini. Exploratory search: from finding to understanding. *Comm. ACM*, 49(4):41 – 46, April 2006.

[5] R. White, B. Kules, S. Drucker, and M. Schraefel. Supporting exploratory search: Special issue. *Comm. ACM*, 49(4), 2006.

Supplying Collaborative Source-code Retrieval Tools to Software Developers

Juan M. Fernández-Luna
Departamento de Ciencias de la Computación e Inteligencia Artificial, CITIC-UGR.
Universidad de Granada, 18071 Granada, Spain
jmfluna@decsai.ugr.es

Juan F. Huete
Departamento de Ciencias de la Computación e Inteligencia Artificial, CITIC-UGR.
Universidad de Granada, 18071 Granada, Spain
jhg@decsai.ugr.es

Julio C. Rodríguez-Cano
Centro de Desarrollo Territorial Holguín. Universidad de las Ciencias Informáticas, 80100 Holguín, Cuba
jcrcano@uci.cu

ABSTRACT

Collaborative information retrieval (CIR) and search-driven software development (SDD) are both new emerging research fields; the first one was born in response to the problem of satisfying shared information needs of groups of users that collaborate explicitly, and the second to explore source-code retrieval concept as an essential activity during software development process. Taking advantages of the recent contributions in CIR and SDD, in this paper we introduce a plug-in that can be added to the *NetBeans IDE* in order to enable remote teams of developers to use collaborative source-code retrieval tools. We also include in this work experimental results to confirm that CIR&SDD techniques give out better search results than individual strategies.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and presentation (e.g., HCI)**]: Group and Organization Interfaces; H.3.3 [**Information Storage and Retrieval**]: Search Process.

General Terms

Design, Human Factors.

Keywords

Collaborative Information Seeking and Retrieval, Search-driven Software Development, Multi-user Search Interface.

1. INTRODUCTION

"Collaboration" seems to be the buzzword this year, just like "knowledge management" was last year.
– David Coleman

In the last few years, Information Retrieval (IR) Systems have become critical tools for software developers. Today we can use vertical IR systems focused in integrated development environment (IDE) extensions for source-code retrieval as such *Strathcona* [5], *CodeConjurer* [6], and *CodeGenie* [1], but these only allow an individual interaction from the team developers's perspective.

EuroHCIR '11 Newcastle, UK

One of the reasons that the existing IR systems do not adequately support collaboration is that there are not good models and methods that describe users' behavior during collaborative tasks. To address this issue, the community has adopted CIR as an emerging research field in charge to establish techniques to satisfy the shared information needs of group members, starting from the extension of the IR process with the knowledge about the queries, the context, and the explicit collaboration habits among group members. CIR community identifies four fundamental features in this multidisciplinary field that can enhance the value of collaborative search tools: user intent transition, awareness, division of labor, and sharing of knowledge [2].

In addition, SDD is a new research area motivated by the observation that software developers spend most of their time searching pertinent information that they need in order to solve their tasks at hand. We identified that SDD context was a very interesting field where collaborative IR features could be greatly exploited. For this reason we use the phrase collaborative SDD to refer to the application of different collaborative IR techniques in the SDD process [3].

It's known than some IDE incorporate tools with support for developer's collaboration practices, but without making emphasis in source-code retrieval. In this sense, the objective of this paper is to present the results of the comparison of traditional SDD and collaborative SDD. In both search scenarios, we use the *NetBeans IDE* plug-in *COSME* (COllaborative Search MEeting) with the appropriate configurations. *COSME* endows *NetBeans IDE* with traditional and collaborative source-code retrieval tools.

This paper is organized as follows: The first section presents a brief overview of related works and place our research in context. Then, we describe our software tool and method, explaining the different aspects of our experimental evaluation. Finally we discuss the results and present some conclusion remarks.

2. RELATED WORK

There is a small body of work that investigates methods to join collaborative information retrieval and search-driven software development. On the one hand, some researchers have identified different search scenarios where it is necessary to extend IR systems with collaborative capabilities. For example, in the Web context, *SearchTogether* [8] is a system which enables remote users to synchronously or asynchronously collaborate when searching the Web. It supports

collaboration with several mechanisms of group awareness, division of labor, and persistence. On the other hand, the SDD community presents different prototypes and systems. For example, *Sourcerer* [1] is an infrastructure for large-scale indexing and analysis of open source code. *Sourcerer* crawls Internet looking for *Java* code from a variety of locations, such as open source repositories, public web sites, and version control systems.

CIR systems can be applied in several domains, such as travel planning, organizing social events, working on a homework assignment or medical environments, among many others. We identified software development as another possible application field where much evidence of collaboration among programmers on a development task can be found. For example, concurrent edition of models and processes require synchronous collaboration between architects and developers who can not be physically present at a common location [7].

However, current SDD systems do not have support for explicit collaboration among developers with shared technical information needs, which frequently look for additional documentation on the API (Application Programming Interface), read posts for people having the same problem, search the company's site for help with the API, or looking for source code examples where other people successfully used the API. Fortunately, in the last few years, some researchers have realized that collaboration is an important feature, which should be analyzed in detail in order to be integrated with operational IR systems, upgrading them to CIR systems.

As an approach to these situations, we propose in this work the *COSME* plug-in [4]. It makes the contribution in current SDD providing explicit support for teams of developers, enabling developers to collaborate on both the process and results of a search. *COSME* provides collaborative search functions for exploring and managing source-code repositories and documents about technical information in the software development context.

In order to support such CIR techniques, *COSME* provides some collaborative services in the context of SDD:

- The embedded chat tool enables direct communication among different developers.
- Relevant search results can be shared with the explicit recommender mechanisms.
- Another important feature is the automatic division of labor. By implementing an effective division of labor policy the search task can be split across team developers, thereby avoiding considerable duplication of effort.
- Through awareness mechanisms all developers are always informed about the team activities to save effort. Awareness is a valuable learning mechanism that help the less experienced developers to view the syntax used by their teammates, being an inspiration to reformulate their queries.
- All search results can be annotated, either for personal use, like a summary, or in the team context, for discussion threads and ratings.

3. THE COSME PLUG-IN

To improve software developers with shared technical information needs we implemented the *COSME* front-end as a *NetBeans IDE* plug-in. The principal technologies that we used to implement it include the *CIRLab* framework [2], *NetBeans IDE* platform, *Java* as programming language, and *AMENITIES* (A MEthodology for aNalysis and desIgn of cooperaTIve systEmS) as software engineering methodology. *COSME* is designed to enable either synchronous or asynchronous, but explicit remote collaboration among teams of developers with shared technical needs. In the following section we are going to outline *COSME*.

3.1 Current Features

Figure 1 is a screenshot showing various features of our *COSME* plug-in. We refer to the circled numbers in the following text.

1. Search Control Panel: It is integrated in turn for three collapsible panels; **(a)** configuration, where the developers can select the search options and engines to accomplish the search tasks; **(b)** filters show the user's interest field according to the collection contents; and **(c)** collection type permit to specify the type of search result's items.

2. Search Results Window: The search results can be classified according to three different source-code localization: **(d)** results can be obtained as a consequence of division of labor techniques introduced by the collaborative search session (CoSS) chairman. A CoSS is a group of end-users working together to satisfy their shared information needs. One CoSS only can have one developer in the roll of chairman; **(e)** or by explicit recommendations accomplished for group members of their CoSS; **(f)** finally, search results also can be obtained by individual search.

3. Item Viewer: It shows full item content in different formats, e.g. pdf, plain text, and *Java* source-code files. All item formats are showed to the developers within the *NetBeans IDE*.

4. CoSS Portal: Developer can use the chat tool embedded in the CoSS Portal to negotiate the creation of a collaborative search session or to join at any active CoSS. For each CoSS, the chairman can to establish the integrity criteria, membership policy, and division of labor principles.

4. EXPERIMENTAL EVALUATION

In this section we are going to show how collaborative features applied to SDD improves the traditional operation without them. Then if we consider the null hypothesis (H_0) that $A_{TSDD} \geq A_{CSDD}$, our alternative hypothesis (H_1) is that the collaborative work should help to improve the retrieval performance in a SDD task: $A_{TSDD} < A_{CSDD}$, where TSDD stands for Traditional SDD and CSDD for Collaborative SDD. To evaluate our proposal we compare 10 group interactions in two different kinds of search scenarios (SS) on SDD, SS_{2k+1} and $SS_{2(k+1)}; k \in 0, \ldots, 9$. SS_{2k+1} represents a team of developers that use a conventional IR system, this means that developers do not have access to techniques of division of labor, sharing of knowledge, or awareness (traditional SDD – TSDD), while $S_{2(k+1)}$ represents a team of developers that uses a CIR system. Then, 5 teams worked in a TSDD context (those with odd subindexes) and the other 5 with CSDD (even subindexes). In both search scenarios, we used *COSME* with the appropriate configurations for both settings.

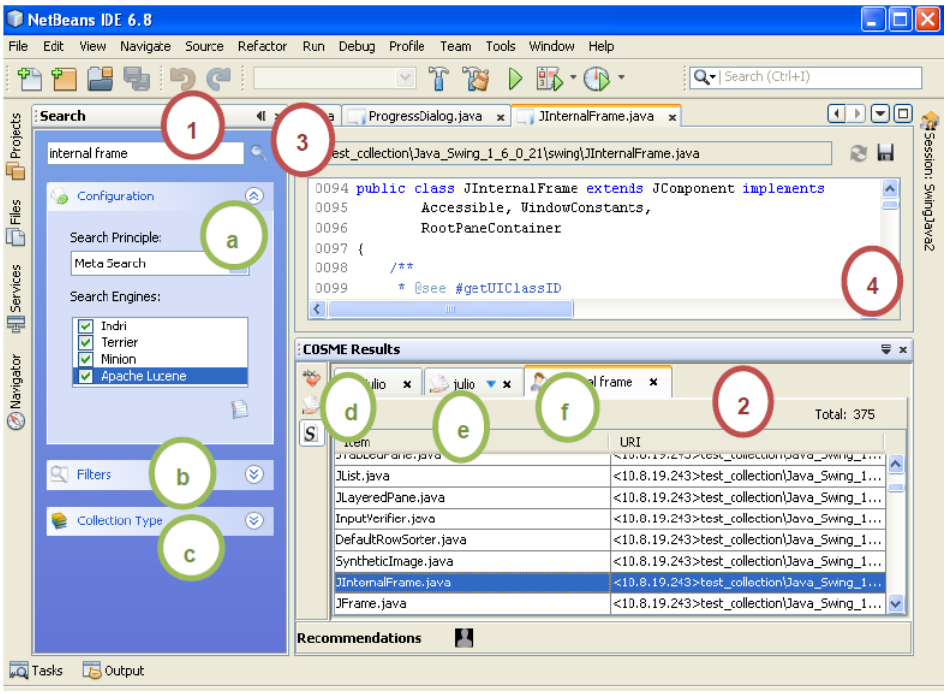


Figure 1: Screenshot of *NetBeans IDE* with *COSME* plug-in installed

The search scenario was a common task proposed to a group of developers without *Java* background: select the most relevant classes to manage GUI (Graphical User Interface) components using different *Java* API with a total of 2420 files. Specifically, *Jidesoft* (634), *OpenSwing* (434)), *SwingX* (732)) and *Swing* (620). We have focussed on these API because they are directly related to the context of the experiment although they are not complete: we have only considered their most relevant API packages for the experiment.

For evaluation purposes, we created our own test collection: a group of 10 experts proposed a set of 100 topics strongly related to the objective of the experimentation, then their corresponding queries were submitted to each of the following search engines: *Lucene*, *Minion*, *Indri* and *Terrier*. A document pool was obtained by ranking fusion and later the experts, grouped in pairs, determined the relevant documents for each topic.

In collaborative SDD, it is very important to analyze the interaction among group members, therefore, unlike the evaluation of a traditional SDD system, we can not fix the queries. Then each participating group could freely formulate their queries to the search engine. In order to compare team results, the search engine identified the most similar queries formulated by the members of the teams with respect to those formulated by experts. If the system found enough similarity and if they occur in all the groups, then these queries are considered that deals with the same topic and selected for group comparison purposes. The similarity measure between queries is calculated by Equation 1. A user query (q_u) and an expert query (q_e) are considered to be the same if they are within a given similarity threshold. A new query $q_u{}'$ is obtained applying the Porter stemmer algorithm to q_u's terms, and analogously, we would obtain $q_e{}'$.

$$sim(q_u, q_e) = \frac{\mid q_u{}' \bigcap q_e{}' \mid}{\mid q_u{}' \bigcup q_e{}' \mid} = \lambda \qquad (1)$$

In Equation 1, λ is a value between 0 and 1. For this experiment we assumed that there exists an expert's relevance judgement to q_u only if $\exists\ \lambda \geq \frac{\frac{N}{2}+1}{N}$, where $N = \mid q_u{}' \bigcup q_e{}' \mid$, selecting the relevance judgements that correspond to λ_{max} for each q_e.

In order to measure the effectiveness of the described SS_{TSDD} and SS_{CSDD} scenarios, we considered as evaluation measures the metrics proposed by Pickens et al. in [9], i.e. selected precision (P_s, the fraction of documents judged relevant by the developer that were marked relevant in the ground truth), and selected recall (R_s) as their dependent measures. To summarize effectiveness in a single number we use F_{1s} measure.

According to the documents that each team selected for each common topic, F_{1s} measure was computed. In order to accomplish the statistical analysis of the results, we use the non parametric test of Wilcoxon (all against all). The Monte Carlo method was used and adjusted with the 99% trust intervals and 10000 signs. It was considered the existences of significance (Sig.) as appear in Table 1.

We could notice significative differences between TSDD and CSDD groups, considered two by two. As F_{1s} values for CSDD groups are better than those computed from TSDD groups for those cases, then we could conclude that when teams works supported by collaborative tools, they obtain better results. From Table 1, we could realize that apart from SS_5, each SS_{TSDD} has got at least one SS_{CSDD} with significant difference values of F_{1s}. With this results we accept H_1, because $A_{TSDD} < A_{CSDD}$.

	SS_1	SS_2	SS_3	SS_4	SS_5	SS_6	SS_7	SS_8	SS_9
F_{1s}									
SS_2	$0,062$								
SS_3	$0,180$	$0,051$							
SS_4	$0,022^{\dagger}$	$0,212$	$0,038^{\dagger}$						
SS_5	$0,272$	$0,069$	$0,152$	$0,054$					
SS_6	$0,045^{\dagger}$	$0,201$	$0,080$	$0,290$	$0,056$				
SS_7	$0,215$	$0,031^{\dagger}$	$0,340$	$0,090$	$0,206$	$0,042^{\dagger}$			
SS_8	$0,053$	$0,131$	$0,061$	$0,190$	$0,072$	$0,158$	$0,070$		
SS_9	$0,243$	$0,072$	$0,201$	$0,029^{\dagger}$	$0,344$	$0,068$	$0,238$	$0,042^{\dagger}$	
SS_{10}	$0,065$	$0,098$	$0,041^{\dagger}$	$0,290$	$0,072$	$0,235$	$0,045^{\dagger}$	$0,132$	$0,058$
†: significant difference ($0,01 \leq Sig < 0,05$) ‡: highly significant difference ($Sig < 0,01$)									

Table 1: Wilcoxon Test Results.

5. CONCLUSIONS AND FUTURE WORKS

Collaboration in SDD is just being recognized as an important research area. While in some cases collaborative SDD can be handled by conventional search engines, we need to understand how the collaborative nature of source-code retrieval affects the requirements on search algorithms. Research in this direction needs to adopt the theories and methodologies of SDD and CIR, and supplement them with new approach constructs as appropriate. In this work we present *COSME* as a collaborative SDD tool that helps team developers to find better sources than searching with traditional SDD strategies, as well as an experimental approach that confirms our hypotheses.

Our ongoing work focuses on the *COSME* back-end which poses fundamental research challenges as well as provides new opportunities to let group members collaborate in new ways:

(i) Profile Analysis. We aim to analyze the user-generated data using various techniques from the study of different collaborative virtual environments and recommender systems. With the results, our goal is to provide better personalized search results, support the users while searching and recommend users to relevant trustworthy collaborators.

(ii) P2P/hybrid-network Retrieval. Due to scalability and privacy issues we favor a distributed environment by means of a P2P (peer-to-peer) retrieval feature based on hybrid architecture to store the user-generated data and collections (*CASPER* – CollAborative Search in PEer-to-peer netwoRks). The main challenges in this respect are to ensure a reliable and efficient data analysis.

6. ACKNOWLEDGMENTS

This work has been partially supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), the Spanish MICIN project TIN2008-06566-C04-01 and the Andalusian Consejería de Innovación, Ciencia y Empresa project TIC-04526. We also would like to thank Carmen Torres for support and discussions and for all of our experiment participants.

7. REFERENCES

[1] S. Bajracharya, J. Ossher, and C. Lopes. Sourcerer: An internet-scale software repository. In *SUITE '09: Proceedings of the 2009 ICSE Workshop on Search-Driven Development-Users, Infrastructure, Tools and Evaluation*, pages 1–4, Washington, DC, USA, 2009. IEEE Computer Society.

[2] J. M. Fernández-Luna, J. F. Huete, R. Pérez-Vázquez, and J. C. Rodríguez-Cano. Cirlab: A groupware framework for collaborative information retrieval research. *Information Processing and Management*, 44(1):256–273, 2009.

[3] J. M. Fernández-Luna, J. F. Huete, R. Pérez-Vázquez, and J. C. Rodríguez-Cano. Improving search–driven development with collaborative information retrieval techniques. In *HCIR '09: IIIrd Workshop on Human–Computer Interaction and Information Retrieval*, Washington DC, USA, 2009.

[4] J. M. Fernández-Luna, J. F. Huete, R. Pérez-Vázquez, and J. C. Rodríguez-Cano. Cosme: A netbeans ide plugin as a team–centric alternative for search driven software development. In *Group 2010: Ist Workshop on Collaborative Information Seeking*, Florida, USA, 2010.

[5] R. Holmes. Do developers search for source code examples using multiple facts? In *SUITE 2009: First International Workshop on Search-Driven Development Users, Infrastructure, Tools and Evaluation*, Vancouver, Canada, 2009.

[6] W. Janjic. Lowering the barrier to reuse through test-driven search. In *SUITE 2009: First International Workshop on Search-Driven Development Users, Infrastructure, Tools and Evaluation*, Vancouver, Canada, 2009.

[7] M. Jiménez, M. Piattini, and A. Vizcaíno. Challenges and improvements in distributed software development: A systematic review. 2009.

[8] M. R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In *UIST '07: Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 3–12, New York, NY, USA, 2007. ACM.

[9] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt, and M. Back. Algorithmic mediation for collaborative exploratory search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322, New York, NY, USA, 2008. ACM.

Interactive Analysis and Exploration of Experimental Evaluation Results

Emanuele Di Buccio
University of Padua, Italy
dibuccio@dei.unipd.it

Marco Dussin
University of Padua, Italy
dussinma@dei.unipd.it

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Ivano Masiero
University of Padua, Italy
masieroi@dei.unipd.it

Giuseppe Santucci
Sapienza University of Rome, Italy
santucci@dis.uniroma1.it

Giuseppe Tino
Sapienza University of Rome, Italy
tino@dis.uniroma1.it

ABSTRACT

This paper proposes a methodology based on discounted cumulated gain measures and visual analytics techniques in order to improve the analysis and understanding of IR experimental evaluation results. The proposed methodology is geared to favour a natural and effective interaction of the researchers and developers with the experimental data and it is demonstrated by developing an innovative application based on Apple iPad.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Search process]; H.3.4 [**Systems and Software**]: [Performance evaluation (efficiency and effectiveness)]

General Terms

Experimentation, Human Factors, Measurement, Performance

Keywords

Ranking, Visual Analytics, Interaction, Discounted Cumulated Gain, Experimental Evaluation, DIRECT

1. INTRODUCTION

The Information Retrieval (IR) field has a strong and long-lived tradition, that dates back to late 50s/early 60s of the last century, as far as the assessment of the performances of an IR system is concerned. In particular, in the last 20 years, large-scale evaluation campaigns, such as the Text REtrieval Conference (TREC)[1] in the United States and the Cross-Language Evaluation Forum (CLEF)[2] in Europe, have conducted cooperative evaluation efforts involving hundreds of research groups and industries, producing a huge amount of valuable data to be analysed, mined, and understood.

The aim of this work is to explore how we can improve the comprehension of and the interaction with the experimental results by IR researchers and IR system developers. We imagine the following scenarios: (i) a researcher or a developer is attending the workshop of one of the large-scale evaluation campaigns and s/he wants to explore and understand the experimental results as s/he is listening at the presentation discussing them; (ii) a team of researchers or developers is working on tuning and improving an IR system and they need tools and applications that allow them to investigate and discuss the performances of the system under examination in a handy and effective way.

These scenarios call for: (a) proper metrics that allow us to understand the behaviour of a system; (b) effective analysis and visualization techniques that allow us to get an overall idea of the main factors and critical areas which have influenced performances in order to be able to dig into details; (c) for tools and applications that allow us to interact with the experimental result in a both effective and natural way.

To this end, we propose a methodology which allows us to quickly get an idea of the distance of an IR system with respect to both its own optimal performances and the best performances possible. We rely on the (normalized) *discounted cumulated gain* (n)DCG family of measures [7] because they have shown to be especially well-suited not only to quantify system performances but also to give an idea of the overall user satisfaction with a given ranked list considering the persistence of the user in scanning the list.

The contribution of this paper is to improve on the previous work [7,11] by trying to better understand what happens when you flip documents with different relevance grades in a ranked list. This is achieved by providing a formal model that allows us to properly frame the problem and quantify the gain/loss with respect to an optimal ranking, rank by rank, according to the actual result list produced by an IR system.

The proposed model provides the basis for the development of Visual Analytics (VA) techniques that give us the possibility to get a quick and intuitive idea of what happened in a result list and what determined its perceived performances. Visual Analytics [8, 10, 14] is an emerging multi-disciplinary area that takes into account both ad-hoc and classical Data Mining (DM) algorithms and Informa-

[1] `http://trec.nist.gov/`
[2] `http://www.clef-campaign.org/`

tion Visualization (IV) techniques, combining the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where human beings and machines cooperate using their respective distinct capabilities for the most effective results. Decisions on which direction analysis should take in order to accomplish a certain task are left to final user. While IV techniques have been extensively explored [4,13], combining them with automated data analysis for specific application domains is still a challenging activity [9]. Moreover, the Visual Analytics community acknowledges the relevance of interaction for visual data analysis, and that the current research activities very often focus only on visual representation, neglecting the interaction design, as clearly stated in [14]. This refers to two different typologies of interaction: 1) interaction within a visualization and, 2), closer to the paper contribution, interaction between the visual and the analytical components.

The idea of exploring and applying VA techniques to the experimental evaluation in the IR field is quite innovative since it has never been attempted before and, due to the complexity of the evaluation measures and the amount of data produced by large-scale evaluation campaigns, there is a strong need for better and more effective representation techniques. Moreover, visualizing and assessing ranked list of items, to the best of the authors' knowledge, has not been addressed by the VA community. The few related proposals, see, e.g., [12], use rankings for presenting the user with the most relevant visualizations, or for browsing the ranked result, see, e.g., [5], but do not deal with the problem of observing the ranked item position, comparing it with an ideal solution, to assess and improve the ranking quality. A first attempt in such a direction is in [1], where the authors explored the basic issues associated with the problem, providing basic metrics and introducing a VA web based system that allows for exploring the quality of a ranking with respect to an optimal solution.

On top of the proposed model, we have built a running prototype where the experimental results and data are stored in a dedicated system accessible via standard Web services. This allows for the design and development of various client applications and tools for exploiting the managed data. In particular, in this paper, we have started to explore the possibility of adopting the Apple iPad[3] as an appropriate device to allow users to easily and naturally interact with the experimental data and we have developed an initial prototype that allows us for interactively inspecting the actual experimental data in order to get insights about the behaviour of a IR system.

Overall, the proposed model, the proposed visualization techniques, and the implemented prototype meet all the (a-c) requirements for the two scenarios introduced above.

The paper is organized as follows. Section 2 introduces the model underlying the system together with its visualization techniques; Section 3 describes the interaction strategies of the system, Section 4 describes the overall architecture of the system, and Section 5 concludes the paper, pointing out ongoing research activities.

2. THE PROTOTYPE

According to [7] we model the retrieval results as a ranked vector of n documents V, i.e., $V[1]$ contains the identifier of the document predicted by the system to be most relevant, $V[n]$ the least relevant one. The ground truth GT function assigns to each document $V[i]$ a value in the relevance interval $\{0..k\}$, where k represents the highest relevance score, e.g. $k = 3$ in [7]. The basic assumption is that the greater the position of a document the less likely it is that the user will examine it, because of the required time and effort and the information coming from the documents already examined. As a consequence, the greater the rank of a relevant document the less useful it is for the user. This is modeled through a discounting function DF that progressively reduces the relevance of a document, $GT(V[i])$ as i increases:

$$DF(V[i]) = \begin{cases} GT(V[i]), \ if\, i \leq x \\ GT(V[i])/\log_x(i), \ if \ i > x \end{cases} \quad (1)$$

The quality of a result can be assessed using the discounted cumulative gain function $DCG(V,i) = \sum_{j=1}^{i} DF(V[j])$ that estimates the information gained by a user that examines the first i documents of V.

The DCG function allows for comparing the performances of different search engines, e.g., plotting the $DCG(i)$ values of each engine and comparing the curve behavior.

However, if the user's task is to improve the ranking performance of a single search engine, looking at the misplaced documents (i.e., ranked too high or too low with respect to the other documents) the DCG function does not help: the same value $DCG(i)$ could be generated by different permutations of V and it does not point out the loss in cumulative gain caused by misplaced elements. To this aim, we introduce the following definitions and novel metrics.

We denote with $OptPerm(V)$ the set of optimal permutations of V such as that $\forall OV \in OptPerm(V)$ it holds that $GT(OV[i]) \geq GT(OV[j]) \forall i,j <= n \bigwedge i < j$, that is, OV maximizes the values of $DCG(OV,i) \forall i$. In other words, $OptPerm(V)$ represents the set of the optimal rankings for a given search result. It is worth noting that each vector in $OptPerm(V)$ is composed by $k+1$ intervals of documents sharing the same GT values. As an example, assuming a result vector composed by 12 elements and $k = 3$, a possible sequence of GT values of an optimal vector OV is <3,3,3,3,2,2,2,2,1,1,0,0>; according to this we define the $max\_index(V,r)$ and $min\_index(V,r)$ functions, with $0 \leq r \leq k$, that return the greatest and the lowest indexes of elements in a vector belonging to $OptPerm(V)$ that share the same GT value r. As an example, considering the above 12 GT values, $min\_index(V,2) = 5$ and $max\_index(V,2) = 8$.

Using the above definitions we can define the relative position $R\_Pos(V[i])$ function for each document in V as follows:

$$\begin{cases} 0, \ if\, min\_index(V, GT(V[i])) \leq i \leq max\_index(V, GT(V[i])) \\ min\_index(V, GT(V[i])) - i, \ if\, i < min\_index(V, GT(V[i])) \\ max\_index(V, GT(V[i])) - i, \ if\, i > max\_index(V, GT(V[i])) \end{cases}$$

$R\_Pos(V[i])$ allows for pointing out misplaced elements and understanding how much they are misplaced: 0 values denote documents that are within the optimal interval, negative and positive values denote elements that are respectively below and above the optimal interval. The absolute value of $R\_Pos(V[i])$ gives the minimum distance of a misplaced element from its optimal interval.

According to the actual relevance and rank position, the same value of $R\_Pos(V[i])$ can produce different variations of the DCG function. We measure the contributions of mis-

[3] http://www.apple.com/ipad/

Figure 1: The iPad prototype interface.

placed elements with the function $\Delta\_Gain(V, i)$ that compares $\forall i$ the actual values of $DF(V[i])$ with the corresponding values in OV, $DF(OV[i])$: $\Delta\_Gain(V, i) = DF(V[i]) - DF(OV[i])$.

3. INTERACTION

A multi-touch prototype interface based on the model presented in section 2 has been designed for the iPad device. It has been developed and tested on the iOS 4.2[4] with the integration of the Core Plot[5] plotting framework for the graphical visualization of data. The interface allows the end user for comparing the curve of the ranked results, for a given experiment/topic, with the optimal one and with the ideal one. This facilitates the activities of failure analysis, easily locating misplaced elements, blue or red items, that pop up from the visualization together with the extent of their displacement and the impact they have on *DCG*.

Figure 1 shows a screenshot of the current interface: the main list on the left represents the top $n = 200$ ranked result for a given experiment/topic and it can be easily scrolled by the user. Each row corresponds to a document ID, a short snippet of the content is included in the subtitle of each cell and more information on a specific result (i.e. relevance score, *DCG*, *R_Pos*, Δ*_Gain*) can be viewed by touching the row. On the right side there are two coloured vectors which show the *R_Pos* and Δ*_Gain* functions. The *R_Pos* vector presents the results using different color shadings: light green, light red and light blue respectively for documents that are within, below and above the optimal interval. It allows for locating misplaced documents and, thanks to the shading, understanding how they are far from the optimal position. Similarly, the Δ*_Gain* vector codes the function using colors: light blue refers to positive values, light red codes negative values, and green 0 values. Moreover, if the user touches a specific area of the *R_Pos* vector (that is simulated by the gray round in Figure 1), the main results list automatically scrolls back, providing the end user with a detailed view on the corresponding documents. The rightmost part of the screen shows the *DCG* graphs of the ideal, the optimal and the experiment vector, i.e. the ranking curves. The navigation bar displays a back button on the right which let the user visualize the results for a different topic.

4. ARCHITECTURE

The design of the architecture of the system benefits from what has been learned in ten years of work for the CLEF and in the design and implementation of Distributed Information Retrieval Evaluation Campaign Tool (DIRECT), the system developed in CLEF since 2005 to manage all the aspects of an evaluation campaign [2, 3].

The approach to the architecture is the implementation of a modular design, as sketched in Figure 2, with the aim to clearly separate the logic entailed by the application into three levels of abstraction – data, application, and interface logic – able to reciprocally communicate, easily extensible and implementable using modular and reusable components. The *Data Logic* layer, depicted at the bottom of Figure 2, deals with the persistence of the information coming from the other layers. From the implementation point of view, data stored into databases and indexes are mapped to resources and communicate with the upper levels through the mechanism granted by the Data Access Object (DAO) pattern[6] — see point *(1)* in Figure 2. The *Application Logic*

[4] `http://developer.apple.com/`
[5] `http://code.google.com/p/core-plot/`
[6] `http://java.sun.com/blueprints/corej2eepatterns/`

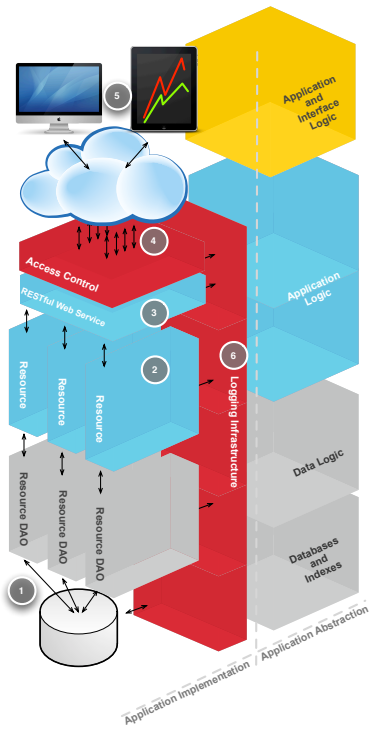


Figure 2: The Architecture of the Application.

layer is in charge of the high-level tasks made by the system, such as the enrichment of raw data, the calculation of metrics and the carrying out of statistical analyses on experiments. These resources *(2)* are therefore accessible via HTTP through a RESTful Web service [6], sketched at point *(3)*. After the validation of credentials and permissions made by the access control mechanism *(4)*, it is possible for remote devices such as web browsers or custom clients *(5)* to create, modify, or delete resources attaching their representation in XML[7] or JSON[8] format to the body of an HTTP request, and to read them as response of specific queries. A logging infrastructure *(6)* grants the tracking of all the activities made at each layer and can be used to obtain information about the provenance of all the managed resources.

5. CONCLUSIONS

We have presented a model and a prototype which allow users to easily interact with the experimental results and to work together in a cooperative way while actually accessing the data. This first step uncovers new and interesting possibilities for the experimental evaluation and for the way in which researchers and developers usually carry out such activities. For example, the proposed techniques may alleviate the burden of certain tasks, such as failure analysis, which are often overlooked due to their demanding nature, thus making easier and more common to perform them and, as a consequence, improving the overall comprehension of system behaviour. This will be explored in the future work.

Patterns/DataAccessObject.html
[7]http://www.w3.org/XML/
[8]http://www.ietf.org/rfc/rfc4627.txt

Acknowledgements

The work reported in this paper has been partially supported by the PROMISE network of excellence (contract n. 258191), as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

6. REFERENCES

[1] N. Ferro, A. Sabetta, G. Santucci, G. Tino, and F. Veltri. Visual comparison of ranked result cumulated gains. In *Proc. of EuroVA 2011*. Eurographics, 2011.

[2] M. Agosti, G. Di Nunzio, M. Dussin, and N. Ferro. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In *Proc. of EVIA 2010*, pages 16–24. Tokyo, Japan, 2010.

[3] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods*. Chandos Publishing, Oxford, UK, 2009.

[4] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *Proc. of InfoVis '97*, pages 92–99, Washington, DC, USA, 1997. IEEE Computer Society.

[5] M. Derthick, M. G. Christel, A. G. Hauptmann, and H. D. Wactlar. Constant density displays using diversity sampling. In *Proc. of the IEEE Information Visualization*, pages 137–144, 2003.

[6] R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. *ACM TOIT*, 2:115–150, 2002.

[7] K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM TOIS*, 20(4):422–446, October 2002.

[8] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Information visualization. chapter Visual Analytics: Definition, Process, and Challenges, pages 154–175. Springer-Verlag, Berlin, Heidelberg, 2008.

[9] D. Keim, J. Kohlhammer, G. Santucci, F. Mansmann, F. Wanner, and M. Schäfer. Visual Analytics Challenges. In *Proc. of eChallenges 2009*, 2009.

[10] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proc. of IV'06*, pages 9–16, 2006.

[11] H. Keskustalo, K. Järvelin, A. Pirkola, and J. Kekäläinen. Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In *Proc. of SIGIR '08*, pages 675–681. ACM Press, NY, USA, 2008.

[12] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. In *Proc. of the IEEE Information Visualization*, pages 65–72, 2004.

[13] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. of the 1996 IEEE Symposium on Visual Languages*, pages 336 –343, 1996.

[14] J. J. Thomas and K. A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26:10–13, 2006.

A Taxonomy of Enterprise Search

Tony Russell-Rose
UXLabs Ltd.
London
UK
+44 7779 936191
tgr@uxlabs.co.uk

Joe Lamantia
Endeca
101 Main St.
Cambridge, USA
+1 617 674 6000
jlamantia@endeca.com

Mark Burrell
Endeca
101 Main St.
Cambridge, USA
+1 617 674 6000
mburrell@endeca.com

ABSTRACT

Classic IR (information retrieval) is predicated on the notion of users searching for information in order to satisfy a particular "information need". However, it is now accepted that much of what we recognize as search behaviour is often not informational per se. For example, Broder (2002) has shown that the need underlying a given web search could in fact be navigational (e.g. to find a particular site or known item) or transactional (e.g. to find a sites through which the user can transact, e.g. through online shopping, social media, etc.). Similarly, Rose & Levinson (2004) have identified consumption of online resources as a further category of search behaviour and query intent.

In this paper, we extend this work to the enterprise context, examining the needs and behaviours of individuals across a range of search and discovery scenarios within various types of enterprise. We present an initial taxonomy of "discovery modes", and discuss some initial implications for the design of more effective search and discovery platforms and tools.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Search process;
H.3.5 [Online Information Services]: Web-based services

General Terms

Human Factors.

Keywords

Enterprise search, information seeking, user behaviour, knowledge workers, search modes, information discovery, user experience design.

1. INTRODUCTION

To design better search and discovery experiences we must understand the complexities of the human-information seeking process. Numerous theoretical frameworks have been proposed to characterize this complex process, notably the standard model (Sutcliffe & Ennis 1998), the cognitive model (Norman 1998) and the dynamic model (Bates, 1989). In addition, others have investigated search as a strategic process, examining the various problem solving strategies and tactics that information seekers employ over extended periods of time (e.g. Kuhlthau, 1991).

In this paper, we examine the needs and behaviours of varied individuals across a range of search and discovery scenarios within various types of enterprise. These are based on an analysis of the scenarios derived from numerous engagements involving the development of search and business intelligence solutions utilizing the Endeca Latitude software platform. In so doing, we extend the classic IR concept of information-seeking to a broader notion of discovery-oriented problem solving, accommodating the much wider range of behaviours required to fulfil the typical goals and objectives of enterprise knowledge workers.

Our approach to enterprise discovery is an activity-centred model inspired by Don Norman's Activity Centred Design, which "organizes according to usage" whereas "...traditional human centred design organizes according to topic, in isolation, outside the context of real, everyday use." (Norman 2006). This approach is an extension of previous activity-centred modelling efforts which focused on a "captur[ing] a systematic and holistic view of what users need to accomplish when undertaking information retrieval tasks more complex than searching" (Lamantia 2006), employing Grounded Theory to provide methodological structure (Glaser 1967).

In this context, we present an alternative model focused on *information discovery* rather than information seeking per se, which has at its core an initial taxonomy of the "modes of discovery" that knowledge workers employ to satisfy their information search and discovery goals. We then discuss some initial implications of this model for the design of more effective search and discovery platforms and tools.

2. INFORMATION RETRIEVAL MODELS

The classic model of IR assumes an interaction cycle consisting of four main activities: the identification an information need, the specification of an appropriate query, the examination of retrieval results, and reformulation (where necessary) of the original query. This cycle is then repeated until a suitable result set is found (Salton 1989).

In both the above models, the user's information need is assumed to be static. However, it is now acknowledged that information seekers' needs often change as they interact with a search system. In recognition of this, alternative models of information seeking have been proposed. For example, Bates (1989) proposed the dynamic "berry-picking" model of information seeking, in which the information need (and consequently the query) changes throughout the search process This model also recognises that information needs are not satisfied by a single, final result set, but

by the aggregation of results, insights and interactions along the way.

Bates' work is particularly interesting as it explores the connections between the dynamic model and the search strategies and tactics that professional information-seekers employ. In particular, Bates identifies a set of 29 individual tactics, organised into four broad categories (Bates, 1979). Likewise, O'Day & Jeffries (1993) examined the use of information search results by clients of professional information intermediaries and identified three distinct "search modes" or major categories of search behaviour: (1) Monitoring a known topic or set of variables over time; (2) Following a specific plan for information gathering; (3) Exploring a topic in an undirected fashion.

O'Day and Jeffries also observed that a given search would often evolve over time into a series of interconnected searches, delimited by certain triggers and stop conditions that indicate the transitions between modes or individual searches executed as part of an overall enquiry or scenario. Moreover, O'Day & Jeffries also attempted to characterise the analysis techniques employed by the clients in interpreting the search results, identifying the following six primary categories: (1) Looking for trends or correlations; (2) Making comparisons; (3) Experimenting with different aggregations/scaling; (4) Identifying critical subsets; (5) Making assessments; (6) Interpreting data to find meaning.

More recent investigations into the relationship between information needs and search activities include that of Marchionini (2005), who identifies three major categories of search activity, namely "Lookup", "Learn" and "Investigate".

3. A TAXONOMY OF ENTERPRISE SEACH AND DISCOVERY

The primary source of data in this study is a set of user scenarios captured during numerous engagements involving the development of search and business intelligence solutions utilizing the Endeca Latitude software platform. These scenarios take the form of a simple narrative that illustrates the user's end goal and the primary task or action they take to complete it, followed by a brief description of their job function or role, for example:

- "I need to understand a portfolio's exposures to assess portfolio-level investment mix" (Portfolio Manager)
- "I need to understand the quality performance of a part and module set in manufacturing and the field so that I can determine if I should replace that part" (Engineering)

These scenarios were manually analyzed to identify themes or modes that appeared consistently throughout the set. For example, in each of the scenarios above there is an articulation of the need to develop an understanding or comprehension of some aspect of the data, implying that "comprehending" may constitute one such discovery mode. Inevitably, this analysis process was somewhat iterative and subjective, echoing the observations made by Bates (1979) in the identification of her search tactics: "*While our goal over the long term may be a parsimonious few, highly effective tactics, our goal in the short term should be to uncover as many as we can, as being of potential assistance. Then we can test the tactics and select the good ones. If we go for closure too soon, i.e., seek that parsimonious few prematurely, then we may miss some valuable tactics*."

There are however some guiding principles that we can apply to facilitate convergence on a stable set. For example, an ideal set of modes would exhibit properties such as: Consistency (they represent approximately the same level of abstraction); Orthogonality (they operate independently to each other); and Comprehensiveness (they address the full range of discovery scenarios).

The initial set of discovery modes to emerge from this analysis consists of a set of nine, arranged into three top-level categories consistent with those of Marchionini (2005). The nine modes are as follows, each shown with a brief definition:

1. Lookup

1a. **Locating**: To find a specific (possibly known) item; 1b. **Verifying**: To confirm or substantiate that an item or set of items meets some specific criterion; 1c. **Monitoring**: To maintain awareness of the status of an item or data set for purposes of management or control.

2. Learn

2a. **Comparing**: To examine two or more items to identify similarities & differences; 2b. **Comprehending**: To generate insight by understanding the nature or meaning of an item or data set; 2c. **Exploring**: To proactively investigate or examine an item or data set for the purpose of serendipitous knowledge discovery.

3. Investigate

3a. **Analyzing**: To critically examine the detail of an item or data set to identify patterns & relationships; 3b. **Evaluating**: To use judgment to determine the significance or value of an item or data set with respect to a specific benchmark or model; **Synthesizing**: To generate or communicate insight by integrating diverse inputs to create a novel artefact or composite view.

Evidently, the output of this process has been optimized for the current data set and in that respect represents an initial interpretation that will need to evolve further. For example, "monitoring" may appear to be a lookup activity when considered in the context of a simple alert message, but when viewed as a strategic activity performed by an executive in the context of an organisational dashboard, a much greater degree of interaction and complexity is implied. Conversely, "exploring" is a concept whose level of abstraction may prove somewhat higher than the others, thus breaking the consistency principle suggested above.

However, the true value of the modes will be realised not by their conceptual purity or elegance but by their utility as a design resource. In this respect, they should be judged by the extent to which they facilitate the design process in capturing important characteristics common to enterprise search and discovery experiences, whilst flexibly accommodating arbitrary variations in domain, information resources, etc.

4. MODE SEQUENCES AND PATTERNS

A further interesting observation arising from the above analysis is that the mapping between scenarios and modes is not one-to–one. Instead, some scenarios are seen to involve a number of modes, sometimes with a primary or dominant mode, and often with an implied linear sequence. Moreover, certain sequences of modes tend to re-occur more frequently than others, forming specific "mode chains" or patterns, analogous to higher-level syntactic units. These patterns provide a framework for

understanding the transitions between modes (echoing the triggers identified by O'Day & Jeffries), and allude to the existence of natural seams that can be used be used to provide further insight into information enterprise search and discovery behaviour.

These mode chains echo the above-mentioned efforts to create goal-based information retrieval models, which yielded modes and a set of broadly applicable "information retrieval patterns that describe the ways users combine and switch modes to meet goals: Each pattern is assembled from combinations of the same four [elemental] modes" (Lamantia 2006).

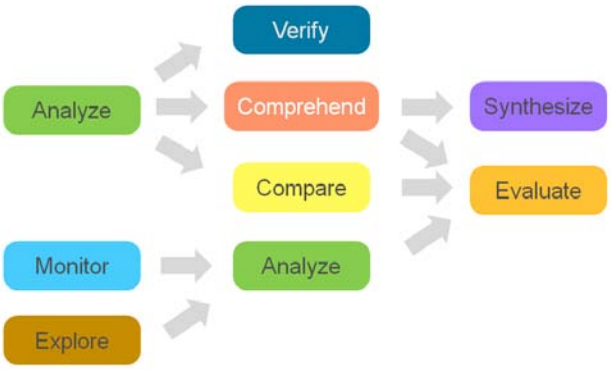


Figure 1. Discovery mode network

The five most frequent mode patterns are listed below. These have been assigned descriptive (if somewhat informal) labels to aid their characterisation, along with the sequence of modes they represent and an associated example scenario:

1. **Comparison-driven optimization**: (Analyze-Compare-Evaluate) e.g. "*Replace a problematic part with an equivalent or better part without compromising quality and cost*"
2. **Exploration-driven optimization**: (Explore-Analyze-Evaluate) e.g. "*Identify opportunities to optimize use of tooling capacity for my commodity/parts*"
3. **Strategic Insight** (Analyze-Comprehend-Evaluate) e.g. "*Understand a lead's underlying positions so that I can assess the quality of the investment opportunity*"
4. **Strategic Oversight** (Monitor-Analyze-Evaluate) e.g. "*Monitor & assess commodity status against strategy/plan/target*"
5. **Comparison-driven Synthesis** (Analyze-Compare-Synthesize) e.g. "*Analyze and understand consumer-customer-market trends to inform brand strategy & communications plan*"

Further insight may be derived by examining how the mode patterns combine across all the scenarios to the form of a "mode network", as shown in Figure 1. Evidently, some modes act as "terminal" nodes, i.e. entry points or exit points to a discovery scenario. For example, Monitor and Explore feature only as entry points at the initiation of a scenario, whilst Synthesize and Evaluate feature only as exit points to a scenario.

5. DESIGN PRINCIPLES FOR SEARCH AND DISCOVERY SOLUTIONS

The modes establish a 'taskonomy' or collection of defined discovery activities which are structurally consistent, domain and scale independent, orthogonal, semantically distinct, conceptually connected, and flexibly sequenceable. Such a profile -- analogous to notes in the musical scale, or the words and phrases we assemble into sentences -- should allow the modes to serve as a language for the design of variable scale activity-centered discovery solutions through common constructive mechanisms such as concatenation, combination and nesting. And if the modes do act as an elementary grammar for discovery, then sustained use as a functional and interaction design language should result in the creation of larger and more complex units of meaning which offer cumulative value.

Professional experience with employing the modes as both an analytical framework for understanding discovery needs and as a design grammar for the definition of discovery solutions suggests that both implications are valid. Further, our observations of using the modes suggest the existence of recognizable patterns in the design of discovery solutions. We will briefly discuss some of the patterns observed, doing so at three common levels of solution scale: on the level of a single functional or interface element, for whole screens or interfaces composed of multiple functional elements, and for applications comprising multiple screens.

5.1 Single element patterns

5.1.1 Comparison Views

One of the most common design patterns is to support the need for the Compare mode by creating A/B type comparison views that present two display panes - each containing data display charts or tables; or single items or groups of items - side by side to emphasize similarities and differences.

5.1.2 Contextual Views

Another common design pattern supports the Analysis mode by allowing a fore-grounded view of a single chart, table, item, or list, accompanied by its contextual 'halo' - the full body of information available about the element such as status, origin, format, relationships to other elements; annotations; etc.

5.2 Whole screen patterns

5.2.1 Dashboard

One of the most common screen-level design patterns is to support the Monitoring and Synthesis modes by presenting a collection of metrics which in aggregate provide the status of independent processes, groups, or progress versus goals in a 'dashboard' style screen.

5.2.2 Visual Discovery Screen: 4-Dimensions

A second common screen-level design pattern for discovery experiences is the visual discovery screen, which supports modes such Exploration, Evaluation, and Verification by layering views that present visualizations of several dimensions of a single axis of focus such as a core process, organizational unit, or KPI. When switching between layered views, the axis in focus remains the same, but the data and presentation in the dimensions adjusts to match the preferred discovery mode.

5.3 Application-level patterns

5.3.1 Differentiated Application

The 'Differentiated Application' pattern assembles a collection of individual screens whose distinct compositions and designs support individual discovery modes of Analysis, Comparison, Evaluation and Monitoring in aggregate to address the 'Strategic Oversight' mode sequence. Application-level patterns often

address a spectrum of discovery needs for a group of users with differing organizational responsibilities, such as management vs. detailed analysis.

6. DISCUSSION

The above analysis is predicated on the notion that the user scenarios provide a unique insight into the information needs of enterprise knowledge workers. However, a number of caveats apply to both the data and the approach.

Firstly, the scenarios were originally generated to support the development of a specific implementation rather than for the analysis above. Therefore, the principles governing their creation may not faithfully reflect the true distribution or priority of information needs among the various end user populations. Secondly, the particular sample we selected for this study was based on a number of pragmatic factors (including availability), which may not faithfully represent the true distribution or priority among enterprise organizations. Thirdly, the data will inevitably contain some degree of subjectivity, particularly in cases where scenarios were generated by proxy rather than with direct end-user contact. Fourthly, the data will inevitably contain some degree of inconsistency in cases where scenarios were documented by different individuals.

We should also acknowledge a number of caveats concerning the process itself. In inductive work with foundations in qualitatively centered frameworks such as Grounded Theory, it is expected that a number of iterations of a "propose-classify-refine" cycle will be required for the process to converge on a stable output (e.g. Rose & Levinson, 2004). In addition, those iterations should involve a variety of critical viewpoints, with the output tested and refined using a separate, independent sample on each iteration. Likewise, the process by which scenarios are classified would benefit from further rigour: this is a critical part of the process and of course relies on human judgement and inference, but that judgement needs to go beyond simple word matching and be consistently applied to each scenario so that subtle distinctions in meaning and intent can be accurately identified and recorded.

That said, some interesting comparisons can already be made with the existing frameworks. For example, the first and third of the search modes suggested by O'Day and Jeffries have also been identified as distinct discovery modes in our own study, and the second (arguably) maps on to one or more of the mode chains identified above. Likewise, the search results analysis techniques that O'Day & Jeffries identified also present some interesting parallels.

7. CONCLUSIONS AND FUTURE DIRECTIONS

To design better search and discovery experiences we must understand the complexities of the human-information seeking process. In this paper, we have examined the needs and behaviours of varied individuals across a range of search and discovery scenarios within various types of enterprise. In so doing, we have extended the classic IR concept of information-seeking to a broader notion of discovery-oriented problem solving, accommodating the much wider range of behaviours required to fulfil the typical goals and objectives of enterprise knowledge workers.

In addition, we have proposed an alternative model focused on *information discovery* rather than information seeking which has at its core a taxonomy of "modes of discovery" that knowledge workers employ to satisfy their information search and discovery goals. We have also examined some of the initial implications of this model for the design of more effective search and discovery platforms and tools.

Suggestions for future work include further iterations on the "propose-classify-refine" cycle using independent data. This data should ideally be acquired based on a principled sampling strategy that attempts where possible to address any biases introduced in the creation of the original scenarios. In addition, this process should be complemented by empirical research and observation of knowledge workers in context to validate and refine the discovery modes and triggers that give rise to the observed patterns of usage.

8. REFERENCES

[1] Bates, Marcia J. 1979. "Information Search Tactics." *Journal of the American Society for Information Science* 30: 205-214

[2] Bates, Marcia J. 1989. "The Design of Browsing and Berrypicking Techniques for the Online Search Interface." *Online Review* 13: 407-424.

[3] Broder, A. 2002. A taxonomy of web search, *ACM SIGIR Forum*, v.36 n.2, Fall 2002

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

[5] Lamantia, J. 2006. "10 Information Retrieval Patterns" JoeLamantia.com, http://www.joelamantia.com/information-architecture/10-information-retrieval-patterns

[6] Glaser, B. & Strauss, A. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter.

[7] Marchionini, G. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49(4): 41-46

[8] Norman, Donald A. 1988. *The psychology of everyday things*. New York, NY, US: Basic Books.

[9] Donald A. Norman. 2006. Logic versus usage: the case for activity centered design. *Interactions* 13, 6

[10] O'Day, V. and Jeffries, R. 1993. Orienteering in an information landscape: how information seekers get from here to there. *INTERCHI 1993*: 438-445

[11] Rose, D. and Levinson, D. 2004. Understanding user goals in web search, *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA

[12] Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

[13] A.G. Sutcliffe and M. Ennis. Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10:321–351, 1998.

Back to MARS: The unexplored possibilities in query result visualization

Alfredo Ferreira
INESC-ID/IST/TU Lisbon
Lisbon, Portugal
alfredo.ferreira@ist.utl.pt

Pedro B. Pascoal
INESC-ID/IST/TU Lisbon
Lisbon, Portugal
pmbp@ist.utl.pt

Manuel J. Fonseca
INESC-ID/IST/TU Lisbon
Lisboa, Portugal
mjf@inesc-id.pt

ABSTRACT

A decade ago, Nakazato proposed 3D MARS, an immersive virtual reality environment for content-based image retrieval. Even so, the idea of taking advantage of post-WIMP interfaces for multimedia retrieval was no further explored for content-based retrieval. Considering the latest low-cost, off-the-shelf hardware for visualization and interaction, we believe that is time to explore immersive virtual environments for multimedia retrieval. In this paper we highlight the advantages of such approach, identifying possibilities and challenges. Focusing on a specific field, we introduce a preliminary immersive virtual reality prototype for 3D object retrieval. However, the concepts behind this prototype can be easily extended to the other media.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Interaction Styles, Input Devices and Strategies*

Keywords

Multimedia Information Retrieval, 3D Object Retrieval, Immersive Virtual Environment

1. INTRODUCTION

Despite advances on multimedia information retrieval (MIR), this field still on its infancy. Especially when compared to its textual counterpart. Actual textual search engines are maturely developed and its widespread use makes them familiar to most users. The current scenario in MIR is quite different. Indeed, existing content-based MIR solutions are far from being largely used by the common user.

A few exceptional systems were able to strive with relative success, such as **Retrievr**[1], a search tool for **Flickr**[2] based on visual queries. However, most existing solutions still face major drawbacks and challenges to be tackled. Among others, extensively identified in Datta's survey [5], we highlight two. First, queries rely mostly on meta-information, often keyword-based. This means that, in a closer analysis, searches can be reduced to text information retrieval of multimedia objects. Second, the result visualization follows the traditional paradigm, where the results are presented as a list of items on a screen. These items are usually thumbnails, but can be just filenames or metadata. Such methodology greatly hinders the interpretation of query results on collections of videos or 3D objects.

[1]http://labs.systemone.at/retrievr/
[2]http://www.flickr.com/

Notably, a decade ago, a new visualization system for content-based image retrieval(CBIR) was proposed by Nakazato and Huang from the University of Illinois. The 3DMARS [11] was an immersive virtual reality (VR) environment to perform image retrieval. It worked on the NCSA CAVE [4] which provided fully immersive experience and later on desktop VR systems. However, despite this ground-breaking work and recent developments in the interaction domain, little advantages have been taken by the multimedia information retrieval community from immersive virtual environments.

In this paper we bring up the work of Nakazato and Huang as a starting point to the exploration of new possibilities for result visualization in multimedia information retrieval. With the spreading of stereoscopic viewing and last generation interaction devices outside lab environment and into our everyday lives, we believe that in a short time users will expect richer results from multimedia search engines than just a list of thumbnails. Following this rationale, and despite it could be applied to any type of media, we will focus our approach on 3D object retrieval (3DOR).

2. TRADITIONAL 3DOR APPROACHES

The first and most noticeable 3D search engine, at least within researchers working on this area, is the Princeton 3D Model Search Engine[8]. This remarkable work provide content-based retrieval of 3D models from a collection of more than 36000 objects. Four query specification options are available: text based; by example; by 2D sketch; and by 3D sketch. The results of this queries are presented as an array of model thumbnails.

Additionally to queries by example and sketch-based queries, the FOX-MIIRE search engine[1] introduced the query by

photo. This was the first tool capable of retrieve a 3D model from a photograph of a similar object. However, and similarly to Princeton engine, the results are displayed as a thumbnail list.

Outside the research field, **Google 3D Warehouse**[3] offers a text-based search engine for the common user. This online repository contains a very large number of different models, from monuments to cars and furniture, humans and spaceships. However, searching for models in this collection is limited by textual queries or, when models represent real objects, by its georeference. On the other hand, the results are displayed by model images in a list, with the opportunity to manipulate a 3D view of a selected model.

Generally, the query specification and visualization of results in commercial tools for 3D object retrieval, usually associated with 3D model online selling sites, did not differ much from those presented above. The query is specified through keywords or by example and results are presented as a list of model thumbnails.

These traditional approaches to query specification and result visualization do not take advantage of latest advances of neither computer graphics or interaction paradigms. Current hardware and software are capable of handling millions of triangles per frame and generating complex effects in real-time. Additionally, the growingly common use of new human-computer interaction (HCI) paradigms and devices brought new possibilities for multi-modal systems.

3. NEW PARADIGMS IN HCI

The recent dissemination among common users of new HCI paradigms and devices (e.g. Nintendo Wiimote[4] or Microsoft Kinect[5]) brought new possibilities for multi-modal systems. For decades, the "windows, icons, menus, pointing device" (WIMP) interaction style prevailed outside the research field, while post-WIMP interfaces were being devised and explored [16], but without major impact in everyday use of computer systems.

Particularly, the use of gestures to interact with system has been part of the interface scene since the very early days. A pioneering multimodal application was "Put-that-there" [2], by Bolt. In "Put-that-there", the user commands simple shapes on a large-screen graphics display surface. This approach combined gestures and voice commands to interact with the system. However, just recently such interaction paradigm have been introduced in off-the-shelf commodity products.

Recent technological advances allowed development of low-cost, lightweight, easy to use systems. With limited resources, novel and more natural HCI can be developed and explored. For instance, Lee [10] used a Wiimote and took advantage of its high resolution infra-red camera to implement multipoint interactive whiteboard, finger tracking and head tracking for desktop virtual reality displays. Post-WIMP finally arrived to the masses.

[3]http://sketchup.google.com/3dwarehouse/
[4]http://www.nintendo.com/wii/console/controllers
[5]http://www.xbox.com/en-US/kinect

Figure 1: The interface of 3D MARS.

Generally, post-WIMP approaches abandoned the traditional mouse and keyboard combination, favouring devices with six degrees of freedom (DoF). Unlike traditional WIMP interaction style, where it is necessary to map the inputs from a 2D interaction space to a 3D visualization space, six DoF devices allow straightforward direct mapping between device movements and rotations and corresponding effects on the three-dimensional space. This represents an huge leap to the concept of direct manipulation, which, according to Shneiderman [14], rapidly increments operations and allows the immediate visualization of effects on an manipulated object. This helps making the interaction more comprehensible, predictable and controllable.

Combining six DoF devices with stereoscopy, it is possible to make a multi-modal immersive interaction with direct and natural manipulation of objects shapes within virtual environments. This may be experienced using immersive displays (e.g., HMDs, CAVEs) [7] or desktop [15].

Despite the growing interest around the application of this new paradigms in HCI, no relevant efforts were made to explore the latest technological advances for multimedia information retrieval. Indeed, to the extent of our knowledge, there has not been presented any research or new solution that take advantage of immersive virtual environments for information retrieval since Nakazato's 3D MARS [11] .

4. 3D MARS

The 3D MARS system demonstrates that the use of 3D visualization in multimedia retrieval has two benefit. First, more content can be displayed at the same time without occluding one another. Second, by assigning different meanings to each axis, the user can determine which features are important as well as examine the query result with respect to three different criteria at the same time.

Nakazato focused his work on query result visualization. Thus 3D MARS supports only query-by-example mechanism to specify the search. The user select one image from a list and the system retrieves and displays the most similar images from the image database in a 3D virtual space. The image location on this space is determined by its distance

to the query image, where more similar images are closer to the origin of the space. The distance in each coordinate axis depend on a pre-defined set of features. The X-axis, Y-axis and Z-axis represent color, texture and structure of images respectively.

The interaction with the query results is done through a wand that the user holds while freely walking around the CAVE, as depicted in Figure 1. By wearing shutter glasses, the user can see a stereoscopic view of the world, which provides a full immersive experience. In such solution, visualizing query results goes far beyond scrolling on a list of thumbnails. The user navigates among the results in a three-dimensional space.

The 3D MARS was a catalyst for the incitement proposed in this paper: explore immersive visualization systems for multimedia information retrieval. Following that idea, we devised an immersive 3D virtual reality system for the display of query results of queries for 3D object Retrieval.

5. IMMERSIVE 3DOR

Taking advantage of the new paradigms in HCI, we propose an immersive VR system for 3D object retrieval (**Im-O-Ret**). The version of the system presented in this paper relies on a large-screen display, the LEMe Wall [6], and the a six DoF interaction device, the SpacePoint Fusion, an off-the-shelf device developed by *PNI Sensor Corporation*. However, minimal effort is required in order to have the system working in a context with HMD glasses or stereoscopic glasses, as well as using other input devices, such as Wiimote or Kinect.

Regardless of the hardware details, the **Im-ORet** allows the user to browse the results of a query to collection of 3D objects in an immersive virtual environment. The objects are distributed in the virtual 3D space according to their similarity. This is measured by the distance of each result to the query, which stands in the origin of the coordinates. To each of the three axis is assigned a different shape matching algorithm. The similarity to the query returned by the corresponding algorithm determines the coordinate. Current version of **Im-O-Ret** uses the Lightfield Descriptors [3] on the X-axis, the Coord and Angle Histogram [13] for the Y-axis, the Spherical Harmonics Descriptor [9] for the Z-axis. Figure 2 illustrates a user browsing the results of a query.

5.1 Possibilities

Similar to the 3D MARS, this work opens a myriad of new possibilities. By assigning different shape matching algorithms to each axis, one can adapt the query mechanism to specific domains, producing more precise results. Applying transparency to results, it is possible to overlay results of distinct queries. Adding effects to results, such as glow or special colors, it order to convey additional information.

Since query results are not images or thumbnails, but three-dimensional models, it is possible to navigate around them in the virtual environment and even manipulate them. Moreover, instead of a static view of the result, displaying it as a 3D object that can be rotating over one axis, offers a better perception of the model. Adding stereoscopy will improve even more the visualization since the user gains depth perception over the environment.

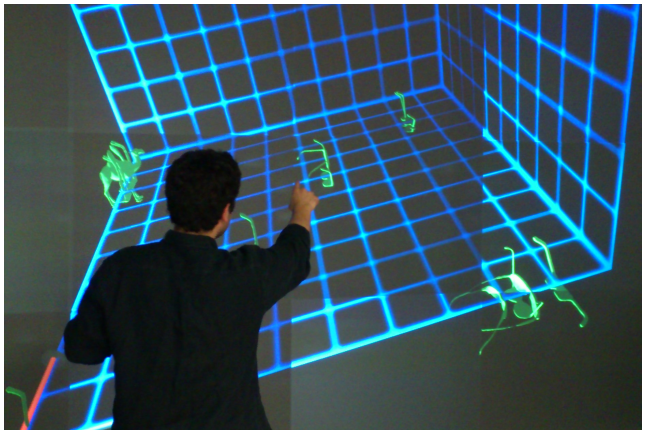


Figure 2: User exploring query results in Im-O-Ret

The combined use of VE and devices with six DoF, provides a more complete visualization and makes interaction more natural, comprehensible and predictable. Their use, will also add some challenges to the implementation of such system.

5.2 Challenges

While in traditional 3DOR systems the query results are represented and ordered as a list of thumbnails ordered by a given similarity measure, when we move to a virtual environment, the distribution of results in a 3D space becomes a challenge. How query results should arranged in 3D space to be meaningful to the user remains an open question. In our approach we select three shape descriptors and assigned each one to a coordinate axis, but this is a preliminary approach. We believe that a final solution is more complex that this. Further investigation on this topic is clearly required.

On the other hand, the way users navigate and interact with objects in an immersive environment and interact with it still an open issue. Norman[12] stated that gesturing is a natural, automatic behaviour, but the unintended interpretations of gestures can create undesirable states. Having this in mind, it is important to aim for an interface that is both predictable and easy to learn.

Above all, an important challenge remains open. No easy query specification mechanism has been presented, neither in traditional search engines, nor with new HCI paradigms. Although sketch-based queries apparently provide good results, they greatly depend on the ability of the user to draw a 3D model, which hinders the goal of a widely used, content-based, 3D search engine.

6. CONCLUSIONS

We believe that recent advances in low-cost, post-WIMP enabler technology, can be seen as an opportunity to overcome some drawbacks of current multimedia information retrieval solutions. Combined with the dissemination of stereoscopic visualization as a commodity, these interaction paradigms will acquaint common users with immersive virtual reality environments.

In this paper we highlight that such scenario is a fertile ground to be explored by search engines for multimedia information retrieval. In that context, we identified two major research topics: query result visualization and query specification. While the latest requires further study, we already started tackling the first one.

We developed a novel visualization approach for 3D object retrieval. The **Im-O-Ret** offers the users an immersive virtual environment for browsing results of a query to a collection of 3D objects. The query results are displayed as 3D models in a 3D space, instead of the traditional list of thumbnails. The user can explore the results, navigating in that space and directly manipulating the objects.

Looking back to 3D MARS, the initial work proposed by Nakazaro, we realize it was a valid idea that fell almost into obliviousness. We expect that our preliminary work, which lies over concepts introduced by 3D MARS, could prove the goodness of our incitement to explore the possibilities offered by immersive virtual environments to the multimedia information retrieval.

7. ACKNOWLEDGMENTS

The work described in this paper was partially supported by the Portuguese Foundation for Science and Technology (FCT) through the project 3DORuS, reference PTDC/EIA-EIA/102930/2008 and by the INESC-ID multiannual funding, through the PIDDAC Program funds.

8. REFERENCES

[1] T. F. Ansary, J.-P. Vandeborre, and M. Daoudi. 3d-model search engine from photos. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 89–92, New York, NY, USA, 2007. ACM.

[2] R. A. Bolt. Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '80, pages 262–270, New York, NY, USA, 1980. ACM.

[3] D.-Y. Chen, X.-P. Tian, Y. te Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. volume 22 of *EUROGRAPHICS 2003 Proceedings*, pages 223–232, 2003.

[4] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. Surround-screen projection-based virtual reality: the design and implementation of the cave. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '93, pages 135–142, New York, NY, USA, 1993. ACM.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40:5:1–5:60, May 2008.

[6] B. R. de AraÃžjo, T. Guerreiro, R. J. Costa, J. A. P. Jorge, and J. M. Pereira. Leme wall: Desenvolvendo um sistema de multi-projecção. 13Âž Encontro PortuguÃłs de ComputaÃğÃčo GrÃąfica, Vila Real, Portugal, 2005.

[7] T. DeFanti, D. Acevedo, R. Ainsworth, M. Brown, S. Cutchin, G. Dawe, K.-U. Doerr, A. Johnson, C. Knox, R. Kooima, F. Kuester, J. Leigh, L. Long, P. Otto, V. Petrovic, K. Ponto, A. Prudhomme, R. Rao, L. Renambot, D. Sandin, J. Schulze, L. Smarr, M. Srinivasan, P. Weber, and G. Wickham. The future of the cave. *Central European Journal of Engineering*, 1:16–37, 2011. 10.2478/s13531-010-0002-5.

[8] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3d models. *ACM Trans. Graph.*, 22:83–105, January 2003.

[9] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, SGP '03, pages 156–164, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[10] J. Lee. Hacking the nintendo wii remote. *Pervasive Computing, IEEE*, 7(3):39 –45, july-sept. 2008.

[11] T. S. H. Munehiro Nakazato. 3d mars: Immersive virtual reality for content-based image retrieval. In *Proceedings of 2001 IEEE International Conference on Multimedia and Expo (ICME2001)*, 2001.

[12] D. A. Norman. Natural user interfaces are not natural. *interactions*, 17:6–10, May 2010.

[13] E. Paquet and M. Rioux. Nefertiti: a query by content software for three-dimensional models databases management. In *NRC 97: Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, page 345, Washington, DC, USA, 1997. IEEE Computer Society.

[14] B. Shneiderman. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd international conference on Intelligent user interfaces*, IUI '97, pages 33–39, New York, NY, USA, 1997. ACM.

[15] B. Sousa Santos, P. Dias, A. Pimentel, J.-W. Baggerman, C. Ferreira, S. Silva, and J. Madeira. Head-mounted display versus desktop for 3d navigation in virtual reality: a user study. *Multimedia Tools Appl.*, 41:161–181, January 2009.

[16] A. van Dam. Post-wimp user interfaces. *Commun. ACM*, 40:63–67, February 1997.

The Mosaic Test: Benchmarking Colour-based Image Retrieval Systems Using Image Mosaics

William Plant
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

Joanna Lumsden
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

Ian T. Nabney
School of Engineering and Applied Science
Aston University
Birmingham, U.K.

ABSTRACT

Evaluation and benchmarking in content-based image retrieval has always been a somewhat neglected research area, making it difficult to judge the efficacy of many presented approaches. In this paper we investigate the issue of benchmarking for *colour-based image retrieval* systems, which enable users to retrieve images from a database based on low-level colour content alone. We argue that current image retrieval evaluation methods are not suited to benchmarking colour-based image retrieval systems, due in main to not allowing users to reflect upon the suitability of retrieved images within the context of a creative project and their reliance on highly subjective ground-truths. As a solution to these issues, the research presented here introduces the *Mosaic Test* for evaluating colour-based image retrieval systems, in which test-users are asked to create an image mosaic of a predetermined target image, using the colour-based image retrieval system that is being evaluated. We report on our findings from a user study which suggests that the Mosaic Test overcomes the major drawbacks associated with existing image retrieval evaluation methods, by enabling users to reflect upon image selections and automatically measuring image relevance in a way that correlates with the perception of many human assessors. We therefore propose that the Mosaic Test be adopted as a standardised benchmark for evaluating and comparing colour-based image retrieval systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*; H.2.8 [**Database Management**]: Database Applications—*Image Databases*

Keywords

Image databases, content-based image retrieval, image mosaic, performance evaluation, benchmarking.

1. INTRODUCTION

Colour-based image retrieval systems such as Chromatik [1], MultiColr [5] and Picitup [10] enable users to retrieve images from a database based on colour content alone. Such a facility is particularly useful to users across a number of different creative industries, such as graphic, interior and fashion design [6, 7]. Surprisingly, however, little research appears to have been conducted into evaluating colour-based image retrieval systems. Currently, there is no standardised measure and image database to evaluate the performance of an image retrieval system [8]. The most commonly applied evaluation methods are those of *precision and recall* [8] and the *target search* and *category search* tasks [11]. The precision and recall measure is used to evaluate the accuracy of image results returned by a system in response to a query, whilst the target search and category search tasks are both user-based evaluation strategies in which test-users are asked to retrieve images from a database that are relevant to a given target, using the image retrieval system that is being evaluated.

In this research, we argue that the image retrieval system evaluation strategies listed above are not suitable for evaluating and benchmarking colour-based image systems for two fundamental reasons. Firstly, none of the above evaluation methods allow test-users to perform an important process often conducted by creative users, known as *reflection-in-action* [12]. In reflection-in-action, a creative project is modified by a user and then reviewed by the user after the modification. After assessing their modification, the creative individual will then decide whether to maintain or discard the modification to the project. As an example, a graphic designer will add an image to a web page before making an assessment as to its aesthetic suitability. Secondly, the category search and precision and recall measures require an image database and associated ground-truth (a manually generated list pre-defining which images in the database are similar to others) for defining image relevance during a system evaluation. Such human-based definitions of similarity, however, can often be highly subjective resulting in retrieved images being incorrectly assessed as irrelevant.

As a result of these drawbacks, no method currently exists for reliably evaluating colour-based image retrieval systems. The following section introduces the Mosaic Test which has been developed to address the current problem, providing a reliable means for benchmarking colour-based image retrieval systems.

2. THE MOSAIC TEST

For the Mosaic Test, participants are asked to manually create an image mosaic (comprising 16 cells) of a predetermined target image. An image mosaic (first devised by Silvers [14]) is a form of art that is typically generated automatically through use of content-based image analysis. A target image is divided into cells, each of which is then replaced by a small image with similar colour content to the corresponding cell in the target image. Viewed from a distance, the smaller images collectively appear to form the target image, whilst viewing an image mosaic close up reveals the detail contained within each of the smaller images. An example of an automatically generated image mosaic is shown in Figure 1.

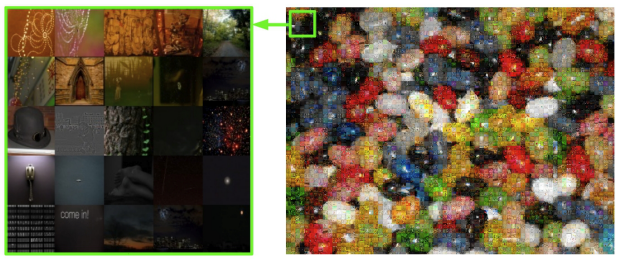


Figure 1: An example of an image mosaic. The region highlighted green in the image mosaic (right) has been created using the images shown (left).

For target images in the Mosaic Test, photographs of jelly beans are used. The images of jelly beans produce a bright, interesting target image for participants to create in mosaic form and the generation of an image mosaic that appears visually similar to the target image is also very achievable. More importantly, retrieving images from a database comprising large areas of a small number of distinct colours is a practise commonly performed by users in creative industries.

To complete their image mosaics, participants must identify the colours required to fill an image mosaic cell (by inspecting the corresponding region in the target image), and retrieve a suitably coloured image from the 25,000 contained within the MIRFLICKR-25000 image collection [4] using the colour-based evaluation system under evaluation. When selecting images for use in their image mosaic, users can add, move or remove images accordingly to assess the suitability of images within the context of their image mosaic. It is in this way that the Mosaic Test overcomes the first major drawback of existing evaluation methods, by enabling participants to perform the creative practise of reflection-in-action [12]. Upon completion of an image mosaic, the time required by the user to finish the image mosaic is recorded, along with the visual accuracy of their creation in comparison with the initial target image. Through analysing the accuracy of user-generated image mosaics (in a manner which correlates with the perception of a number of different human assessors), the Mosaic Test is able to overcome the second drawback associated with existing evaluation techniques. This is because it does not rely on a highly subjective image database ground-truth. The image mosaic accuracy measure adopted for use with the Mosaic Test is discussed further in Section 3.1. Additionally, participants are asked to indicate their subjective experience of workload (using the NASA TLX scales [2]) post test.

The time (number of seconds), subjective workload (user NASA-TLX ratings) and relevance (image mosaic accuracy) measures achieved by colour-based image retrieval systems evaluated using the Mosaic Test can be directly compared and used for benchmarking. When comparing the Mosaic Test measures achieved by different systems, the more effective colour-based image retrieval system will be the one that enables users to create the most accurate image mosaics, fastest and with the least workload.

2.1 Mosaic Test Tool

To support users in their manual creation of image mosaics using the Mosaic Test, we have developed a novel software tool in which an image mosaic of a predetermined target image can be created using simple drag and drop functions. We refer to this as the *Mosaic Test Tool*. The Mosaic Test Tool has been designed so that it can be displayed simultaneously with the colour-based image retrieval system under evaluation (as can be seen in Figure 2). This removes the need for users to constantly switch between application windows, and permits users to easily drag images from the colour-based image retrieval system being tested to their image mosaic in the Mosaic Test Tool. It is important to note that the facility to export images through drag and drop operations is the only requirement of a colour-based image retrieval system for it to be compatible with the Mosaic Test Tool and thus the Mosaic Test.

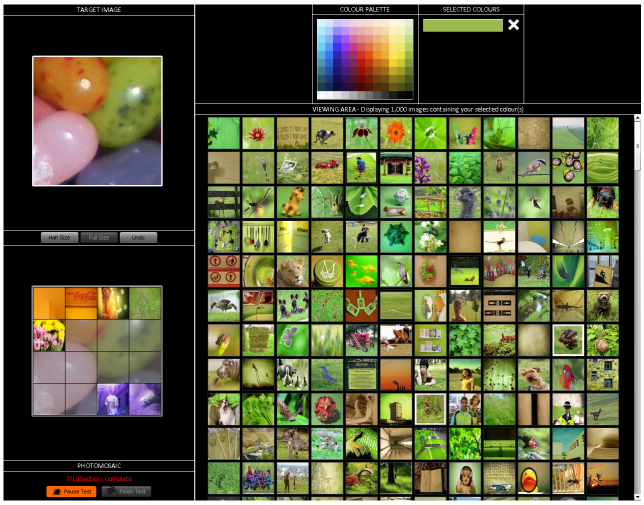


Figure 2: The Mosaic Test Tool (left) and an image retrieval system under evaluation (right) during a Mosaic Test session.

The target image and image mosaic are displayed simultaneously on the Mosaic Test Tool interface to allow users to manually inspect and identify the colours (and colour layout) required for each image mosaic cell. As can be seen in Figure 2, the target image (the image the user is trying to replicate in the form of an image mosaic) is displayed in the top half of the Mosaic Test Tool. Coupled with the ease in which images can be added to, or removed from, image mosaic cells, users of the Mosaic Test Tool can simply as-

sess the suitability of a retrieved image by dragging it to the appropriate image mosaic cell and viewing it alongside the other image mosaic cells.

3. USER STUDY

To evaluate the Mosaic Test, we recruited 24 users to participate in a user study. Participants were given written instructions explaining the concept of an image mosaic and the functionality of the Mosaic Test Tool. A practise session was undertaken by each participant, in which they were asked to complete a practise image mosaic using a small selection of suitable images. Participants were then asked to complete 3 image mosaics using 3 different colour-based image retrieval systems. To ensure that users did not simply learn a set of database images suitable for use in a solitary image mosaic, 3 different target images were used. These target images were carefully selected so that the number of jelly beans (and thus colours) in each were evenly balanced, with only the colour and layout of the jelly beans varying between the target images. To also ensure that results were not effected by a target image being more difficult to create in image mosaic form than another, the order in which the target images were presented to participants remained constant whilst the order in which the colour-based image retrieval systems were used was counter balanced. After completing the 3 image mosaics, participants were asked to rank each of their creations in ascending order of 'closeness' to its corresponding target image.

We wanted to investigate whether the Mosaic Test does overcome the drawbacks of existing evaluation strategies so that it may be adopted as a reliable benchmark of colour-based image retrieval systems. Firstly, we hypothesised that users in the study would perform reflection-in-action and so we wanted to observe whether this was indeed true for participants when judging the suitability of images retrieved from the database. Secondly, we were eager to investigate which method should be adopted for measuring the accuracy of an image mosaic in the Mosaic Test.

3.1 Assessing Image Mosaic Accuracy

As an image mosaic is an art form intended to be viewed and enjoyed by humans, it seems logical that the adopted measure of image mosaic accuracy - i.e., how close an image mosaic looks to its intended target image - should correlate with the inter-image distance perceptions of a number of human assessors. An existing measure for automatically computing the distance between an image mosaic and its corresponding target image is the *Average Pixel-to-Pixel* (APP) distance [9]. The APP distance is expressed formally in Equation (1), where i is 1 of a total n corresponding pixels in the mosaic image M and target image T, and r, g and b are the red, green and blue colour values of a pixel.

$$APP = \frac{\sum_{i=0}^{n} \sqrt{(r_M^i - r_T^i)^2 + (g_M^i - g_T^i)^2 + (b_M^i - b_T^i)^2}}{n} \quad (1)$$

We were eager to compare the existing APP image mosaic distance measure with a variety of image colour descriptors (and associated distance measures) commonly used for content-based image retrieval, to discover which best correlates with human perceptions of image mosaic distance. To do this, we calculated the image mosaic distance rankings according to the existing measure and several colour descriptors (and their associated distance measures), and then calculated the Spearman's rank correlation coefficient between each of the tested distance measures and the rankings assigned by the users in our study.

For the image colour descriptors (and associated distance measures), we firstly tested the global colour histogram (GCH) as an image descriptor. A colour histogram contains a normalised pixel count for each unique colour in the colour space. We used a 64-bin histogram, in which each of the red, green and blue colour channels (in an RGB colour space) were quantised to 4 bins (4 x 4 x 4 = 64). We adopted the Euclidean distance metric to compare the global colour histograms of the image mosaics and corresponding target images. We also tested local colour histograms (LCH) as an image descriptor. For this, 64-bin colour histograms were calculated for each image mosaic cell (for the image mosaic descriptor), and its corresponding area in the target image (for the target image descriptor). The average Euclidean distance between all of the corresponding colour histograms (in the image mosaic and target image LCH descriptors) was used to compare LCH descriptors. Finally, we tested (along with their associated distance measures) the MPEG-7 colour structure (MPEG-7 CST) and colour layout (MPEG-7 CL) descriptors [13], as well as the auto colour correlogram descriptor (ACC) [3].

The auto colour-correlogram (ACC) of an image can be described as a table indexed by colour pairs, where the k-th entry for colour i specifies the probability of finding another pixel of colour i in the image at a distance k. For the MPEG-7 colour structure descriptor (MPEG-7 CST), a sliding window (8×8 pixels in size) moves across the image in the HMMD colour space [13] (reduced to 256 colours). With each shift of the structuring element, if a pixel with colour i occurs within the block, the total number of occurrences in the image for colour i is incremented to form a colour histogram. The distance between two MPEG-7 CSTs or two ACCs can be calculated using the L_1 (or city-block) distance metric. Finally, the MPEG-7 colour layout descriptor (MPEG-7 CL) [13] divides an image into 64 regular blocks, and calculates the dominant colour of the pixels within each block [13]. The cumulative distance between the colours (in the YC_bC_r colour space) of corresponding blocks forms the measure of similarity between 2 MPEG-7 CL descriptors.

Accuracy Measure	$\mathbf{r_s}$	**Significant (5%)**
MPEG-7 CST	0.572	YES
APP	0.275	NO
GCH	0.242	NO
MPEG-7 CL	0.198	NO
LCH	0.176	NO
ACC	0.154	NO

Table 1: The Spearman's rank correlation coefficients (r_s) between the image mosaic distance rankings made by humans and the rankings generated by the tested colour descriptors.

4. RESULTS

Table 1 shows the Spearman's rank correlation coefficients (r_s) calculated between the human-assigned rankings and each of the rankings generated by the tested colour descriptors. We compare the r_s correlation coefficient for each measure tested with the critical value of r, which at a 5% significance level with 22 d.f. $(24 - 2)$ equates to **0.423**. Any r_s value greater than this critical value can be considered a significant correlation at a 5% level.

5. DISCUSSION

We observed the actions taken by the participants of the user study when creating their image mosaics. It was clear that the majority of users performed reflection-in-action when assessing the relevance (or suitability) of images retrieved from the database for use in their image mosaics. As participants of a Mosaic Test were able to perform this reflection-in-action [12], it is clear that the Mosaic Test also overcomes the first of the two major drawbacks present in current image retrieval evaluation methods. As shown in Table 1, the MPEG-7 colour structure descriptor (MPEG-7 CST) was the only colour descriptor (and associated distance measure) we found to correlate with human perceptions of image mosaic distance at the 5% significance level. Therefore, by measuring the L_1 (or city-block) distance between the MPEG-7 CSTs of the target image and user-generated image mosaics, the Mosaic Test can automatically calculate the relevance of retrieved images in a manner that correlates with human perception, thus overcoming the second major drawback of existing image retrieval evaluation methods for benchmarking colour-based image retrieval systems (the reliance on a highly subjective image database ground-truth).

6. CONCLUSION

Current image retrieval system evaluation methods have two fundamental drawbacks that result in them being unsuitable for evaluating and benchmarking colour-based image retrieval systems. These evaluation strategies do not enable users to perform the practise of reflection-in-action [12], in which creative users assess project modifications within the context of the creative piece he/she is working on. The existing image retrieval system evaluation methods also rely heavily upon highly subjective image database ground-truths when assessing the relevance of images selected by test users or returned by a system. As a result of these drawbacks, no method currently exists for reliably evaluating and benchmarking colour-based image retrieval systems. In this paper, we have introduced the Mosaic Test which has been developed to address the current problem, by providing a reliable means by which to evaluate colour-based image retrieval systems.

The findings of a user study reveal that the Mosaic Test overcomes the two major drawbacks associated with existing evaluation method used in the research domain of image retrieval. As well as also providing valuable effectiveness data relating to efficiency and user workload, the Mosaic Test enables participants to reflect on the relevance of retrieved images within the context of their image mosaic (i.e., perform reflection-in-action [12]). The Mosaic Test is also able to automatically measure the relevance of retrieved images in a manner which correlates with the perceptions of multiple human assessors, by computing MPEG-7 colour structure descriptors from the user-generated image mosaics and their corresponding target images, and calculating the L_1 (or city-block) distance between them. As a result of our findings, we propose that the Mosaic Test be adopted in all future research evaluating the effectiveness of colour-based image retrieval systems. Future work will be to publicly release the Mosaic Test Tool and procedural documentation for other researchers in the domain of content-based image retrieval.

7. REFERENCES

[1] Exalead. Chromatik. Accessed December 1, 2010, at: `http://chromatik.labs.exalead.com/`.

[2] S. G. Hart. NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pages 904–908, 2006.

[3] J. Huang, S. R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In *Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[4] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *ACM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[5] idée Inc. idée MultiColr Search Lab. Accessed November 2, 2010 at `http://labs.ideeinc.com/multicolr`.

[6] Imagekind Inc. Shop Art by Color. Accessed November 2, 2010, at: `http://www.imagekind.com/shop/ColorPicker.aspx`.

[7] T. K. Lau and I. King. Montage : An Image Database for the Fashion, Textile, and Clothing Industry in Hong Kong. In *Third Asian Conference on Computer Vision*, pages 410–417, 1998.

[8] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[9] S. Nakade and P. Karule. Mosaicture: Image Mosaic Generating System Using CBIR Technique. In *International Conference on Computational Intelligence and Multimedia Applications*, pages 339–343, 2007.

[10] Picitup. Picitup. Accessed January 21, 2011, at: `http://www.picitup.com/`.

[11] W. Plant and G. Schaefer. Evaluation and Benchmarking of Image Database Navigation Tools. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, pages 248–254, 2009.

[12] D. A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, 1983.

[13] T. Sikora. The MPEG-7 Visual Standard for Content Description - An Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 2001.

[14] R. Silvers. Photomosaics: Putting Pictures in their Place. Master's thesis, Massachusetts Institute of Technology, 1996.

Evaluating the Cognitive Impact of Search User Interface Design Decisions

Max L. Wilson
Future Interaction Technology Labs
Department of Computer Science, College of Science
Swansea University, UK
m.l.wilson@swansea.ac.uk

ABSTRACT
The design of search user interfaces has developed dramatically over the years, from simple keyword search systems to complex combinations of faceted filters and sorting mechanisms. These complicated interactions can provide the searcher with a lot of power and control, but at what cost? Our own work has seen users experience a sharp learning curve with faceted browsers, even before they begin interacting. This paper describes a forthcoming period of work that intends to investigate the cognitive impact of incrementally adding features to search user interfaces. We intend to produce search user interface design recommendations to help designers maximize support for searchers while minimizing cognitive impact.

Author Keywords
Search, Exploratory Search, User Interface Design, Cognitive Load Theory

ACM Classification Keywords
H5.2. Information interfaces and presentation (User Interfaces): evaluation/methodology, screen design. H3.3. Information search and retrieval: Search process.

INTRODUCTION

User Interface (UI) Designers are always concerned with supporting users effectively and intuitively, but a common recent focus for Search User Interface (SUI) designers has been to increase the interactive power and control that searchers have over results. As a community, we want to support users in exploring, discovering, comparing, and choosing results that meet their needs. SUI designers, therefore, are concerned with maximizing the use of powerful interface features while maintaining a clear and intuitive design.

In our prior work, we developed mSpace [7] as a faceted browser that lets searchers use combinations of orthogonal metadata filters to narrow their search. We developed advanced interactions for faceted browsers that took advantage of visual location within the SUI, and highlighted options in unused filters that were related to guide searchers [10]. Frequently, however, we informally noted that searchers spent increasing periods of time on visually comprehending the interface before making their first move. In follow up studies, we saw minimal interaction with facets during the first visit, but recorded a significant increase in the use of faceted features during subsequent return visits. It is the hypothesis of our forthcoming work that this non-use of such powerful features is caused by an increased cognitive load created by the associated increased complexity of the SUI. It is this cognitive impact that we believe can be measured and attributed to specific design decisions.

mSpace is one specific faceted browser, but the principle of faceted browsing can be implemented in many different ways [2]. We also hypothesize that not only the presence, but also the subsequent design of SUI features can also have an impact. The following sections cover some related work before describing our plans to evaluate the cognitive impact that adding features to SUIs can have.

RELATED WORK

SUI design is affected by many factors. Interaction designers can decide how best to support searchers, but designs may be limited by the metadata that is available about the possible results. Both the underlying data and the graphical design may also have an impact, then, on how the chosen interaction will look and feel. As perhaps the most recognized SUI for many users around the world, Google has always maintained a very clean and clear white design[1], and make very incremental careful design changes that stay within that design. Competitor search engines have notably changed over the years, with many now being very similar to Google in terms of interaction design, while trying to keep their own visual design consistent.

For more exploratory websites that sell a wide range of products, or provide large collections of information or documents, there are now many different features that support people, from tabular or dropdown-based sorting

[1] http://searchengineland.com/qa-with-marissa-mayer-google-vp-search-products-user-experience-10370

mechanisms, to categories, clusters, filters, and facets. Some websites that provide these features are frustrating and difficult to use, while others are simple, intuitive, and successful. In these systems it is often the way that the ideal support has been developed that has affected their success. In a study of the success of different faceted browser implementations, Capra et al [1] directly compared two faceted browsers to a government website, all over the same hierarchical government dataset, and discovered that the customized hierarchical design of the original website supported searchers far better than the functionally more powerful faceted browsers.

Both the choice of content and the visual design have both been shown to have an impact on usability. White et al showed that the text that includes the search terms is best, and that highlighting these terms also improves search [12]. Similarly, Lin et al. have shown that simply highlighting the domain name in the URL bar significantly reduces the chances that users will be caught be fishing attacks [4]. Zheng et al [13] have also shown that users can make often-accurate snap judgments about the credibility of websites within half a second. Further, Wilson et al [10] noted that the success of adding guiding highlights to their faceted browser was affected by the choice of highlight-colour and its implied meaning.

The choice of SUI features within a single implementation has also been shown to have an impact on search success. Diriye et al compared a keyword search interface with a revised version that also included query suggestions [3]. Their results showed that such features slowed down searchers who were performing simple lookup tasks, but supported those who were performing more complicated exploratory tasks. Similarly, Wilson and Wilson have also found early results indicating that the simple presence, without interaction, of a keyword cloud provides additional support, where subsequent interaction provides very little gain [11] during exploratory tasks. Wilson and Wilson's results suggest that searchers can learn more about the result set from seeing the terms in the keyword cloud, than actually using them to filter the results.

The location of features within a SUI has also been shown to have an impact. Morgan and Wilson studied the visual layout of search thumbnails, predicting that having a rack of thumbnails at the top of the user interface would allow searchers to make faster judgments when trying to re-find pages [5]. Their results showed that a rack of thumbnails was significantly more disruptive to searchers when the target page was not in the results, than the support it provided when it was.

The studies above indicate that the success of SUIs can be attributed to the appropriateness of the functionality provided, where unnecessary functionality can slow users down. Further, the studies indicate that the success of SUIs can be determined by simple visual or spatial changes that do not necessarily impact functionality. Consequently, where two systems provide the same support, one may be harder or easier to use because of its simple visual design. Our conclusion is that to understand the success of a SUI, we must analyse both the support in terms of functionality, and the cognitive impact is creates. Being able to understand and predict these two things would help us to design and build better SUIs

EVALUATING THE SUPPORT PROVIDED BY SUIS

Beyond the common practice of performing task-oriented user studies, my own doctoral work focused on the design of an analytical evaluation metric for SUIs, called the Search Interface Inspector[2] (Sii). Sii calculates the support for different types of users based upon the set of features in the interface, and how many interactions they take to use [9]. To analyse a SUI, the evaluator catalogues the features of the design and calculates how many interactions are required to perform a set of known search tactics. The method then interpolates the likely support for different types of searchers (explorers or searchers that know what they are looking for, for example), based upon the types of tactics they are likely to perform. Sii can be used to compare several designs and produces a series of 3 interactive graphs that allow evaluators perform an investigative analysis of the results.

Sii is based on detailed established information seeking theory and rewards the design of search functionality that has simple interaction. Consequently, however, Sii rewards the addition of new simple functionality, without being able to estimate the increasing complexity of the SUI as new features are added. To remedy this problem, a chapter of the thesis investigated Cognitive Load Theory and initially specified a similar metric that calculated the cognitive load of a UI. This second measure of intrinsic cognitive load was proposed for inclusion in Sii, estimated the intrinsic cognitive load of a SUI. Similar to how the original metric was correlated with study results, one aim of the work described below is to further refine and validate this analytical measure of the cognitive impact of SUIs.

Cognitive Load Theory highlights that capacity for learning is affected by three aspects: intrinsic, extrinsic, and germane cognitive load. Intrinsic cognitive load is created by the materials providing the learning experience, or in our case the SUI. Extrinsic cognitive load is created by the complexity in the task at hand. Germane cognitive load is then required to process what is learned and commit it to long-term memory. If intrinsic load and extrinsic load are too high, then there may not be enough space load left for germane cognitive load. Although, it is commonly accepted that effort can increase overall capacity, the aim should still be to reduce intrinsic cognitive load by improving the design of learning materials or SUIs [6]. Reducing intrinsic load creates space for users to perform increasingly

[2] http://mspace.fm/sii

complex tasks, or opens-up germane cognitive load so that what is being learned can be retained.

EVALUATING THE COGNITIVE IMPACT OF SUIS

The general structure of the studies we are planning is to use brain scanners to record the cognitive impact that different SUIs have on a user. The initial phases will focus on identifying and measuring such responses to significant and obvious differences, before trying to capture changes to more subtle designs and, hopefully, in-situ. Initially, we will be using EPOC Emotiv headsets[3], as shown in Figure 1, to take readings. These headsets are commercialized versions of EEG scanners, but are designed for use in more natural contexts. EEG scanners, as with many other brain scanning systems, are typically affected by simple body movements and so are often restricted to confined conditions. Such scanners, therefore, are often not suitable for task-based evaluations, which require action and movement. In psychology, EEG scanners are typically used in constrained environments where users are only allowed to move their thumbs to answer yes or no. Consequently, this work requires scanners that can be used in more natural contexts while performing everyday searching tasks. In the future, funding permitting, we also intend to buy an fNIR scanner, which has been shown to be suitable for task-based evaluation conditions [8]. We intend to use these measurements to understand the impact of design decisions, in order to make clear recommendations to SUI designers.

Figure 1: EPOC Emotiv Headset

Phase 1 – the impact of additional features

Beginning this summer, with two summer interns, we will be performing our first studies, which will simply display SUIs of incremental complexity to participants. We will begin with a simple keyword search design, and add features such as recommendations and filters. The order that interfaces are shown to participants will be randomized to avoid learning and familiarity bias. The aim of this phase is to prove that the learning curves experienced by users exist and the cognitive load can be measured objectively. We hope that the results will show initial insight into the amount of impact that different features have, which may in turn help us make hypotheses about design issues. This phase will help us identify the cost of adding a feature, where task success would allow us to measure their benefit.

Phase 2 – capturing impact in the context of tasks

Where the first phase above allows us to learn to recognize the signs from EEG signals, we intend to try and detect cognitive load in situ, and in the context of a task. We will be setting participants specific simple and exploratory tasks, whilst controlling the type of user interface features they see, to capture the cognitive impact as they start. This phase will help us identify whether the impact of a search user interface is affected by task context.

Phase 3 – the impact of different implementations

While adding features creates an obvious change in the user interface, different features can be put in different places in the SUI and also be implemented differently. Google, for example, puts suggested refinements at the bottom of the page, while Bing has them on the side. Bing also chooses to provide a mix of refinements and alternative directions. In Phase 2 we intend to analyse both of these kinds of variables to see if they have significant impacts on cognitive load. This phase will help us identify whether the cost of adding SUI features can be minimized by refining their design.

Discussion

There are many challenges remaining in this planned work. So far, we have planned very controlled comparisons of SUI changes, but in real life these systems are used in the context of complex tasks and for extended periods of time. Controlled situations will help identify cause and effect, but other similar objective measurements, like eye trackers, still require interpretation. We hope to expand on these methods, and the findings of existing brain scanning HCI research [8], by addressing this issue over time. Finally, although this research is primarily interested in the development of SUI interfaces and how they affect people learning to use powerful search features, there are many other things that can be distracting in general UI design. These methods will likely expand to help address other design questions; we, however, are particularly aiming to answer questions about encouraging exploratory search and learning, by increasing the power of SUIs, while reducing their impact on searchers.

CONCLUSIONS

This work has yet to begin formally, but we intend to learn more about the impact that very simple design decisions can have on searchers. From previous experience of searcher success in evaluations, both industry and academia know that such changes can seriously impact the success of a search user interface. This work will use objective measurements of brain response to help us identify the factors that make search user interfaces hard to comprehend. We hope that such measurements will a) help us analyse the cost-benefit trade-off of adding additional

[3] http://www.emotiv.com/

support to search user interfaces, and b) help us develop design recommendations for implementing search user interface features so that they have minimal impact.

REFERENCES

1. Robert Capra, Gary Marchionini, Jung Sun Oh, Fred Stutzman, and Yan Zhang. 2007. Effects of structure and interaction style on distinct search tasks. In *Proc. JCDL '07*. ACM, New York, NY, USA, 442-451.
2. Edward C. Clarkson, Shamkant B. Navathe, and James D. Foley. 2009. Generalized formal models for faceted user interfaces. In *Proc. JCDL '09*. ACM, New York, NY, USA, 125-134.
3. Abdigani Diriye, Ann Blandford, and Anastasios Tombros. 2010. Exploring the impact of search interface features on search tasks. In *Proc. ECDL'10*.
4. Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, John Aycock. Does Domain Highlighting Help People Identify Phishing Sites. In *Proc. CHI2011* (in press).
5. Rhys Morgan and Max L. Wilson. 2010. The Revisit Rack: grouping web search thumbnails for optimal visual recognition. In *Proc. ASIS&T '10*.
6. Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proc. MULTIMEDIA'06*. ACM, New York, NY, USA, 871-880.
7. m.c. schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. 2006. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Commun. ACM* 49, 4 (April 2006), 47-49.
8. Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M. Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert J.K. Jacob. 2009. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST '09*. ACM, New York, NY, USA, 157-166.
9. Max L. Wilson, M. C. schraefel, and Ryen W. White. 2009. Evaluating advanced search interfaces using established information-seeking models. *J. Am. Soc. Inf. Sci. Technol.* 60, 7 (July 2009), 1407-1422.
10. Max L. Wilson, Paul André, and mc schraefel. 2008. Backward highlighting: enhancing faceted search. In *Proc UIST '08*. ACM, New York, NY, USA, 235-238
11. Wilson, M. J. and Wilson, M. L. Tag Clouds and Keyword Clouds: evaluating zero-interaction benefits. In *Ext. Abstract CHI'11*.
12. Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proc. SIGIR '02*. ACM, New York, NY, USA, 57-64.
13. Xianjun Sam Zheng, Ishani Chakraborty, James Jeng-Weei Lin, and Robert Rauschenberger. 2009. Correlating low-level image statistics with users - rapid aesthetic and affective judgments of web pages. In *Proc. CHI '09*. ACM, New York, NY, USA, 1-10.

The potential of Recall and Precision as interface design parameters for information retrieval systems situated in everyday environments

Ayman Moghnieh
Universitat Pompeu Fabra
C/Tanger 122-140, E-08018
Barcelona, Spain
ayman.moghnie@upf.edu

Josep Blat
Universitat Pompeu Fabra
C/Tanger 122-140, E-08018
Barcelona, Spain
josep.blat@upf.edu

ABSTRACT
In this paper, we investigate ways for a tighter integration of IR and HCI in new urban contexts, as HCI expands its reach outside the workplace towards environments where efficiency and performance no longer constitute the backbone of interaction requirements. In particular, we propose to use Recall and Precision as design parameters to describe the information settings and performance of situated interfaces acting as retrieval systems in these environments. To explore this notion, we follow an inductive design research process by which different prototypes are designed, developed, and evaluated. Our experience shows that Recall and Precision, as design parameters, help to reflect the information requirements onto the interface design, and contribute to adapting IR to the contemporary challenges it faces, although more work is needed to consolidate its role vis-à-vis the growing ubiquity of computer technologies.

Categories and Subject Descriptors
H.5.2 User Interfaces.

General Terms
Design, Experimentation, Human Factors, Theory.

Keywords
Information Retrieval, Human-Information Interaction, Situated Interfaces, Interface and Interaction Design

1. INTRODUCTION
As computer technologies become more ubiquitous and versatile, and get further integrated in human environments, several genres of situated information interfaces (e.g. interactive peripheral displays, ambient displays, and interactive surfaces) are starting to assume a mediating role between people and digital information spaces in different environments. From an HCI perspective, these situated interfaces, primarily found in public and semi-public environments such as malls, public transportation, building entrances, and public squares, represent new border zones that maintain connectivity and mutual presence between the real and the digital worlds, and actively sustain flows of useful or relevant information towards nearby people who in-turn search, discover, and interact with the displayed information.

The human interaction with information via situated interfaces creates new challenges for conventional information retrieval (IR) systems: first, the relationship between people and digital information spaces becomes more explicit and the technology that supports it more ubiquitous. Second, the human interaction with information spaces adopts a more direct approach supported by the coming of age of new interaction paradigms (e.g. touch, gesture, speech) that emulate the manipulation of objects. Third, the information space hosted by a situated interface tends to be specialized in subjects and themes befitting the environment where the interface is situated, and the goals and interests of the people present in it. Fourth, the interaction properties may vary considerably in terms of interaction duration and the amount of user attention delegated to the situated interface [1].

These challenges, among others [2], justify the search for a tighter coupling of interface and interaction design, and IR systems, by which IR as a supporting technology for interacting with information contributes to making the interface design more transparent and the human-information interaction more fluid and direct. Therefore, we reason that the performance of situated interfaces as IR systems ought to be attuned according to the nature of each specific interaction scenario, given that a maximization of IR performance, may not be adequate for answering the interaction design requirements in all kinds of user experiences with situated interfaces [5, 10]. Consequently, IR performance tilts towards becoming a design issue that determines some of the characteristics of situated interfaces that mediate this interaction.

Currently, two metrics (Recall and Precision) are used to assess the performance of IR systems in response to user queries [3]. Recall is the fraction of retrieved information elements from the entire existing set of elements that are relevant to the user query in the information space. Precision is the fraction of retrieved elements found relevant with respect to the user query, over the entire set of retrieved elements. However, the query as a middleman between humans and information spaces goes against the transparent design of situated interfaces that support a direct interaction with information spaces. In addition, the information spaces hosted by situated interfaces are usually predetermined or pre-queried in accordance with the specific interests of potential

users and the characteristics or nature of the environments where the hosting interfaces are situated. Instead of querying, the explicit momentarily needs of users are answered by direct interaction with the visualized information. This superlatively converts the relevance of the displayed information to the user interests from a performance factor to a design issue.

Therefore, we argue that the definition of Recall and Precision can be loosened or reinterpreted to respectively describe the quantity of retrieved information elements and their visual diversity as displayed on the interface, since relevance is no longer a performance factor from an HCI stance. These two metrics can consequently act as parameters that bind the design and performance of situated interfaces as retrieval systems to the informational expectations of users, by controlling the amount and diversity of visualized information in order to maximize the transparency of their designs to support a direct human-information interaction.

In order to explore this idea further, we followed a line of inductive design research by conceptualizing, designing, and evaluating experimental prototypes. We first introduce two sets of prototypes devised to understand how users perceive the quantity and visible diversity of information objects. We then define parameterization scales for Recall and Precision based on these experiments. In order to develop a thoughtful understanding of how Recall and Precision, which we will consecutively refer to as R and P, can act as design parameters for situated interfaces, we use them in the analysis, design, and evaluation of five different situated interfaces. Next, we investigate how these two parameters can be dynamically controlled by users through the design of two interactive interfaces for searching and browsing news articles. We conclude by assessing our experience and discuss the viability and implications of our approach.

2. RECALL AND PRECISION FROM A PERCEPTUAL STANCE

Figure 1. An instance of the InformationCasserole prototypes

InformationCasserole is a series of video prototypes (figure 1) designed to study the effect that the number of visualized elements (R) has on the way humans perceive the information revealed on the interface. They show classified ads from magazines and newspaper floating on different levels in a glass container filled with slowly moving water. Therefore, their settings emulate a transparent interface design and foster a direct relationship between the human and digital information spaces.

Miller's Law argues that the total number of different objects that humans can simultaneously hold in their working memory is approximately seven [4]. This affects the manner by which information is perceived when the cardinality of the visualized set of objects increases. In particular, there is a natural observable tendency to perceptually cluster or group these objects recursively whenever the perceivable number exceed Miller's threshold. To observe this phenomenon, eight 10 minutes long think-aloud sessions were organized with eight different university students that watched InformationCasserole showing magazine ads progressively being added to the water container, and commented on how the number of ads shown in the casserole affects the way they perceive the set of visualized ads.

We observed that when one object is shown, it tends to engage the subjects in a prolonged and detailed examination. This changes when two to seven objects are displayed since subjects become more interested in identifying relations among the objects and comparing them. The interest in object relations abates with a higher object number, and instead the relations among clusters or collections of objects start to proportionally grab attention. When the number of visualized objects crosses a certain threshold, which we estimate at Miller's number squared, the casserole becomes perceptually saturated and the subjects begin to treat the set of ads as a space, reasoning about different regions in it. In conclusion, we find that the quantity of visualized objects (R) is perceived in four different density thresholds, and to each we accord a parameter value: R=0 for visualizing no or a single object; R=1 for a single collection of seven or less objects; R=2 for seven or less collections; and R=3 for single information space or more than seven squared objects. This is reflected in figure 2.

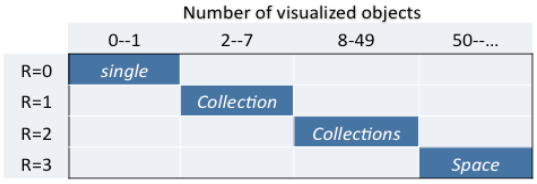


Figure 2. R as a design parameter

In order to study the effects that the visible diversity of information objects (P) has on the manner by which people perceive information, eight paper-based prototypes similar to the InformationCasserole were conceived. Each prototype shows a combination of twelve to fifteen information objects from different genres (e.g. classified ads, news headlines, blog posts, news pictures, movie posters, youtube videos, secondhand goods, and city events). The object genre was emphasized and differentiated by aesthetic design. The visible object diversity encourages people to search for relations among visualized objects [6]. Therefore, the combinations, ranging from one to eight genres, were designed to encourage subjects to search for patterns and relations among the objects. Six twenty minutes think-aloud sessions were organized with subjects whom were asked to search for and identify different genres of objects in each of the eight combinations presented in random order.

As expected, the subjects perceptually clustered the objects primarily in accordance to their genre. However, they sometimes tended to search for inner-divisions in objects of the same genre (e.g. clustering movies according to their cinematic kind or news articles in familiar news categories), or to merge related genres as a single genre (e.g. news articles and blog posts, or movie posters and news pictures). In total, the subjects perceived the diversity of

objects (P) in four different levels, and to each level we accord a corresponding parameter value inversely proportional to the number of visible object genres: the first level is a single-genre diversity (P=3); the second level is a diversity of two to three genres (P=2); the third level refers to diversity of three to four genres (P=1); the fourth level describes a diversity of five to seven genres of objects (P=0). Figure 3 shows the number of visible genres of objects in each of the eight combinations as seen by the subjects, and the P value of each of the four identified diversity levels.

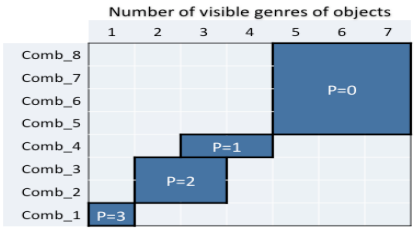


Figure 3. P as a design parameter

3. SITUATED INTERFACES AS IR SYSTEMS

In order to assess how R and P act as design parameters for the information settings of situated interfaces, the following five interfaces that act as retrieval systems in real-world environments were analyzed, and for each a corresponding design was developed and evaluated in settings that resemble or emulate its deployment environment.

The *Arts&Movies* is a situated interface intended for movie theatre lobbies to support the search and discovery of new interesting movies through an animated visualization that draws attention to relationships between movies and concepts. The *DigiJuke* is installed inside a bar to allow people to browse and select music songs on the touch-screen, and play their video clips accompanied by related images on the projection display. The *YouServe* prototype is collocated in a university library lobby to assist people in familiarizing themselves with the available library services, and finding a service relevant to specific needs. The *NewsWall* is a large display situated in the news production room of a broadcasting corporation. The prototype subtly visualizes the constantly evolving news information space on the web. The *MetroWindow* is designed for metro wagons and broadcasts summarized local news about cultural and civic events in the city of Barcelona.

In related works [7, 8] we have argued how R and P, as design parameters, can be quantified during requirement analysis and used alongside other aspects to conceptualize the design of information interfaces. For each situated interface, a couple of designers analyzed the characteristics of three entities being: the deployment environment, the humans present in it, and the adequate information space, which was defined based on an understanding of the needs and goals of the humans alongside the nature of the environment and the information and activity flows that it hosts. Based on this analysis, the designers qualified the values of R and P for each situated interface, and consequently described its information settings, being the quantity of information to visualize and its visible diversity. This qualification of R and P was defined in accordance with several non-disjoint or co-dependent situational aspects of human-information interaction such as:

- The amount of available user attention (e.g. MetroWindow disposes of little attention in contrast with DigiJuke).
- The duration of human interaction with information (e.g. NewsWall remains in contact for prolonged durations, while the interaction with YouServe is more momentarily).
- The convergence or divergence of the information seeking tasks (e.g. YouServe supports finding a specific library service, while Arts&Movies is designed to acquaint people with many movies).

Table 1. Values of R and P parameters for each interface

Situated interface	Recall	Precision
Arts&Movies	2	1
DigiJuke	3	3
YouServe	1	2
NewsWall	1	1
MetroWindow	0	3

The results of this R and P qualification are summarized in table 1. They show how R and P can characterize, from a perceptual stance, the role of a situated interface as an information retrieval engine, and parameterize the design of its information settings accordingly. For example, when the user objectives are to search for specific objects (e.g. *YouServe*), R is minimized, while P can be maximized when the search converges on specific genres (e.g. *MetroWindow*) or minimized when it diverges to cover many genres (e.g. *NewsWall*). A maximized R signals that the interaction tackles a large number of objects. In this case, when P is maximized (e.g. *DigiJuke*), it determines that this large number is a single collection of similar objects, or, when it is minimized (e.g. *Arts&Movies*), it signals that this large number of objects is a visually diversified information space.

The designers also developed the interfaces information architecture and aesthetic design, but these activities lies outside the scope of this paper. The final designs are shown in figure 4.

Figure 4. The situated interfaces final designs

4. USER CONTROL OVER R AND P

Based on the discerned ability of R and P to describe the information settings of situated interfaces and consequently their performance as information retrieval systems, we explored the possibility of allowing users to control them dynamically in classic search and retrieval scenarios. Therefore, we designed two experimental prototypes (figure 5) for querying a large information space of news articles, by which users can set and control the values of both R and P. The prototypes were evaluated to assess the feasibility of this approach and its utility.

The NewSearch prototype collocates two slide-bars adjacently to the query textbox for setting R and P explicitly, and returns an equivalent clustered visualization of news articles. Users control the number of clusters (discerned by color) by P and their average cardinality by R. The 3DQuery prototype uses a tag-map as a new concept for defining user queries, and shows a corresponding map of news articles. The tag-map is a rectangular box where users can place different tags of distinct sizes. The position of each tag determines that of the corresponding cluster of news articles, and the tag size the cluster cardinality.

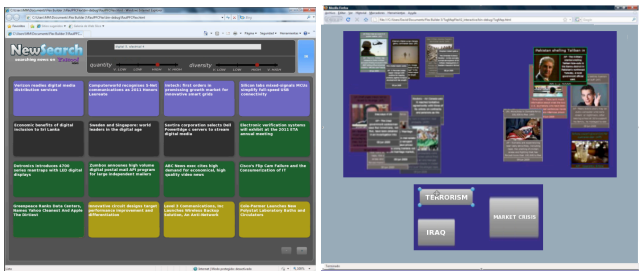


Figure 5. NewSearch (left) and 3DQuery (right) prototypes

Each prototype was evaluated by a different group of ten subjects in the lab. The subjects were asked to browse and read the collection of news articles for fifteen minutes, and then answer a set of open-ended questions concerning their utility and usability. The user evaluations of both prototypes showed that their learning curve is not negligible. Subjects were not naturally inclined to use the slide-bars of NewSearch to control the information settings. An explanation for this may well be that they are accustomed to a given query paradigm and the difficulty lies in making the paradigm change [9]. However, this issue requires further investigations. Subjects found it easy to use the tag-map paradigm in general, but it was deemed too complicated for simple queries and more useful for prolonged search and exploration since it allows users to dynamically adjust queries and therefore eliminates or reduces the need for re-querying.

The experience and knowledge gathered with the design and evaluation of these two prototypes would be used for developing future prototypes that intent to delegate more intuitively a dynamic control over the information settings of information retrieval interfaces to their users.

5. CONCLUSIONS

During the course of this paper we have explored ways to tightly integrate IR and HCI in a variety of human-information interaction scenarios where interfaces act as information retrieval systems. In particular, we studied how R and P as design parameters can describe the information settings of these interfaces. Both aspects were parameterized on a 0-3 scale on the basis of conducted experiments to analyze different possible information settings. Consequently, five situated interfaces were designed and analyzed to discern how R and P are qualified during requirement analysis, and how together they describe the information settings of situated interfaces, and therefore help reflect the interaction requirements onto the interface design.

Finally, we investigated the feasibility and utility of delegating control of R and P dynamically to users during classic search and retrieval scenarios, and concluded that while this approach is clearly advantageous for exploration tasks and tasks that require re-querying, a more profound study should be conducted for further analysis. Such endeavor will constitute the essence of our future work.

6. DISCUSSION

The approach that we presented in this paper demonstrates that a tighter integration of HCI and IR is possible, by exploring the potential of R and P as design parameters for the information settings of situated interfaces. The use of these two performance metrics as design parameters may be seen as controversial, however, it is justified given that efficiency and information relevance no longer constitute the backbone of user expectations in all cases of human-information interaction. Instead, new aspects of human-information interaction (e.g. emotional, cognitive, experiential, situational, and cultural) are affecting the manner by which we conceptualize information systems. Our approach does not comprehensively address all these aspects, and therefore can be complemented by introducing new parameters to reflect with a higher affinity the aspects of human-information interaction onto the system design.

7. ACKNOWLEDGEMENTS

The authors would like to thank Oriol Galimany and other members of the Interactive Technology Group at Universitat Pompeu Fabra for their support.

8. REFERENCES

[1] Vogel, D. and Balakrishnan, R. 2004. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. Proceedings of UIST '04, pp. 137- 146.

[2] NJ Belkin. Some (what) grand challenges for information retrieval. ACM SIGIR Forum, 2008

[3] R.A. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[4] Miller G. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review, 1956.

[5] L. Hallnäs and J. Redström. 2001. Slow Technology, Designing for Reflection. *Personal Ubiquitous Comput.* 5, 3 (January 2001), 201-212.

[6] Koffa, K. (1935): *Principles of Gestalt Psychology*. London, Routledge & Kegan Paul Ltd.

[7] Moghnieh, A., & Blat, J. (2009). A basic framework for integrating social and collaborative applications into learning environments. Proceedings of m-ICTE'09 Vol. 2 (pp. 1057-1061), 2009.

[8] Moghnieh, A., Sayago, S., Arroyo, E., Sopi, G., and Blat, J. Parameterized User-Centered Design for Interacting with Multimedia Repositories. In Proc. MMEDIA '09, IEEE.

[9] B. Buxton. 2007. Sketching User Experiences: Getting the Design Right and the Right Design. Morgan Kaufmann Publishers Inc. CA, USA.

[10] S. Bødker. 2006. When second wave HCI meets third wave challenges. In Proceedings of NordiCHI '06.

Towards User-Centered Retrieval Algorithms

Manuel J. Fonseca
Department of Computer Science and Engineering
INESC-ID/IST/Technical University of Lisbon
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
mjf@inesc-id.pt

ABSTRACT

Nowadays almost all retrieval algorithms (for text, images, drawings, etc.) are mainly concerned in achieving good system-centered measures, such as precision and recall. However, these systems are used by users, who try to achieve goals through the execution of tasks. To better satisfy the users' needs we must involve them in the development process of the retrieval systems.

In this paper, we argue that a user-centered approach, where users are included in the development cycle of the overall retrieval system, can lead to improved retrieval algorithms and also to a better user satisfaction while using the system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - Graphical user interfaces (GUI)

General Terms

Design, Human Factors

Keywords

User-Centered Design, User-centered approach, Retrieval algorithms

1. INTRODUCTION

The majority of the retrieval algorithms, whether they are for text, images, drawings, 3D objects, audio, video, etc., are mainly interested in performing well for system-centered measures, like for instance precision and recall. However, these systems are used by users who want to perform specific tasks and achieve specific goals. We can develop a good retrieval system, that performs well against a predefined ground truth, but when we delivery it to users they may not be able to find what they want or they may not even be able to submit a query to the system.

For illustration purposes let us consider the following hypothetic scenario: "We developed a system for retrieving generic complex vector drawings, like for instance technical drawings, architectural plants or clipart drawings. We evaluated it using query-by-example and a set of predefined drawings, achieving a good precision and recall measure. Afterwards, when we delivered the system to users, we noticed that they were not able to use it, because they could not find the (first) drawing that they must use as query to find the desired drawing. Moreover, users do not want to search for the complete drawing, but only by a subpart of the drawing."

This scenario could be avoided if before we developed the retrieval system we asked users what were their needs, what did they want to perform on the system and how they want to do it. To collect all this information we need to apply a user-centered approach where users are involved in the development of the retrieval system and algorithms.

In this paper we defend an user-centered approach as a way to create better retrieval algorithms and improve the overall retrieval system. We start by shortly describe the user-centered approach and the iterative cycle used in the user interface design. In Section 3 we describe our application of the user-centered approach in the development of retrieval algorithms. Finally, we present some conclusions.

2. USER-CENTERED DESIGN

The user-centered design (UCD) is a design methodology, where the needs, skills and limitations of the users are taken into account during all stages of the development of the system. The key premise of the user-centered design is that the active involvement of the users in the development process as well as in the evaluation of the interactive products can lead to well-designed systems that best meet the desired usability goals. These systems will take advantage of users skills, will be relevant to their work and activities, and will help them rather than constrain their actions.

One of the principles from the UCD [4] states that we first need to identify who the users will be (profile, skills limitations, etc.) and what tasks they perform and/or wish to perform. The second principle mentions that the systems should be exposed to users in the early stages of development to collect feedback from them. Finally, the third principle is iterative design. The results and feedback from user testing should be used to fix and improve the system. The UCD assumes an iterative cycle with identification of the users' needs, design of the solution and evaluation, repeated as

EuroHCIR '11 Newcastle, UK
.

often as necessary, as depicted in Figure 1.

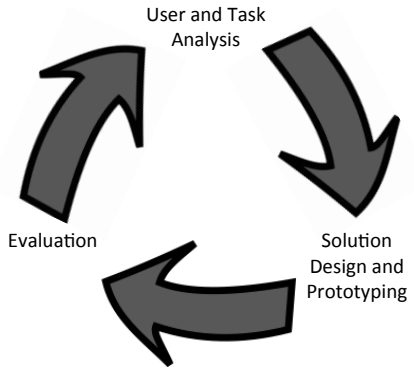


Figure 1: User-centered design iterative cycle.

3. USER-CENTERED RETRIEVAL

Typically when we want to develop a new retrieval approach, we look at the media to retrieve (text, audio, video, drawings, images, etc.), identify the features that better describe the media, create a matching algorithm and finally we compute precision and recall. Although this methodology allows us to create retrieval systems, we believe that by including the user in the development cycle will allow us to deliver better and more usable retrieval systems, that will allow users to achieve their goals and not only systems that have a good precision and recall performance.

Moreover, we should not develop retrieval systems, and that includes descriptor computation, matching algorithms and presentation of the results, without first identifying a set of user needs and functional requirements (first step in the user-centered design). We need to know our users, their skills, their background, their profile. We must identify their needs and requirements, their goals and how they achieve them. In summary, we need to do an user and task analysis before we start developing our retrieval system. User and task analysis should not only influence the design of the user interface, but also the design of the retrieval approach or algorithm.

For instance, users could use various strategies to perform a search in a drawing retrieval system. They could use a drawing that they already have, in a file, to search for similar drawings using query-by-example, or they could draw a sketch of the drawing that they want to find. As we can see, the retrieval solution (feature extraction, indexing and matching algorithms) will be different on each case. While in the first case we only need to compare two drawings of the same complexity and with the same characteristics (sets of lines and polygons), in the second case we need to compare complex drawings with sketches (typically simpler and with less elements). Thus, the way users perform the task to achieve their goal influence the retrieval approach that we should develop.

After developing the retrieval solution based on the user requirements, we should evaluate the retrieval system, using not only system-centered measures, but also user-centered measures, such as time to complete tasks, error rates, satisfaction, etc. As in the user-centered design of interactive systems, results from the evaluation of the retrieval system (system and user centered measures) should be used to improve the system and to refine the user and functional requirements of the retrieval system.

One of the things that we observed in one evaluation session with users, was that users did not care about where in the order of retrieval the intended drawing appears, the important fact being that it was there. One of the users produced this comment "It [the system] found it [the drawing]! That is what counts!" However, when we evaluate retrieval systems, the majority of the existing measures and ground truth datasets privilege precision. Of course this system-centered evaluation is important, but we should also take into account the users perspective, where they privilege recall.

3.1 An Example

Involving the users can affect the way we develop the retrieval algorithms. In recent years we developed a generic approach for complex vector drawing retrieval, based on the topology and geometry of the elements present in the drawing. These two features were used to describe the content of the drawings, and during matching, we first compare the drawings using topology and them we compare the geometry of those with similar topologies, giving the same weigh to both features (for more details see [1]). This generic retrieval approach was used to develop one system for retrieving technical drawings [3] and another for retrieving clipart drawings [2].

Before we developed this solution and the two retrieval systems, we performed user and task analysis to understand how users wanted to make queries to this type of systems. We notice that they prefer to draw sketches of the drawing that they were looking for than to submit an existing drawing to perform a query-by-example. Moreover, most of the times they do not have a drawing similar to the one that they are looking for.

The two systems were both evaluated with users, and from those evaluations we observed that the way users search for technical drawings was different from the way they search for clipart drawings [6]. While in the case of technical drawings users draw more complete sketches with several visual elements, and consequently defining a richer topological con-

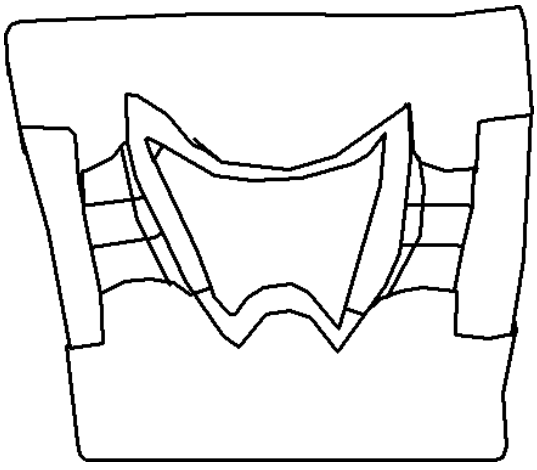


Figure 2: Sketch specifying a query to find a technical drawing.

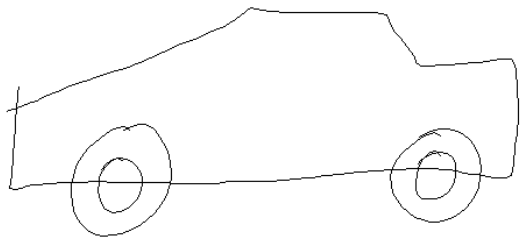


Figure 3: Sketch specifying a query to find a clipart drawing.

figuration, as illustrated in Figure 2; for clipart drawings, users produced simpler sketches, with fewer elements and with a poorer topological description (see Figure 3).

Due to this observation during tests with users, we refine our retrieval algorithm for retrieving clipart drawings [5], putting more emphasis on the geometry than on topology. With this change we were able to achieve a better precision and recall measure for clipart drawings, and we adapted our retrieval system to the users' way of sketching queries.

3.2 Discussion

We can not develop our retrieval algorithms without involving our users into the development cycle. As in the design of interactive systems, also in the development of retrieval systems we must involve the users.

They must be involved in the initial phase, so we can understand how they search for the information, what are their knowledge, what are their limitations and what is their profile. With this we are able to identify users needs and functional requirements.

Later on, during the development of the algorithms we should take into account this input and adapt the algorithms to provide "good results" for "our" users, and not for the users in general, or for the system.

Finally, during the evaluation stage, besides computing the traditional system-centered measures, for a set of datasets defined as ground truth, we should also involve users in the evaluation to collect quantitative and qualitative measures. Information gather during evaluation should be used to improve the retrieval algorithms and the overall retrieval system, in the next iteration of the iterative cycle of the user-centered approach.

4. CONCLUSIONS

In this paper we defended a user-centered approach for the development of retrieval systems. As in the case of user interfaces design, also for retrieval systems is important to know our users, adapt the algorithms to them, and involve the users in the evaluation of the system.

We believe, and we had confirmed, that the involvement of the user in the development cycle of retrieval systems can conduct to better systems that satisfy users needs and are more adapted to them.

5. ACKNOWLEDGMENTS

This work was supported by FCT through the PIDDAC Program funds (INESC-ID multiannual funding) and the Crush project, PTDC/EIA-EIA/108077/2008.

6. REFERENCES

[1] M. J. Fonseca. *Sketch-Based Retrieval in Large Sets of Drawings*. PhD thesis, Instituto Superior Técnico / Technical University of Lisbon, July 2004.

[2] M. J. Fonseca, B. Barroso, P. Ribeiro, and J. A. Jorge. Retrieving clipart images by content. In *Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR'04)*, volume 3115 of *Lecture Notes in Computer Science*, pages 500–507. Springer-Verlag, Dublin, Ireland, July 2004.

[3] M. J. Fonseca, A. Ferreira, and J. A. Jorge. Content-based retrieval of technical drawings. *International Journal of Computer Applications in Technology (IJCAT)*, 23(2–4):86–100, 2005.

[4] J. D. Gould and C. Lewis. Designing for usability: key principles and what designers think. *Commun. ACM*, 28(3):300–311, 1985.

[5] P. Sousa and M. J. Fonseca. Geometric matching for clip-art drawing retrieval. *Journal of Visual Communication and Image Representation (JVCI)*, 20(2):71–83, February 2009.

[6] P. Sousa and M. J. Fonseca. Sketch-based retrieval of drawings using spatial proximity. *Journal of Visual Languages and Computing (JVLC)*, 21(2):69–80, April 2010.

Design Thinking for Search User Interface Design

Arne Berger
Chemnitz University of Technology
Strasse der Nationen 62
09107 Chemnitz
Germany

arne.berger {at} informatik.tu-chemnitz.de

ABSTRACT
The paper describes with the help of a brief example how design methods, namely those formed in design thinking can help search user interface design to innovate throughout the software development process.

Categories and Subject Descriptors
H.5.2 [**Ergonomics, Evaluation/methodology**]: Design Methods in Search User Interface Design

General Terms
Measurement, Documentation, Performance, Design, Human Factors, Experimentation

Keywords
Design Thinking, User Interface Design, Design Methods, Qualitative Studies

1. INTRODUCTION
Since Tim Browns ingenious talk on TED [1.], Design Thinking (DT) had a huge impact on the business and design world. By injecting the way designers think into accustomed business processes, CEOs hoped to gain an advantage in competition. Designers on the other hand hoped their overall influence might increase. However, the field has more to offer than bringing creative techniques to supposedly uncreative domains. The first publications on the matter appeared as early as the late 1960s [2., 3., 4.] as a way to externalize the enigmatic design process. Since then, the creative application of design methods (DM) has proven its effectiveness, fun and relevance countless times. [5., 6.] Despite its persistent application in typical creative domains, the radical application of DM for digital age products is still a young discipline.

1.1 Design Thinking vs. Design Methods
The difference between DT coined and developed at Stanford [7.] and DM as defined by Jones amongst many others [3.] needs to be precised in another publication. For now, the author (a Designer) is grateful to see the broad spectrum of DM finally being brought to attention due to the success of DT. However, there are way more methods to use than the 51 methods as suggested by DT [8.] and there are way more feasible design processes than defined in DT. Because of the briefness of this paper and for the sake of a better understanding, DT is used as an expression for the design process, while DM is used as an expression for any design method from the DT or any other DM toolbox.

2. CURRENT STATE OF DESIGN METHODS IN SEARCH USER INTERFACE DESIGN
The possibilities of DM are still badly implemented into product development. However, a subset of DM, namely User Centered Design (UCD) is fairly well implemented in the domain of interface design, including that of search user interface design. UCD significantly helps evaluating user needs but often fails to innovate. UCD methods mainly consist of a relatively strict set of methods compared to what DT and DM have to offer [9.]. Those methods are capable of gaining insight and evaluating interfaces but do not encourage an innovation process for future user interfaces.

As an user interface design professional working in an academic development environment that is mainly formed by information retrieval experts, the following description of a typical workflow abstracts the prototypical UCD process of developing search user interfaces.

2.1 Current Process of Search User Interface Design
1. Users tasks and problems are observed via Site Visits or Website Analytics [10.]. Those methods help to gain insight into specific user problems. The combination of both nowadays is the holy grail of gaining insight into users issues [10.].

2. Information retrieval experts and search user interface designers use methods like brainstorming to plan a software product. It is used mainly as a conversation starter, but also functions as a way to frame the current state of technical possibilities.

3. Users problems (step 1.) are interpreted and tried to be solved with the help of the technical possibilities (step 2.) which are then implemented.

4. The usability of the search user interface proposed in 3. is evaluated via user studies comparable to the ones in step 1.

Iterations: The abovementioned steps are iteratively repeated several times. With the help of prototypes the interface is refined before a final implementation takes place. However these steps only help to streamline the interface. They are not fully useful for innovating an interface according to DTs possibilities.

2.2 Critics of the Current Process

We believe that the process of nailing down the problem and suggesting a vital solution after framing technical possibilities and observing users is insufficient. Those well established methods have the main advantage of providing hard numerical measures. Which is even more so, when measures like precision and recall are used to learn how efficient a system is. Via those standardized measurements a comparison between different solutions is easy to draw. Relying on those hard measures only shows insights, which can be formulated in numbers and concluded from those.

On the other hand, soft properties of a search user interface like »what user really want«, »fun of use«, »suitability to unusual tasks« and in parts »user satisfaction« are next to impossible to measure via hard numbers. Although efforts exist [11.] measurability of qualitative soft properties is hard to be standardized. Outcomes therefore are less clear cut and often fail to be comparable via statistics. As the academic viewpoint in the field tends to analytic comparison, soft properties are seldom explored, described and measured. Therefore subsequent findings often fail to be implemented.

Based on the before mentioned, we propose the radical application of DT in search user interface design via »participatory prototypes«. This concept integrates users and developers alike. We demonstrate its process briefly in the next chapter and explain its application in three following examples.

3. PROPOSED DESIGN THINKING PROCESS FOR SEARCH USER INTERFACES

In the business world (see introduction) DT is foremost a process used for innovating new products.

The DT process is defined as following [8.]

Understand: Understand problem and context.

Observe: Externalize future users problems via e.g. extreme user interviews or empathy maps.

Define: Interpreting and weighting the gained knowledge from the previous steps via e.g. ad-hoc personas.

Ideate: Using common or uncommon creative techniques, e.g. body storming for generating many ideas.

Prototype: Visualize and communicate ideas with the help of fast and cheap prototypes with paper, Lego bricks or the product box method.

Test: Future users test those prototypes, via e.g. story telling techniques.

We believe that DT can and should be incorporated in any possible stage of a development cycle. Interface design prototypes are extraordinary easy to manufacture and cost next to nothing.

We suggest to apply the DT process more closely to the development of search user interfaces to benefit from its many advantages, esp. to force the pace of innovation.

3.1 Prototype Categories

As the label »prototype« may be misleading, we tend to think of anything capable of producing feedback as a prototype. To make further understanding easier we classify prototypes as following in the order of their advancement:

3.1.1 Very Low-Fi Prototype (Conceptual Model)

Generated by: user

Function: none, may not be technically feasible

Workflow: only conceptual

Visual Design: none

Medium: analog

Modality: any

Usually user generated, often not understandable without the creators explanations. It only describes a preliminary workflow of operations and functions and is not necessarily technically feasible.

3.1.2 Low-Fi Prototype (e.g. Paper Prototype)

Generated by: user, designer

Function: none, may not be technically feasible

Workflow: preliminary, mimicking operations

Visual Design: none

Medium: analog

Modality: any

Usually presented via the Wizard-Of-Oz technique it incorporates as many operations as possible and always fakes function.

3.1.3 Mock-Up

Generated by: designer

Function: none, may not be technically feasible

Workflow: mimicking operations closely

Visual Design: none

Medium: digital

Modality: any

Is often (and should be) visually unapealing, mimicking operations closely, but fakes function.

3.1.4 Dummy (often refered to as Click Dummy)

Generated by: designer

Function: none, may not be technically feasible

Workflow: mimicking operations

Visual Design: existing, often visually polished

Medium: digital

Modality: any

Incorporates a polished visual design, mimicking operations, but fakes function. May or may not incorporate the proposed interaction paradigm. The most common implementation of the later is a browser based click dummy that fakes the functions off a mobile touchscreen device.

3.1.5 High-Fi Prototype

Generated by: designer, developer

Function: incorporates some or most of the proposed functions

Workflow: mimicking operations

Visual Design: existing, often visually polished

Medium: digital

Modality: same as end product

Is similiar to a Dummy but also incorporates some of the proposed functions. It also incorporates the proposed interaction paradigm.

3.1.6 Alpha Grade Version

Generated by: developer

Function: incorporates some or most of the proposed functions

Workflow: mostly operational

Visual Design: may or not be existing

Medium: digital

Modality: any

A prototype proposed by developers that demonstrates most basic functions, usually does not feature a polished design.

3.1.7 Beta Version

Generated by: developer

Function: incorporates some or most of the proposed functions

Workflow: fully operational

Visual Design: existing

Medium: digital

Modality: same as end product

A visually polished prototype most often proposed by developers is a functioning program that may have bugs or quirks and is mainly used in order to get rid of those.

3.2 Observations for Prototypes

As this brief listing suggests most of the prototyping work in search user interface design is done by a designer. Thus helping to maintain a conversation between what users want and what developers can implement.

There are usually no direct prototypes from the users. Users comments or observations are interpreted multiple times. First they are made operable via prototypes, crafted by designers, which subsequently are interpreted by the developers.

Prototypes from the perspective of a developer are used only for evaluation during the end of the implementation cycle. As a lot of code and effort went into these, heavy changes are omitted and hopefully eliminated with earlier prototypes.

While the main goal of DT is to encourage interdisciplinary user groups to create innovative prototypes, it does not focus on direct prototypes from users or developers.

3.3 Implications for Process

We want to continously implement user prototypes into the development and we also encourage a process where developers explain technical feasibility via prototypes even in very draft and early stages.

This realization came through practical usage of various DM in a couple of projects. The following chapter briefly describes how we introduced participatory prototypes to search user interface design for the creation of playlists for mobile video consumption.

Two other successful projects include Design Thinking for a customized faceted navigation and Design Thinking for a multitouch interface for searching in large multimedial repositories.

4. DESIGN THINKING THE CREATION OF PLAYLISTS FOR MOBILE VIDEO CONSUMPTION

We wanted to address a problem, know to many smartphone users on the move. We understand that, weather commuting or going out with friends users usually avoid constructing complex search queries to find suitable content to watch.

To define the problem, we asked users what they miss and want from a mobile TV application. Two main points emerged:

With services like youtube consumers are left having to refine a search query several times or to use non-customized item lists such as »most viewed«. On the other hand, in traditional TV a moderator weaves a golden thread and guides viewers via this potentially emotional connection through a series of video clips. After an ideate session the most promising prototype was a mixed breed of playlists, woven together by emotional metadata. To gain insight into users mindsets regarding the construction of those personalized playlists we applied various DM.

To find out which emotional content attributes users are looking for, we asked participants to map out a virtual space of content properties and show how they thought to navigate within it. This method usually helps to discover pathways and interests in which people make sense of a particular content space. The results eventually help to make sense of how to construct queries for filter specification.

Users were asked to individually draw a map or diagram of what comes to their mind when being on the move and having a mobile video handset available, whether sitting on public transportation alone or being in a pub with friends. The six users had 15 minutes time to draw a map or scheme and were asked to freely associate parameters to form a personalized playlist. Given the mindset of being on the move, users formed questions from a simple vocabulary and subsequently wanted to change only certain parameters after watching a few video items. A discussion with all participants followed.

The results lead to the assumption that users are interested in direct mood filters. Most of the user generated maps feature mood clusters or the simple question »how« in a list of questions.

Based on those findings the developers of the future interface with the help of a designer proposed a low fidelity prototype containing a filter named »How« together with more filters based on the four cardinal questions Who, Where, When, What. This was done because all those metadata fields could be filled with metadata readily available in the existing database. To prove the concept it was introduced to twelve users. Users' feedback on this approach was insightful in two ways. On one hand, users at large expressed their general approval on the advantages that might arise by constructing exhaustive content filters with just a few steps of interaction. On the other hand, the pre-structured characteristic was heavily criticized. However, the rigidly defined prototype inspired participants to incredibly rich feedback. This proposal in combination with open ended questions has proved to be a fast

and convenient way to gain user feedback on a large variety of issues without a lot of explanation. The main insight is, that all users found and used the filter option »how«. Most user feedback was given on only this feature. Findings are discussed in depth in [12].

TV Anytime [13.] is a metadata standard that defines metadata for broadcasts. It is common to use in describing video items and also features 53 moods. For the sake of technical interoperability we wanted to stay within the realm of this particular metadata standard but also wanted to make the proposed moods more accessible for users. Based on those technical restrictions and the previous results we individually asked 45 potential users to sort the moods into self-defined categories that made sense to them.

At least two completely different ways of sorting prevailed. One group of users preferred an order that resembles a classification into movie genres, while a second group was interested to sort them according to emotional dependencies. While a number of 45 users was significant enough to reveal two groups, users assigned to the first group were too few to manifest significance. Focusing on the larger group (35 participants) seven mood categories were filled unanimously. Apart from very few moods all other moods are mutually joint to groups. This could make the previous discussed low fidelity prototype more flexible in navigating complete mood sets. Based on those findings, users proposed an interface that asks questions in an order that is more determined by them. A subsequent High-Fi prototype was built, incorporated 1000 video items. It allows the selection of a variety of moods as well as a combination of filters derived from the five cardinal questions. A formal user study is now underway.

5. Acknowledgements

This publication was prepared as a part of the research initiative sachsMedia (http://sachsmedia.tv), which is funded by the German Federal Ministry of Education and Research under the grant reference number 03IP608. The authors take sole responsibility for the contents of this publication.

References

[1] http://www.ted.com/talks/tim_brown_urges_designers_to_think_big.html (accessed Apr 29, 2011)

[2] Archer. Design as a discipline. Design Studies (1979) vol. 1 (1) pp. 17-20

[3] Jones. Design Methods. John Wiley and Sons (1992)

[4] Newell et al. The processes of creative thinking. (1959)

[5] Lawson. How designers think: the design process demystified. (2006) (Elsevier)

[6] Schön. The reflective practitioner: how professionals think in action. Basic Books (1983)

[7] Kelley. The Art of Innovation: Lessons in Creativity from IDEO. Crown Business (2001)

[8] Plattner. Design Thinking. Mi Wirtschaftsbuch (2009)

[9] Cooper. About Face 3. Wiley and Sons (2007)

[10] Hearst. Search User Interfaces. Cambridge University Press (2009)

[11] Hassenzahl et. al. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Proceedings of Mensch & Computer (2003)

[12] Knauf, Berger, et. al. Constraints and simplification for a better mobile video annotation and content customization process. In Workshop Proceedings of the EuroITV. (2010)

[13] TV-Anytime Phase 1: Metadata schemas http://www.etsi.org/deliver/etsi_ts/102800_102899/1028220301/01.02.01_60/ts_1028220301v010201p.pdf (accessed Oct 10, 2010)

The Development and Application of an Evaluation Methodology for Person Search Engines

Roland Brenneke
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
roland.brenneke@gmx.de

Thomas Mandl
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
mandl@uni-hildesheim.de

Christa Womser-Hacker
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
womser@uni-hildesheim.de

ABSTRACT

This paper presents a user oriented evaluation methodology for comparing person search services on the Web. Many established system oriented methods from information retrieval cannot be applied to this domain. Our user oriented methodology is applied to a test comparing the person search engines yasni, pipl.com and 123people. The user study with over 30 participants led to relevant results. The coverage of data object types within the person search engine results is quite different. Especially the amount of pictures and social media network entries which are presented by the systems and which are perceived by the test users differ greatly. The results also revealed a tendency to judge people more positively when more information was found.

1. INTRODUCTION

Person search engines are important specialized search services on the Web. These systems consult other services for information about a person and integrate it in one interface. They can be regarded as meta search services or one point stops for personal information. Mostly, they are tailored for normal people and not for celebrities and other famous people. As such, it is different from named entity search in general.

Especially in the Web 2.0 and its ease of publishing content on the Web, many people deposit much information about them or content they created in various sites. Users need to have the proper information competence to foresee the consequences of such behavior. Often, users are advised not to publish too much information. Online reputation management becomes an important issue. On the side of the users, social networks and person search services lead to information ethical considerations about the use of personal information.

Searching on information about others is a very frequent information need and a reason for using a search service. According to Google Trends, the most popular person search services receive over 200,000 hits per day. However, 90% of the users do not rely on person search engines but they use general Web search or go directly to social networks to find out about people. Nevertheless, 10% is still a significant share and hit rates for person search engines are constantly high. In addition, many of these searches may have a high impact. Many recruiters use person search engines for checking on candidates.

A questionnaire study among 548 enterprises was published in 2010 [5]. This Social Media HR Report 2010, revealed that in 2009 over 59% of the companies have used the internet to check on applicants. Almost 10% had already turned down an application because of information on the Web. Companies who do not use the Web for checking on applicants` state that lack of time and ethical questions are the main reasons not to do so [5].

An international study showed that this behaviour is more widespread in the US than in European countries [3]. Interviews with decision makers in German companies revealed that they are well aware of the potential of retrieving applicant information [11].

The use of person search engines for job applicants is only one potential usage scenario; however, it is a very prominent one. Other than that, there are many reasons for why a user would want to search for a person. And despite the use of a named entity in the search, the information need is rather vague and can be rephrased with “Find out something about person X”.

The success of a person search engine depends on many factors. Person search engines are meta services which extract results from a large variety of different online media. The presentation of these results in the user interface is an essential factor for the success of the search service. If a result is far down on the result page and the user never scrolls there, potentially relevant items cannot be found. That means that the search capability is only one success factor for person search engines. Consequently, our experiment was designed as a user test. We intended to evaluate the user experience and the success with the tool person search engine and neither specific system components nor absolute retrieval performance.

2. RELATED WORK

The evaluation of retrieval systems is central in information retrieval research because the system performance cannot be predicted. The most influential retrieval evaluation methodology is called the *Cranfield* paradigm. Information retrieval research has adopted an evaluation scheme which tries to ignore subjective differences between users in order to be able to compare systems and algorithms. The user is replaced by a prototypical and constant user. Relevance judgments are provided by domain experts [8, 10].

EuroHCIR 2011. The 1st European Workshop on Human-Computer Interaction and Information Retrieval. July 4th 2011. Newcastle, UK

Cranfield evaluations have often been criticised for several reasons. The main objections come from advocates of user oriented studies. The search situation of users depends on many individual and contextual factors which can only be captured in user experiments [6]. The real user experience and the success in a real world situation cannot be measured with the laboratory style experiments based on the *Cranfield* paradigm [12].

Person search engines have a higher chance to succeed than general purpose search services. The retrieval with named entities is known to be easier than searches without names entities [9]. The selection of a person search engine hints the type of result. Consequently, synonymy between names and words are a smaller problem than in general purpose search engines. Synonymy between names, on the other hand, is a big challenge for person search engines.

3. METHODOLOGY

The balance between control and realism is a challenge for each experiment. For the presented study, we chose a user experiment to test person search engines because an approach purely dedicated to retrieval power does not mirror the user experience for person search engines well. It is necessary to limit the realism in a user experiment in order to allow comparison across participants in the test. We selected a job applicant scenario in order to make the experiment interesting for the users. Applicant search is a very prominent usage type. The method was successful in making the experiment attractive. The test users liked the experiment very much and through word of mouth, more applicants wanted to register for the experiment than were needed.

The selection of persons for the task defines the content for the test. It seemed necessary to identify people for whom much information can be found on the Web. If there were no videos, working results like presentations or social network entries, then the performance of the person search engine could not be tested with our experiment. So even if the persons selected are not representative in terms of amount of online information for the whole population or all persons who are indexed in a person search service it increases the validity of the test to select persons with a large amount of online information.

Three people were carefully selected who had similar qualifications. For them, a job profile was developed which was given to the participants together with the names of the people. The users were asked to search for these people who would be interviewed for the position and check if they were appropriate. The job description and the name of each applicant were given to the test persons. Each of the candidates was well qualified for the job but had one negative aspect in his online data. One was an advocate of nuclear power and the job was for offered by an alternative energy company. The second applicant was a serial entrepreneur who portrayed himself on Facebook in pictures with attractive women and sports cars. The third applicant had party photos online where he could be seen smoking cigarettes and he considered himself as lazy in one social network while he had a very business oriented self image in another social network.

Obviously, such a scenario has some limitations. Person search engines need to disambiguate between people with the same name. We decided to choose people who are not ambiguous in order to have the same difficulty for each person. Such issues are evaluated in the system oriented campaign WEPS [1].

We selected people who had posted a large amount of information about themselves in the network. Again, this was done to obtain similar and comparable difficulty for the three test cases. Three person search engines were selected for the comparative test. We chose yasni, pipl.com and 123people because they were very popular at the time of the study according to Google trends. All three companies claim that they exploit only information available on the public Web.

4. STUDY

Students of the University of Hildesheim were recruited through a mailing list of students. Participation was voluntarily and no gratification was given. None of the participants had a computer science background. They all were frequent Internet users and had searched for people before but only 10% had used a person search engine before. The others use Google or social networks to find information on people.

The issue of relevance is always a crucial one in information retrieval evaluation. In our study, any item could contribute to the full picture of the applicant. Despite the clearly defined scenario, it remains vague which information is needed and what type of information is useful. It is difficult to assign relevance to items or even weights to categories. The user interfaces of the person search engines present the items in categories like e.g. social network entries or videos.

A questionnaire study [7] showed that users search mainly for the following items in the order presented when retrieving information about a specific person:

- Contact information
- Profile on a social network
- Photo
- Information about professional accomplishments or interests

The most frequently researched item, contact information does not apply for our scenario because the persons had sent a letter of application. The next two most frequent items are included. The fourth item is rather vague as some of the other items following as far as the categories of person search engines are concerned. As a consequence, the data available does not justify the assignment of weights to some items. In our study, all clicks on items were scored equally. The results will also show which of the items were most popular. The time per applicant was limited to 10 minutes. The entire experiment took 45 minutes on average including the pre- and post questionnaire.

One search service modified the interface after the first two tests. So it was necessary to eliminate three test sessions from the results and recruit further test users. This shows that not only the dynamics of the personal data presents a challenge for the test but also the ongoing modifications of the search engine. Overall, 34 took part in the experiment. Due to the problems of a relaunch of one service, we could consider the experiments of 10 users of 123people, 11 users of Pipl and 10 user of Yasni.

Each test person worked with one search engines on all three applicants. This between groups approach was applied was mainly applied to avoid a long learning phase for each of the person search engines. All tests were recorded with appropriate software.

Figure 1: Popularity of person search engines according to Google Trends

5. RESULTS

The result description focuses on the information perceived by users and the performance of the test users in the application task.

The information items clicked by the users were categorized. It can be seen that the services lead to a similar number of clicks when summed up over all users. Each of the services resulted in between 110 to 120 clicks for the ten test persons. In the case of Pipl, 11 test persons were considered. Each engine leads to a sufficient number of entries and has abundant information on the applicants in our scenario. This was a goal of the test design and was accomplished.

The type of information which was encountered was quite different. It can be easily seen, that 123.people facilitates access to photos whereas Pipl leads more users to social network entries. A comparative analysis for the services for the most popular item types is shown in Table 1.

In the post test questionnaire, users were asked about their subjective impression of the service they had used. In the overall satisfaction, 123people was rated highest. For the page structure, pipl received the best grades and the coverage of different business networks yasni was rated as most successful. In the latter case, the finding from the objective click data was confirmed. Further details on the results are provided in [2].

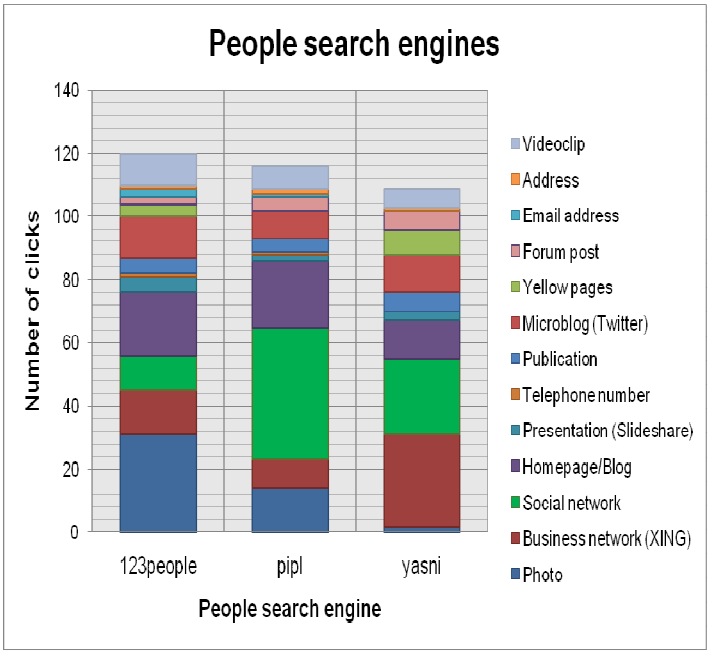


Figure 2: Clicks on items in the three person search engines

Table 1: Comparison of data types encountered

Item	123people	Pipl	Yasni
Photo	+ +	+ –	– –
Business network	–	–	+ +
Social network	–	+ +	+
Homepage/Blog	+	+	+ –
Microblog	+	+ –	+
Yellow pages	+ –	– –	+
Forum post	–	+ –	+
Videoclip	+	+ –	+ –
Publication	Because of a very low number of clicks is no rating possible.		
Presentation			
Email address			
Address			
Phone number			

Perception	
+ +	Excellent
+	Good
+ –	Moderate
–	Poor
– –	Unperceived

For two services, applicant 1 was selected by the majority of the test users. These two services had identified most items for this applicant. For yasni, applicant 2 was chosen as the best applicant despite the fact that the other two services found on average 10 items more for this person. Applicant 3 was given the last place for all three person search services. For each service, he is the applicant with the fewest items. There might be a trend to rate people higher when more information is available online.

6. RESUME

We presented a holistic evaluation methodology for person search engines. The performance of these search services is measured by observing the perception of test users. The test methodology is built on a realistic scenario and use case but it does not cover all the relevant quality aspects of person search engines. The important capability to resolve the ambiguity of names was not dealt with. In future work, it might be promising to develop a performance based test for this task only.

The complete information seeking behaviour and its success is also not measured with our test. In a realistic scenario, people might access the social media networks through a person search engine and continue their search mainly there. This issue could be resolved by observing real behaviour.

In the test, the search engine 123people was the winner. It not only led users to the highest number of items, but it was also subjectively judged to be the best person search engine. However, in several aspects other systems performed better and were judged better. The evaluation showed that the different tools are all based on the freely available data on the Web but that they lead to different results. The most sought items in our test were photos, entries and profiles in social and business networks and personal homepages. Each of the engines exhibited a strength in one of these items, e.g. 123people for photos because they are shown as top results. This is also confirmed by the questionnaire study among American recruiters [7].

For the users who publish information about themselves and who become information providers by doing that the issue of information competence will become more and more important. Personal Online Identity Management is a growing field and several new companies are entering the market.

7. REFERENCES

[1] Artiles, J.; Borthwick, A.; Gonzalo, J.; Sekine, S.; Amigó, E. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In: *CLEF Working Notes* http://nlp.uned.es/weps/weps-3/papers

[2] Brenneke, R. 2010. *Evaluation von Personen-suchmaschinen und Umgang mit persönlichen Daten im Internet*. Master Thesis, University of Hildesheim, Germany. International Information Management.

[3] CrossTab Marketing Services. 2010. *Europäischer Datenschutztag: Studie zur Online Reputation* Trustworthy Computing Group, Microsoft (Hrsg.). http://www.microsoft.com/germany/sicherheit/datenschutzstudie.mspx

[4] Hellmann, R.; Griesbaum, J.; Mandl, T. 2010. Quality in Blogs: How to find the best User Generated Content. In: *13th Intl Conf on Business Information Systems* (BIS 2010) Berlin, 3.-5. May. Berlin et al.: Springer [LNBIP 47] pp. 47-58.

[5] Zur Jacobsmühlen, T. (2010): *Social Media HR Report* 2010 Stepstone.de & HRM.de (eds.). http://www.jacobsmuehlen.de/studie/

[6] Lamm, K.; Greve, W.; Mandl, T.; Womser-Hacker, C. 2010. The Influence of Expectation and System Performance on User Satisfaction with Retrieval Systems. In: Proc *EVIA 2010: The First Intl Workshop on Evaluating Information Access* June 2010 National Institute of Informatics (NII) Tokyo, Japan, June 15-18, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/09-EVIA2010-LammK.pdf

[7] Madden, M.; Smith, A. 2010. *Reputation Management and Social Media: How people monitor their identity and search for others online*. PEW Internet & American Life Project. http://pewinternet.org/Reports/2010/Reputation-Management.aspx

[8] Mandl, T. 2008. Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. In*: Informatica – An Intl. Journal of Computing and Informatics* vol. 32. pp. 27-38.

[9] Mandl, T.; Womser-Hacker, C. 2005. The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: *Proc 2005 ACM SAC Symposium on Applied Computing* (SAC). Santa Fe, New Mexico, USA. March 13.-17. 2005. pp. 1059-1064.

[10] Robertson, S. 2008. On the history of evaluation in IR. In: *Journal of Information Science* 34(4). pp. 439-456

[11] Schäuble, T.; Griesbaum, J.; Mandl, T. 2009. Mehr-wertpotenziale von Online-Social-Business-Netzwerken für die Personalbeschaffung von Fach- und Führungskräften. In: *Informatik 2009 - Beiträge 39. Jahrestagung der Gesellschaft für Informatik e.V.* (GI) Lübeck [LNI P-154] pp. 2166 – 2180.

[12] Tawileh, W.; Mandl, T.; Griesbaum, J. 2010. Evaluation of five web search engines in Arabic language. In: *LWA– Lernen - Wissensentdeckung – Adaptivität*: Proc Work-shopwoche GI, Universität Kassel. Workshop Information Retrieval. http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf