EuroHCIR2012

24th-25th August 2012 – Nijmegen, The Netherlands

Proceedings of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval

A workshop at IIiX2012

Executive Summary

EuroHCIR2012 was the second workshop in the European series focusing on the combined aspects of Human-Computer Interaction and Information Retrieval (HCIR). The MUMIA WG3 supported event saw significant growth in interest from the first year, attracting over 30 submissions. 9 key research and position papers were accepted for Oral Presentation, while a further 13 demos and posters were accepted. All are included in these proceedings so that they can be accessed by those unable to attend the event in Nijmegen.

Organised by

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

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

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

James Kalbach
USEEDSº, Germany
jim.kalbach@gmail.com

MUMIA

Supported as a MUMIA WG3 event

Oral Presentations

Poster Papers

Using card sorts to understand how users think of personal information

Paul Thomas
CSIRO, Canberra
paul.thomas@csiro.au

David Elsweiler
University of Regensburg
david@elsweiler.co.uk

ABSTRACT

Understanding how users think of personal information, and how they mentally categorise or classify the objects they work with, should inform the design of personal information management (PIM) or personal retrieval systems. However, most investigations of this topic predate widespread multimedia, websites, and social media—objects that a contemporary PIM or retrieval system should work with.

We describe a pilot study that has used a variant of card sorts to elicit categories for personal information such as files, email, tweets, and websites. Our early results suggest that there are common categorisations which are not yet supported by PIM software, but which might reward further work. Our results also suggest that—with some caveats—card sorts are useful for understanding users' categories.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces
General Terms: Human Factors
Keywords: Facets, classification, card sorts

1. INTRODUCTION

Tools for personal information management (PIM) and search support the archival, retrieval, and management of "personal" data: the files, email, photos, videos, and other digital objects a person creates or uses [15]. Several studies show that PIM can be challenging [5, 11, 15] and it has been suggested that tools could be easier to use and more useful, if the way they represent objects matches the way users think of them [10, 20].

We are interested in how users think of the wide range of digital objects they interact with—the objects conventionally considered by PIM tools, objects less commonly considered such as websites and applications, and newer objects such as messages from social media. There are three linked questions:

1. What properties do users think "personal" digital objects have? That is, in which ways do users think of the objects they use?
2. Can we expose these properties in a PIM or search tool? Can the properties of an object be determined algorithmically? How should the properties be presented?
3. Assuming we can expose some or all of these properties, would we expect that to make management or retrieval easier?

At present, we are considering the first question. In particular, in this work, we have experimented with card sorts to elicit users' own descriptions of personal information.

2. CLASSIFICATION AND TOOLS

Past work has investigated the properties users assign to files, and elicited categorisation schemes. This work has not however considered as wide a range of object types as we do here; we may expect that with different types, sources, and quantities we would see different categorisations. Existing PIM and file management systems also support, or impose, particular faceting schemes.

2.1 Classifications

There is a rich tradition, in information science and information behaviour, of studies that try to understand how people organise, classify, and think about their information—that is, how people understand their information independent of any particular software capabilities or restrictions.

Three studies of note are by Cole, Kwasnik and Case. Cole [7] studied how 30 office workers classified their document collections. Six aspects of documents were important in filing decisions: "type", "form", "volume", "complexity", "functions", and "levels of information". Similarly, Kwaśnik [19] examined the categorisation behaviour of eight researchers and identified seven dimensions: "situation", "document", "disposition", "order/scheme", "time", "value", and "cognitive state". Case [6] investigated the behaviour of twenty historians and identified three main factors by which objects were classified in offices; "ease of access", "form" and "topic". While there is considerable overlap in the findings of these studies, particularly the criteria "form" and "topic", the studies all predate the rich digital landscape we have today, and focus on physical information objects.

Other research relating to our work has tried to learn about how people think about digital information by investigating how they behave with information in practice. For example, people have been shown to organise email messages and files based on projects [17, 23] and prefer to refind objects by location than using search facilities [2, 3]. These kinds of studies provide strong hints at how people may think about

digital information, but are influenced by the tools they have available to them.

More recently, Gonçalves and Jorge [13] asked participants to tell stories about three of their personal documents by describing each, from memory, in terms of its features, its content and the context in which it was created or used. It was discovered that time, location, and purpose of the document were the most common attributes used in stories. Similarly, Blanc-Brude and Scapin [4] used semi-structured interviews to examine participants' recollection of their documents. They found that location, format, time, keywords and associated events were remembered most frequently, but many of these attributes, particularly keywords, time and location were often only partially remembered or the recollections offered by the participants were incorrect. Both of these studies add a rich understanding of how people perceive their documents by examining a small number of documents in great detail, but do not explore how documents are related.

Our aim here is to add to and complement this previous work by using a technique that can deal with rich variety of information objects we interact with today; be tool agnostic; and allow insight into how different documents can be associated in different ways. We would also like to understand whether this has any impact on the design of PIM tools.

2.2 Tools

Tools for desktop search typically expose not just filesystem attributes such as name, size, and timestamp, but also extracted metadata. For example, Phlat [8] uses title, date, author, recipient, media type and tags; Haystack [1] has extensible facets but the authors have discussed media type, people named in email, text in a document, and URL.

More elaborate PIM tools have exposed other attributes to support different interactions. Some, such as Lifestreams [12], have supported time-based browsing and searching; an interface to Stuff I've Seen [21] extended this by indexing documents according to contemporaneous events. Other tools have taken a more personal view of time, or document lifecycle, and supported information management by context. Here, objects are organised according to tags for the context in which they are used [16, 18] or grouped according to patterns of use [9].

These systems offer variety of projections, across a number of media and storage types, but it is hard to know whether these match the way people naturally think of their objects. Alternative presentations that were natural for users, and easy to implement, would be worth further thought.

3. METHOD

In this work, we have experimented with repeated single-criterion card sorts to elicit users' mental categorisations. Cards represented digital objects on each participant's computer.

3.1 Card sorts

Repeated single-criterion card sorts—or just "card sorts"—are a common technique for eliciting users' categorisations (see e.g. Rugg and McGeorge [22] for an overview). Compared with interview-based techniques, card sorts are less flexible but are very lightweight: in our experience participants grasped the idea very quickly, many found it enjoyable, and the entire protocol took little time. Coding card sorts for later analysis is also relatively straightforward.

In a typical exercise, each participant is given a number of cards, each representing an object or a concept. They are asked to partition these cards according to any criteria they like; the criteria used for the sort, the categories (piles or sets), and the cards in each category are recorded. This is repeated several times, with participants suggesting a different criteria each time. For example, given cards labelled as follows:

1. pig; 2. chicken; 3. snake; 4. horse; 5. spider

a participant may sort cards according to the criterion "raised on a farm", with cards 1, 2, and 4 in category "yes" and cards 3 and 5 in category "no". A second sort, according to the criterion "where eaten", might have cards 1 and 2 in category "almost everywhere", card 3 in category "Asia", card 4 in category "Asia and Europe" and card 5 in category "don't know".

Records of the sorts may then be analysed with qualitative or quantitative methods.

3.2 Our approach

For this early experiment, we recruited a convenience sample of ten participants from two institutions. All were heavy computer users.

As preparation for the experiment participants were asked to select several information objects they had seen, used or created in the recent past. An "information object" was defined by giving as examples computer files, emails, websites, tweets or Facebook updates, documents or articles read, photographs or images, videos, music, and computer applications. However, participants were not restricted to these objects and could choose anything they wanted using these as a guideline. We encouraged participants to label 10 to 15 cards, which we believe balances the need for broad coverage with practical limits on participants' time.

While choosing objects, participants were asked to create index cards with the name of each (or some other reminder of its identity or contents).

With cards made, each participant was introduced to sorts using a set of cards with pictures of buildings; they were taken through some example sorts which included criteria clear from the pictures themselves (colour, material), criteria which were not immediately clear (insulation), criteria which were subjective (good place for a party). "Can't tell" or "don't know" categories were included in these examples.

The participants' own cards were then used for repeated sorts. Participants were asked to make piles according to a criterion of their choosing, and we noted the criterion (sometimes this was implicit), the categories used, and the cards in each category.

After collecting individual classifications from all participants, the full dataset, i.e. the criteria used to associate information objects, was analysed qualitatively using an affinity diagramming technique. This is a group-based process, which allows the discovery and validation of patterns in the data [14]. The researchers, as a team, looked for patterns in the data and grouped related criteria; we then related the formed groups in a way that creates a hierarchical coding scheme.

4. RESULTS OF THE SORTS

The results from this pilot are promising. Card sorts elicited a variety of criteria; some of these differ from those seen before, and many are not well supported by PIM or retrieval tools.

Group	Participants
document lifecycle	3
events	1
object's form	7
object's affective qualities	2
object's cost and value	3
object static/dynamic distinction	2
people and community	4
properties of associated tasks	4
topics covered	5
work/leisure distinction	10
(three other object-related groups)	3

Figure 1: Groups of criteria at the top of our heirarchy. Numbers are the number of participants who used each criteria at least once—note that some participants used some criteria, or criteria in the same group, more than once.

4.1 General observations

Our ten participants provided 64 sorts, a median 5.5 sorts each (first/third quartile 5.0/8.0 sorts each). At the leaves of our hierarchy, there was in general little overlap: 13 of 25 criteria were used by only one participant. However, 12 were used by two or more participants, 8 by three or more, and one criteria (discussed below) was used by every participant for at least one sort. Figure 1 summarises how many participants were represented in each top-level group, that is each group at the top of our hierarchical coding.

The single most common criterion was a distinction between objects used for work and objects used for leisure—all our participants used this criteria, and typically early on. Following this there were four common groups: to do with the object itself, especially the form (data type and other surface features), which 7 participants used at least once; the topics an object is connected with (5 participants); the properties of tasks associated with an object (4); and criteria describing people and community (4). A striking finding is the diversity in the criteria derived by the participants. Although we were able to group the criteria into 13 distinct high-level cateorgies, only two of these, work/leisure and form, were named by more than half the population.

After our initial grouping, which was based on labels proffered by participants and not on any statistics of the sorts themselves, there were no clear correlations between criteria—that is, "work" did not look the same as "important" or "Word files" as "makes me angry".

On our analysis, five of Kwaśnik's seven groups were represented in our data: situation, document, time, value, and cognitive state. However, they were very unevenly distributed: criteria we classified as "situation" were used by all ten participants, for one or two sorts each; "document" was used by nine participants, for a median 3.5 sorts each; while at the other end of the scale, "time" was used by only three participants and "cognitive state" was used by two participants, once each in each case.

4.2 Criteria, groups, and tool support

In many previous studies two groups of criteria—form and topic—were found to be central, and our data reinforces this. There are also, however, notable contrasts.

Work/leisure. To the best of our knowledge, previous studies of classification behaviour have not found a work/leisure distinction. However, every participant in our sample used this criteria. This may be because we most of the objects in our study were digital, not paper documents—it is very easy to mingle work- and leisure-related objects online—but it is clearly important and is not explicitly supported by PIM tools.

One participant reported that he used two top-level folders in his file system, and two email accounts, to keep work and leisure information separate. No other participants reported as clear a distinction, however. It should be possible in a PIM/search tool to tag files, or e.g. learn a classifier, to help maintain this distinction. Distinguishing work from leisure contexts might also allow different technologies to be used in each case.

People. Our participants did associate their objects in terms of specific people, but not in the way we might have expected. Rather than linking objects to particular, specified people, our dimensions relate to relationships with the community: "popular with many people", for example, "things I will/won't talk about", or "involvement of other people". Unfortunately it is not clear how a PIM or search tool could support this sort of classification.

Task, time, and workflow. Users in our study did not group objects by particular tasks—objects related to task A, to task B, etc—but four users did group objects according to whether an object had an associated task, and by properties of that task (state, importance, and cost or difficulty). This could be used to extend the work of Jones and his colleagues [16], who advocate project organisation, but do not allow tasks within projects to be annotated with properties such as cost or importance.

Time was mentioned by four participants. However rather than categorise objects according to time of use (or receipt), as supported by a number of tools, three participants derived criteria from the lifecycle of an object. Criteria such as "when I need to act on this" or "when this is important" will change over time for each object. This is related to Cole's "level of information" dimension. Only one participant used objects' importance to an event, at a particular time, as a criteria. Tools which support an explicit notion of document or task lifecycle, or approximate this e.g. by recording patterns of use, would presumably suit these participants.

5. DISCUSSION OF THIS APPROACH

The results above suggest that card sorts, in this variant, are useful for eliciting criteria: it does seem possible to gain some insight into how users think of personal objects, and how we might support this. This pilot has, however, highlighted some limitations.

The objects represented by each participant's cards were familiar—that is, they tended to choose objects they had used recently or frequently. They were also selected for sharing, since although we did not record the card titles we did see them. We cannot be sure that the chosen objects represent the sorts of things users may search for in a PIM system, and of course they are not representative of *un*familiar objects. We could instead choose objects from a participant's computer, for example by choosing randomly from the file

system and labelling cards with file icons and names, and similar. There is a tradeoff, however: if participants did not recognise these objects, the only possible criteria would be file icons and names, and we would learn little. By allowing users to choose their own objects the cards hopefully acted as prompts for other, richer, associations.

On a related point, some media types, such as video or audio, are difficult to represent on cards. It is not clear what this means for eliciting criteria. We are possibly unlikely to get criteria such as "out of focus" (for photos) or "scratchy bit in the middle" (for audio), but participants' familiarity with the objects may mitigate this to some extent.

There are of course properties that are not captured by this method: links between documents, for example (except implicit links of the type "sorted into the same pile"). It is also possible that our presence, and the apparatus we used, made it hard for participants to think naturally. They may have been inspired to create other categories if prompted (as in Gonçalves and Jorge [13]); on the other hand, our approach has the advantage that we can see which properties were immediately obvious.

We also note that the objects chosen varied greatly from participant to participant, and this may have played a role in the criteria that were chosen—although we did see some overlap, possibly there would have been more if the objects were more similar. We are considering constraining participants more in future, for example by prompting them to make a certain number of cards for each media or perhaps having subpopulations sort a shared set of cards, e.g. emails, web pages etc. they have all seen or received.

6. CONCLUSIONS

We hope to extend this work by scaling to a larger group of participants, but we will consider some methological changes: constraining the objects chosen, for example, or careful prompts to elicit more classifications. Nonetheless card sorts, in the variant here, have proved useful for starting to understand how users think of "personal" digital objects from a wide range of sources and media. Some classifications were both common and expected, but we did observe interesting differences both with the criteria found in earlier studies and with the criteria exposed in PIM and search tools.

7. ACKNOWLEDGEMENTS

We would like to thank our participants for their time.

8. REFERENCES

[1] E. Adar, D. Karger, and L. A. Stein. Haystack: Per-user information environments. In *Proc. CIKM*, pages 413–422, 1999.

[2] D. K. Barreau and B. Nardi. Finding and reminding: File organization from the desktop. *ACM SIGCHI Bulletin*, 27(3):39–43, 1995.

[3] O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved search engines and navigation preference in personal information management. *ACM Trans. Inf. Syst.*, 26(4):1–24, 2008.

[4] T. Blanc-Brude and D. L. Scapin. What do people recall about their documents?: Implications for desktop search tools. In *Proc. IUI*, pages 102–111, 2007.

[5] R. Boardman and M. A. Sasse. "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *Proc. CHI*, pages 583–590, 2004.

[6] D. O. Case. Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *JASIST*, 42(9):657–668, 1991.

[7] I. Cole. Human aspects of office filing: Implications for the electronic office. In *Proc. Human Factors Society*, 1982.

[8] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with Phlat—personal search and organisation made easy. In *Proc. CHI*, pages 261–270, 2006.

[9] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton. Extending document management systems with user-specific active properties. *ACM Trans. Inf. Syst.*, 18(2):140–170, 2000.

[10] D. Elsweiler. *Supporting Human Memory in Personal Information Management*. PhD thesis, The University of Strathclyde, 2007.

[11] D. Elsweiler, M. Baillie, and I. Ruthven. What makes re-finding information difficult? A study of email re-finding. *Proc. ECIR*, 6611:568–579, 2011.

[12] E. Freeman and D. Gelernter. Lifestreams: a storage model for personal data. *SIGMOD Record*, 25(1):80–86, 1996.

[13] D. Gonçalves and J. A. Jorge. Describing documents: what can users tell us? In *Proc. IUI*, pages 247–249, 2004.

[14] J. Hackos and J. Redish. *User and Task Analysis for Interface Design*. 1998.

[15] W. Jones. Personal information management. *Annual Review of Information Science and Technology*, 41(1):453–504, 2007.

[16] W. Jones, H. Bruce, A. Foxley, and C. Munat. The universal labeler: Plan the project and let your information follow. In *Proc. ASIST*, 2005.

[17] W. Jones, H. Bruce, E. Jones, and J. Vinson. How do people keep and re-find project-related information? In *Proc. SIGCHI*, 2010.

[18] V. Kaptelinin. UMEA: translating interaction histories into project contexts. In *Proc. CHI*, pages 353–360, 2003.

[19] B. H. Kwaśnik. How a personal document's intended use or purpose affects its classification in an office. In *Proc. SIGIR*, pages 207–210, 1989.

[20] M. Lansdale. The psychology of personal information management. *Appl Ergon*, 19(1):55–66, 1988.

[21] M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz. Milestones in time: The value of landmarks in retrieving information from personal stores. In *Proc. INTERACT*, pages 184–191, 2003.

[22] G. Rugg and P. McGeorge. The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 22(3):94–107, 2005.

[23] S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *Proc. CHI*, pages 276–283, 1996.

Using semantic differentials for an evaluative view of the search engine as an interactive system

Frances Johnson
Department of Languages, Information & Communications
Manchester Metropolitan University
Geoffrey Manton
+44 161 247 6156
F.Johnson@mmu.ac.uk

ABSTRACT
In this paper, we investigate the use of semantic differentials in obtaining the evaluative view held by users of the search engine. The completed scales of bipolar adjectives were analysed to suggest the dimensions of the user judgment formed when asked to characterize a search engine. These were then used to obtain a comparative evaluation of two engines potentially offering different types of support (or assistance) during a search. We consider the value of using the semantic differential as a technique in the toolkit for assessing the user experience during information interactions in exploratory search tasks.

Categories and Subject Descriptors H3.3 [Information search and retrieval]; Search process. H.5.2 [User interfaces]: Evaluation/methodology

General Terms
Measurement, Performance, Design, Human Factors

Keywords
Semantic Differentials, User Evaluation, Exploratory Search, Information Interaction, User Interface Design,

1. INTRODUCTION

The design of interfaces to support exploratory search seeks to provide users with the tools for and the experience of an interactive and engaging search. This is a departure from the classic model of information retrieval wherein the user submits a keyword query to the system and scans the list of retrieved results for relevance, either stopping with relevant results or refining the query to get results that are closer to the information need. Exploratory search does not necessarily assume that the user has a well defined information need (at least one that can be articulated as a keyword query) or indeed that the query will be 'static' and thus satisfied by a single list of retrieved results.

Accordingly, search engine developments have focused on providing query assistance drawing on contextual aspects to the search, such as personal history and/or current context [9]. At the interface, developments focus on improving the search process via richer information representations and interactions, such as previews and facets through to tools that allow the user to view and explore connections in the results, for example 'the relation browser data analysis tool' [10]. These shifts into HCIR are intended to help in the various stages of search, from starting the task and understanding the query topic, throughout the search in deciding what to do next, and to stopping with a sense of confidence. In short, developments aim to support true exploration of the search and, whilst many efforts may fall short, they will provide some form of user support in query assistance and in improving the search process as an interactive experience.

The context for evaluation is predicated on White and Roth's [3] model of the exploratory search process. This involves the searcher in a dynamic interplay between the cognition of their 'problem space' and their exploratory activities in the iterative search process including the query formulation, results examination and information extraction. Data collected on the searcher's information interactions may confirm this model [7] as well as attempt to systematically evaluate the effectiveness of exploratory search systems. In evaluation, a framework is used to attempt to assess performance during the search stages and to relate aspects of the system to its role in supporting information exploration, including sense making or query visualisation [5]. The challenge for the evaluation of exploratory search is the assumption that the user is willing or able to make an evaluative judgment throughout the search or that valid measures can be found through their actions, for example of usage of query terms. In general, evaluation draws from established HCI measures of effectiveness (can people complete their tasks?) efficiency (how long do people take?), an assessment of the user's overall satisfaction or other affective responses. Where possible, and increasingly so, the user actions are observed and recorded as dependent on the system and/or its interface. In this study we focus on an attempt to obtain the user's evaluative view of the search engine, based on criteria which may be affected by the developments for new and richer interactive designs. It is assumed that this would be part of an assessment which when taken with others will build a picture of the 'user experience' of the system used in exploratory search.

2. USER EVALUATION

In developing an instrument to collect the user assessment effort goes into ensuring that the evaluation is made in the task context. It means little to know that the user is 'satisfied' with the interface without gaining insight into why this assessment has been formed. A variety of questionnaires have been developed for assessing usability of interactive systems, such as search engines. Two well known are the SUS (System Usability Scale) developed at the Digital Equipment Corporation [2] and the QUIS (Questionnaire for User Interaction Satisfaction) from the University of Maryland [4]. Both assess usability from the user perspective with 10 statements and rating scales in the SUS and the QUIS with 27 questions. The QUIS asks the user to respond on a rating scale to statements which address specific usability aspects of the system, such as "use of the terms were consistent throughout the website". The SUS on the other hand focuses on collecting the users' overall reaction to the site/system on statements, such as "I found the website unnecessarily complex". Arguably the QUIS focuses on the concerns that a developer might have when assessing usability whilst the SUS assumes that the user's overall assessment is a reflection on the extent to which their goal directed tasks were facilitated by the system and its design.

Questionnaires, such as SUS, are used in an experimental set up when an explanation of the user's overall assessment is sought. However, the limitations of the questionnaire to capture and provide insight into the complexity of the user's assessment has lead to alternative tools, for example Microsoft's Product Reaction Cards in the "Desirability Toolkit". This invites participants on a usability test to select as many, or as few, words from a list of 118 which best describe their reaction and/or interaction with the system they have just used. Benedek and Miner [1] includes a list of the words used and point out that the approach helps elicit negative comments as well as positive, thus overcoming a problem with questionnaires biased towards positive responses.

Given the potential scope of the users' response (represented in the reaction cards with some 100+ terms) this study sets out to investigate the value in assembling these into a framework (of sorts) for the collection of the users' evaluative judgment of an interactive system based on the technique known as 'semantic differentials'. Specifically the aim of this small preliminary investigation was to begin to determine the extent to which users hold an evaluative view of a 'search engine' and, what are the dimensions (traits or criteria) on which we form this view. If it can be found that this view is strongly held (that is, an attitude is formed which may influence how we behave and interact with the search engine) then it may be feasible to investigate the influence, if any, of a design for information interaction on the evaluative view. In this study the technique of semantic differentials is used to best describe the evaluative view held by its participants. This is then employed to assess two quite different search engines following the completion of two query based searches.

3. SEMANTIC DIFFERENTIALS

Semantic Differentials (SDs) originate from the work of Osgood [8] as a technique for attitude measurement, scaling people on their responses to adjectives in respect to a concept. Typically individuals respond to several pairs of bipolar adjectives scored on a continuum + to – and in doing so differentiate their meaning of the concept in intensity and in direction (in a 'semantic space'). The assumption made here, in the use of SDs on 'search engines' is that users hold an evaluative view which is formed when using the engine to find and/or explore information. The SD is used to investigate the adjectives that best 'conceptualise' the search engine, from the user perspective. Factorial analysis is also used to identify the dimensions of the judgment, in a sense the packaging of the components of the judgment into smaller units of meaning reflecting what is important when responding to the concept 'search engine'.

The design of the SD aims to allow a degree of abstraction in the evaluation so that participants can reflect the complexity of their response. In this study, the adjectives to include on the SD scale were chosen from Microsoft's Product Reaction Cards, these having been collected in previous research, usability studies and in the marketing of web sites and systems. The majority of the terms formed pairs on some continuum and 40 terms (20 pairs) were selected to present in the SD. The selection was subject to the judgment of the researcher. This is a limitation of this exploratory study, however some steps were taken to formalise the selection. A loose grouping of the adjective pairs was made as relating to appearance (such as 'attractive'), judgment ('relevant'), emotive ('boring') and use ('fast'). Five pairs from each of these groupings were made. The pairs were mixed on the SD to avoid having all the positive terms on one side of the scale and only intervals were shown on the scales with the numerical values used only for data entry. This allowed participants to focus on how an adjective pair related to the engine and its characteristics, rather than on 'scoring' it in some way.

3.1 Implementation

The study was conducted on our undergraduates studying BSc Web Development and on a postgraduate cohort studying on MA Library and Information Management or the MSc Information Management. A total of 89 students participated in the study. At the start of the class each participant was asked to think about a search engine, and adjectives they would use to describe the engine, (in other words, "what it means to them"). Each participant was then given the SD to complete. This is referred to as the 'baseline' and the data were analysed to gauge user perceptions of search engines.

In the following lab sessions (about one hour later) each participant was required to perform two search tasks on each of the two search engines - Google, an engine we can assume some familiarity and, a second clustering engine (Yippy, formerly Clusty). The two tasks were as follows

1. Find information on the symptoms for diabetes type II
2. Find information to help write an assignment on the debate 'nurture vs nature'

These were selected to give the participants experience of using the engines for a closed question (find symptoms) and on a more open 'informational' type of query (on the 'nature nurture' debate). A measure of search success was not taken as the aim was simply to get the participants using the engines. The order of use of the two sites was randomized so that approximately half of the participants worked on Google first and half on the clustering engine. All were told to spend no longer than 10 minutes searching on each engine and to complete the SD for each engine immediately after each use.

4. FINDINGS

4.1 Evaluative views

The responses to the baseline (*think of an engine*) were entered into SPSS with the scales coded (7-1) so that the positive adjectives corresponded to the higher numbers. Descriptive statistics of mean, mode and standard deviation were calculated for each of the adjectives. Those with a mean greater than 4 or less than 3 were taken to suggest the adjective pairs that best characterise the participants' view, as follows

attractive	unattractive
powerful	simplistic
valuable	not valuable
relevant	irrelevant
satisfying	frustrating
fast	slow
predictable	unpredictable
intuitive	rigid
easy	difficult

Factor analysis investigates the correlations among subsets of the responses to the bipolar pairs and groups the correlated variables such that each group is largely independent of the others. Exploratory factor analysis was employed to identify the groups which might explain most of the variance in the data. With 20 pairs of adjectives to perform Principal Components Analysis (PCA) in SPSS it is recommended that a minimum of 100 responses are obtained, whilst others recommend that a sample requires approx 5-10 times the number of people as scale pairs [6]. With 89 responses we should use a reduced number of pairs, however the Kaiser-Meyer-Olkin measure of sampling adequacy (.616) is greater than the 0.6 needed to indicate that the correlations matrix may be able to factorise. So with this, PCA was run (with varimax rotation to force items to 'load' with only one factor group), to identify the possible 'factors' or subsets derived from patterns of correlation of the adjective pairs. The following five subsets were obtained (the adjectives from the list above having a low or high mean are shown in bold). The labels were assigned to suggest the evaluative dimension.

Factor 1 *label* USE – Utility

effective, **valuable, satisfy**, **relevant**, **predictable**, intimidating, inspiring, stimulating

Factor 2 *label* QUALITY – Affective

engaging, fun, connected

Factor *3 label* QUALITY - Appearance

high quality, personal, meaningful, **rigid, attractive**

Factor 4 *label* USE – Efficient

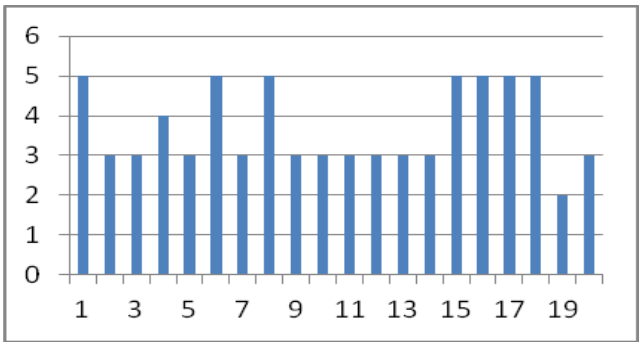
easy, intuitive, fast, powerful

Factor 5 *label* USE - Control

controllable

4.2 Comparative evaluations

Using the same SDs, participants scaled their responses post search using Google and the clustering search engine. These were entered into a worksheet to obtain basic statistics. The mode for each adjective is shown Figure 1 with a note of those with mode >4 and < 3 suggesting a positive or negative response.



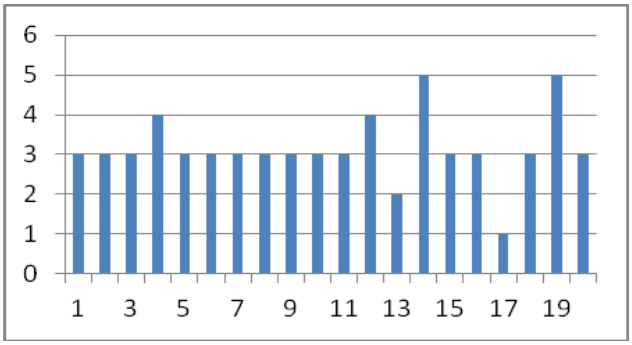
Google (mode > 4 or < 3)

& in bold where mean is also > 4

1**attractive** - , 6valuable - , 8relevant - ,

15**satisfying** - , 16**fast** - , 17predictable - ,18**controllable** -,

and (where mode < 3) 19rigid -



Clustering search engine (mode > 4 or < 3)

& in bold where mean > 4 or < 3

14engaging - , 19intuitive –

and (where mode < 3)

13intimidating - , 17 – **unpredictable**

Figure 1. Responses to the adjectives for both engines

Using the suggested dimensions or aspects of the user evaluation from the factor analysis of the 'baseline' data we can compare the participants' responses on the high or low scoring adjectives across the engines. On `QUALITY – Appearance` Google was rated rigid and attractive and whereas Google was neutral on the factor `QUALITY- Affective`, the clustering search engine obtained a positive score towards the adjective engaging. On the factor labeled `USE- Utility` Google was scored as

predictable, valuable, relevant and satisfying, whereas the clustering engine as unpredictable and towards intimidating. On USE-Efficient Google was rated as fast and the clustering engine appears more intuitive. Google was also rated as controllable.

5. DISCUSSION

This is an exploratory study and it has its limitations. It is questionable whether the selection of the adjectives to use in the SD influenced the results. In particular there is uncertainty in the results that *intuitive to rigid* is on some continuum. Also there is some unease at accepting a factor with 8 out of 20 pairs and one with only one. Perhaps the sample size was too small to attempt factoring. The results also raise questions on how some of the adjectives were interpreted by the participants. These withstanding, the participants in this study did appear to hold an evaluative judgment of the concept 'search engine' and the traits represented in the scale were grouped to suggest the aspects on which an assessment may be formed. It is of particular interest that upon using the search engine Google to conduct a search task the ratings on the SD, on the whole, altered only in the factors of 'controllable' and USE -efficient (easy, intuitive and powerful). Perhaps we can assume that Google was the typical engine when asked to think of an engine in the baseline and, when it came to *use* Google, users shifted their perception with regards to some of the adjectives. Perhaps this is not surprising but it may suggest that we hold an implicit view of search engines, and that this view will be influenced by actual use (and the experience). Our participants may have had less familiarity with the clustering engine, and in the evaluation this appears to have prompted an 'affective' response in finding the engine to be 'engaging' whilst also indicating shifts in the 'use' factors (towards an assessment of the engine as 'unpredictable'). Again the infallibility of some of the terms is highlighted where an 'unpredictable' system may be regarded to be a negative judgment, but if the system is also considered to be engaging the assessment could be highly desirable depending on the user's goals. This study of the use of semantic differentials indicates that it is worth running the test with a new cohort of students to determine the extent to which a consistent view is obtained. As an exploratory study it also suggests that further research on user's perceptions and mental models of search engines is worthwhile. With regards to the challenge of providing an evaluation of the exploratory search, this study falls short as no behavioural data was obtained. However, perhaps, with further design of the SD and use in an experimental set up with honed tasks, a user assessment of the interface may be obtained as dependent on the search interface development and design.

6. REFERENCES

[1] Benedek, J. and Miner, T. "*Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting*." Redmond, WA: Microsoft Corporation, 2002. http://www.microsoft.com/usability/UEPostings/Desirability_Toolkit.doc

[2] Brooke, J. SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor & Francis, 1996 [*www.itu.dk/courses/U/E2005/litteratur/**sus**.pdf#*

[3] Capra, R., and Marchionini, G. The Relation Browser tool for faceted exloratory search. Proceedings of the 2008 Conference on Digital Libraries, Pittsburg, Pennsylvania, June, 2008

[4] Chin, J. P., Diehl, V. A, & Norman, K. Development of an instrument measuring user satisfaction of the human-computer interface, Proceedings of ACM SIGCHI ,1988, pp. 213-218. http://www.cs.umd.edu/hcil/quis/

[5] Daqing He, et al An evaluation of adaptive filtering in the context of realistic task-based information exploration. . *Information Processing Management*, 44(2), 2008 pp. 511-533

[6] Gable, R. K., & Wolf, M. E.. *Instrument development in the affective domain* (2nd ed.). Boston: Kluwer Academic, 1993

[7] Kules, B and Capra, R. Visualizing stages during an exploratory search. Proceedings HCIR October 20th, 2011.

[8] Osgood, C.E, Suci, G., & Tannenbaum, *P The Measurement of Meaning*. University of Illinois Press, 1957

[9] Teevan, J., Dumais,S.T and E. Horvitz. *Potential for Personalization*. ACM Transactions on Computer-Human Interaction special issue on Data Mining for Understanding User Needs, 17(1), 2010 http://people.csail.mit.edu/teevan/work/publications/ papers/tochi10.pdf

[10] White, Ryen W. & Roth., R. A. *Exploratory Search: Beyond the Query-Response Paradigm*, CA: Morgan and Claypool, 2009

Appendix: The Semantic Differential scale

attractive	_	_	_	_	_	_	_	unattractive
impersonal	_	_	_	_	_	_	_	personal
dull	_	_	_	_	_	_	_	fun
powerful	_	_	_	_	_	_	_	simplistic
disconnected	_	_	_	_	_	_	_	connected
valuable	_	_	_	_	_	_	_	not valuable
high quality	_	_	_	_	_	_	_	low quality
irrelevant	_	_	_	_	_	_	_	relevant
effective	_	_	_	_	_	_	_	ineffective
incomprehensible	_	_	_	_	_	_	_	meaningful
stimulating	_	_	_	_	_	_	_	confusing
boring	_	_	_	_	_	_	_	inspiring
intimidating	_	_	_	_	_	_	_	empowering
stressful	_	_	_	_	_	_	_	engaging
satisfying	_	_	_	_	_	_	_	frustrating
fast	_	_	_	_	_	_	_	slow
predictable	_	_	_	_	_	_	_	unpredictable
controllable	_	_	_	_	_	_	_	uncontrollable
intuitive	_	_	_	_	_	_	_	rigid
difficult	_	_	_	_	_	_	_	eas

The fault, dear researchers, is not in Cranfield, But in our metrics, that they are unrealistic.

Mark D. Smucker
Department of Management Sciences
University of Waterloo, Canada
mark.smucker@uwaterloo.ca

Charles L. A. Clarke
School of Computer Science
University of Waterloo, Canada
claclark@plg.uwaterloo.ca

1. INTRODUCTION

As designers of information retrieval (IR) systems, we need some way to measure the performance of our systems. An excellent approach to take is to directly measure actual user performance either in situ or in the laboratory [12]. The downside of live user involvement is the prohibitive cost if many evaluations are required. For example, it is common practice to sweep parameter settings for ranking algorithms in order to optimize retrieval metrics on a test collection. The Cranfield approach to IR evaluation provides low-cost, reusable measures of system performance.

Cranfield-style evaluation frequently has been criticized as being too divorced from the reality of how users search, but there really is nothing wrong with the approach [18]. The Cranfield approach effectively is a simulation of IR system usage that attempts to make a prediction about the performance of one system vs. another [15].

As such, we should really be thinking of the Cranfield approach as the application of models to make predictions, which is common practice in science and engineering. For example, physics has equations of motion. Civil engineering has models of concrete strength. Epidemiology has models of disease spread. Etc. In all of these fields, it is well understood that the models are simplifications of reality, but that the models provide the ability to make useful predictions.

Information retrieval's predictive models are our evaluation metrics.

The criticism of system-oriented IR evaluation should be redirected. The problem is not with Cranfield — which is just another name for making predictions given a model — the problem is with the metrics.

We believe that rather than criticizing Cranfield, the correct response is to develop better metrics. We should make metrics that are more predictive of human performance. We should make metrics that incorporate the user interface and realistically represent the variation in user behavior. We should make metrics that encapsulate our best understanding of search behavior.

In popular parlance, we should bring solutions, not problems, to the system-oriented IR researcher. To this end, we have developed a new evaluation metric, time-biased gain (TBG), that predicts IR system performance in human terms of the expected number of relevant documents to be found by a user [16].

2. TIME-BIASED GAIN

HCI has a long history of automated usability evaluation [10], and indeed, so does IR. Cleverdon designed the Cranfield 2 study carefully in terms of a specific type of user and how this type of user would define relevance [8, p. 9]. Taken together, a test collection (documents, topics, relevance judgments) and an evaluation metric allow for the simulation of a user with different IR systems.

Järvelin and Kekäläinen produced a significant shift in evaluation metrics with their introduction of cumulated gain-based measures [11]. The cumulated gain measures are explicitly focused on a model of a user using an IR system. As long as the user continues to search, the user can continue to increase their gain. The common notion of gain in IR evaluation is the relevant document, but gain can be anything we would like to define it to be.

Cumulated gain can be plotted vs. time to produce a gain curve and compare systems. The curve that rises higher and faster than another curve is the preferred curve. While we can plot gain curves of one system vs. another, it is well-known that users do not endlessly search; different users stop their searches at different points in time for a host of reasons. Given a probability density function $f(t)$ that gives the distribution of time spent searching, we can compute the expected gain as follows:

$$E[G(t)] = \int_0^\infty G(t)f(t)dt, \tag{1}$$

where $G(t)$ is the cumulated gain at time t. Equation 1 represents *time-biased gain* in its general form, i.e. time-biased gain is the expected gain for a population of users.

While it is natural for us to talk about cumulated gain over time, the traditional cumulated gain measures have substituted document rank for time and implicitly model a user that takes the same amount of time to evaluate each and every document. By making time a central part of our metric, we gain the ability to more accurately model behavior. For example, in a document retrieval system, longer documents will in general take users longer to evaluate, and if the retrieval system presents results with document summaries (snippets), we know that users can use summaries to speed the rate at which they find relevant information [14].

Another significant advantage of using time directly in our retrieval metric is that we now make testable predictions of human performance. Our predictions are in the same units as would be obtained as part of a user study. To our knowledge, this alignment between the units of Cranfield-style metrics and user study metrics has not previously existed.

Time-biased gain in the form of Equation 1 makes no mention of ranked lists of documents, for it is a general purpose description of users using an IR system over time. To produce a metric suitable for use in evaluating ranked lists, we followed a process common to development of new simulations [3]:

1. Creation of model.
2. Calibration of model.
3. Validation of model.

Our first step in model creation was to adopt the standard model of a user that works down a result list and move Equation 1 to a form common to cumulated gain measures:

$$\sum_{k=1}^{\infty} g_k D(T(k)), \tag{2}$$

where g_k is the gain at rank k, $T(k)$ is the expected time it takes a user to reach rank k, and $D(t)$ is the fraction of the population that survives to time t and is called the decay function.

Our model for the time it takes a user to reach rank k, $T(k)$, takes into consideration a hypothetical user interface that presents results to the user in the form of document summaries. A click on a document summary takes the user to the full document. We model both the probabilities of clicking on summaries given their NIST relevance and the probability of then judging a viewed full document as relevant. We separately model the time to view summaries and full documents. For the time spent on a full document, we modeled longer documents taking longer with an additional constant amount of spent. We treated duplicate documents as zero length documents. We then calibrated $T(k)$ using data from a user study, and finally we validated that our $T(k)$ provided a reasonable fit to the user study data. Likewise, we modeled $D(t)$ as exponential decay fit to a search engine's log data.

In contrast, older evaluation metrics such as mean average precision [19, p. 59] cannot be calibrated and have only been validated after their creation. For example, the work of Hersh and Turpin [9] is likely the first attempt to validate a metric (average precision). Many recent metrics can be calibrated to actual user behavior [4, 5, 7, 17, 20, 21], but their calibration and validation often come after their release and adoption.

3. CONCLUSION

The Cranfield approach to IR evaluation is merely another name for the development and use of predictive models, which is a fundamental part all science and engineering fields. In particular, IR evaluation fits nicely into the framework of simulation where models are created, calibrated, and validated before being used to make predictions. We have presented time-biased gain as an example of what we believe the correct direction is for IR system evaluation. We are not the only ones to be working on better metrics or taking a simulation based approach [2, 13], and others also consider time an important part of evaluation [1, 6].

Our position is that system-oriented IR research is user-oriented IR research given its use of evaluation metrics that model users. If HCIR researchers can produce better models than exist today — by better, we mean more predictive of human performance — then we can help system development to focus on changes that help users better search.

4. ACKNOWLEDGMENTS

This work was supported in part by the NSERC, in part by GRAND NCE, in part by Google, in part by Amazon, in part by the facilities of SHARCNET, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

5. REFERENCES

[1] L. Azzopardi. Usage based effectiveness measures: monitoring application performance in information retrieval. *CIKM*, pages 631–640, 2009.

[2] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44:35–47, January 2011.

[3] J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol. *Discrete-Event System Simulation*. Prentice Hall, 5th edition, 2010.

[4] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*, pages 611–620, 2011.

[5] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, Hong Kong, 2009.

[6] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *SIGIR*, pp. 206–213. 1997.

[7] G. Dupret. Discounted cumulative gain and user decision models. In *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, pages 2–13, Berlin, Heidelberg, 2011. Springer-Verlag.

[8] D. Harman. *Information Retrieval Evaluation*. Morgan & Claypool, 2011.

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

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

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[12] D. Kelly. *Methods for Evaluating Interactive Information Retrieval Systems with Users*, volume 3. Foundations and Trends in Information Retrieval, 2009.

[13] H. Keskustalo, K. Järvelin, T. Sharma, and M. L. Nielsen. Test collection-based IR evaluation needs extension toward sessions: A case of extremely short queries. In *AIRS*, pp. 63–74, 2009.

[14] R. Khan, D. Mease, and R. Patel. The impact of result abstracts on task completion time. In *Workshop on Web Search Result Summarization and Presentation, WWW'09*, 2009.

[15] J. Lin and M. D. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *SIGIR'08*, pages 19–26. ACM, 2008.

[16] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *SIGIR*, 10 pages, 2012.

[17] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR'09*, pages 508–515. ACM, 2009.

[18] E. M. Voorhees. I come not to bury Cranfield, but to praise it. In *HCIR'09*, pages 13–16, 2009.

[19] E. M. Voorhees and D. K. Harman, editors. *TREC*. MIT Press, 2005.

[20] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *CIKM*, pages 1561–1564, Toronto, 2010.

[21] Y. Zhang, L. A. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13:46–69, February 2010.

A Model of Consumer Search Behaviour

Tony Russell-Rose
UXLabs
London
UK
+44 (0)7779 936191
tgr@uxlabs.co.uk

Stephann Makri
University College London Interaction Centre,
University College London, Gower St.
London, WC1E 6BT, UK
+44 (0)20 7679 0696
s.makri@ucl.ac.uk

ABSTRACT

In order to design better search experiences, we need to understand the complexities of human information-seeking behaviour. In previous work [13], we proposed a model of information behavior based on an analysis of the information needs of knowledge workers within an *enterprise search* context. In this paper, we extend this work to the *site search* context, examining the needs and behaviours of users of consumer-oriented websites and search applications.

We found that site search users presented significantly different information needs to those of enterprise search, implying some key differences in the information behaviours required to satisfy those needs. In particular, the site search users focused more on simple "lookup" activities, contrasting with the more complex, problem-solving behaviours associated with enterprise search. We also found repeating patterns or 'chains' of search behaviour in the site search context, but in contrast to the previous study these were shorter and less complex. These patterns can be used as a framework for understanding information seeking behaviour that can be adopted by other researchers who want to take a 'needs first' approach to understanding information behaviour.

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

Site search, enterprise search, information seeking, user behaviour, search modes, information discovery, user experience design.

1. INTRODUCTION

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 [2] has shown that the need underlying a given web search could in fact be navigational (e.g. to find a particular site) or transactional (e.g. through online shopping, social media, etc.). Similarly, Rose & Levinson [12] have identified the consumption of online resources as a further common category of search behaviour.

In this paper, we examine the needs and behaviours of individuals across a range of site search scenarios. These are based on an analysis of user needs derived from a series of customer engagements involving the development of customised site search applications. In so doing, we extend and validate a model of information behaviours derived from a previous study of enterprise search users [13].

The model is based on a set of 'search modes' that users employ to satisfy their information search and discovery goals. It extends the IR concept of information-seeking to embrace a broader notion of discovery-oriented problem solving, addressing a wider range of information interaction and information use behaviours. The overall structure of the model reflects Marchionini's [9] framework, and consists of three lower-level 'lookup' modes (*locate*, *verify* and *monitor*), three "learn" modes (*compare*, *comprehend* and *explore*) and three higher-level "investigate" modes (*analyze*, *evaluate* and *synthesize*).

We investigate the degree to which the model extends to accommodate the domain of *site search* (i.e. consumer-oriented websites and search applications) and discuss some of the differences between the needs and goals of enterprise search users versus those of site search. We conclude by exploring the ways in which these modes combine to form distinct chains or patterns, and reflect on the value this offers as a framework for expressing complex patterns of behaviour.

2. MODELS OF INFORMATION SEEKING

The framework investigated in this study is influenced by a number of existing models. For example, Bates [1] identified a set of 29 search 'tactics' which she organised into four broad categories, including *monitoring* ("to keep a search on track"). Likewise, O'Day & Jeffries [11] examined the use of information search results by clients of professional information intermediaries and identified three categories of behaviour, including *monitoring a known topic or set of variables over time* and e*xploring a topic in an undirected fashion*. They also observed that a given search scenario would often evolve into a series of interconnected searches, delimited by triggers and stop conditions that signalled transitions between modes within an overall scenario.

Cool & Belkin [3] proposed a classification of interaction with information which included *evaluate* and *comprehend*. They also proposed *create* and *modify,* which together reflect aspects of our *synthesize* mode.

Ellis and his colleagues [4, 5, 6] developed a model consisting of a number of broad information seeking behaviours, including *monitoring* and *verifying* ("checking the information and sources found for accuracy and errors"). In addition, his *browsing* mode ("semi-directed searching in an area of potential interest") aligns

with our definition of *explore*. He also noted that it is possible to display more than one behaviour at any given time. In revisiting Ellis's findings among social scientists, Meho and Tibbo [10] identified *analysing* (although they did not elaborate on it in detail). More recently, Makri et al [8] proposed *searching* ("formulating a query in order to locate information"), which reflects to our own definition of *locate*.

In addition to the research-oriented models outlined above, we should also consider practitioner-oriented views. Spencer [14] suggests four modes of information seeking, including *known-item* (a subset of our *locate* mode) and *exploratory* (which mirrors our definition of explore). Lamantia [7] also identifies four modes, including *monitoring*.

In this paper, we use the characteristics of the models above as a lens to interpret the behaviours found in a new source of empirical site search data. We also explore the combinatorial nature of the modes, extending Ellis's [5] concept of mode co-occurrence to identify and define a set of repeating patterns and sequences.

3. CONSUMER SEARCH BEHAVIOUR

3.1 Data Acquisition

The primary source of data in this study is a set of 277 information needs captured during client engagements involving the development of a number of custom site search applications. These information needs take the form of 'micro-scenarios', i.e. a brief narrative that illustrates the end user's goal and the primary task or action they take to achieve it, for example:

- *Find best offers before the others do so I can have a high margin*.
- *Get help and guidance on how to sell my car safely so that I can achieve a good price*.
- *Understand what is selling by area/region so I can source the correct stock*.
- *See year-on-year ad spend trends for TV and online to supply to the Head of Global Media*.

The scenarios were collected as part of a series of requirements workshops involving stakeholders and customer-facing staff from the respective client organisations. They were generated by participants in individual breakout sessions, and then moderated by the workshop facilitator in a group session to maximise consistency and minimise redundancy or ambiguity. They were also prioritised by the group to identify those that represented the highest value both to the end user and to the client organisation.

This data possesses a number of unique properties. In previous studies of information seeking behaviour (e.g. [5], [10]), the primary source of data has traditionally been interview transcripts that provide an indirect, verbal account of end user *information behaviours*. By contrast, the current data source represents a self-reported account of *information needs*, generated directly by end users (although a proportion were captured via proxy, e.g. through customer facing staff speaking on behalf of the end users). This change of perspective means that instead of using information behaviours to infer information needs and design insights, we can adopt the converse approach and use the stated needs to infer information behaviours and the interactions required to support them.

Moreover, the scope and focus of these scenarios represents a further point of differentiation. In previous studies, (e.g. [8]), measures have been taken to address the limitations of using interview data by combining it with direct observation of information seeking behaviour in naturalistic settings. However, the behaviours that this approach reveals are bounded by the functionality currently supported by existing systems and working practices, and as such do not reflect the full range of *aspirational* or *unmet* user needs encompassed by the scenarios in this study.

Finally, the data is unique in that is constitutes a genuine practitioner-oriented deliverable, generated expressly for the purpose of designing and delivering professional site search systems. As such, it reflects a degree of realism that interview data or other research-based interventions might struggle to replicate.

3.2 Data Analysis

These scenarios were analyzed using the model derived previously for the domain of enterprise search [13]. In this respect, the process was partially deductive, applying the model in a top-down fashion to classify the data. But it was also partially inductive, applying a bottom-up, grounded analysis to identify new types of behaviour not present in the original model or to suggest revised definitions of the existing categories.

Although the original study involved three separate analysts, the behaviours this time were identified by the first author alone. The current analysis approach is therefore much more subjective. However, the first author was also the facilitator at each of the requirements workshops at which the scenarios were generated, and was able to again a deep insight into the needs, goals and motivations of the participants. This allowed him to be as confident as possible in his understanding of the users' information needs and consistent in his interpretation of the information behaviours required to satisfy a particular need.

A number of the scenarios focused on needs that did not involve any explicit information seeking or use behaviour, e.g. "*Achieve a good price for my current car*". These were excluded from the analysis. A further number were incomplete or ambiguous, or were essentially feature requests (e.g. "*Have flexible navigation within the page*"), and were also excluded. This process resulted in further confirmation and validation of the nine search modes identified in the original study, but with revised definitions to reflect a broader scope:

1. **Locate**: *To find a specific (possibly known) item*, e.g. "Find my reading list items quickly". This mode encapsulates the stereotypical 'findability' task that is so commonly associated with site search, consistent with (but a superset of) Spencer's [14] *known item* search mode. This was the most frequent mode in the site search scenarios (120 instances).

2. **Verify**: *To confirm that an item meets some specific, objective criterion*, e.g. "See the correct price for singles and deals". Often found in combination with locating, this mode is concerned with validating the accuracy of some data item, comparable to that proposed by Ellis et al. [5] (39 instances).

3. **Monitor**: *Maintain awareness of the status of an item for purposes of management or control*, e.g. "Alert me to new resources in my area". This activity focuses on the state of asynchronous responsiveness and is consistent with that of Bates [1], O'Day and Jeffries [11], Ellis [4], and Lamantia [7] (13 instances).

4. **Compare**: *To identify similarities & differences within a set of items*, e.g. "Compare cars that are my possible candidates in detail". This mode has not featured prominently in previous models (with the possible exception of Marchionini's), but was found to be a significant component of enterprise search behaviour [13]. Moreover, it is a common feature of product search and navigation on many ecommerce sites. However, it occurred relatively infrequently in the site search scenarios (2 instances).

5. **Comprehend**: *To generate independent insight by interpreting patterns within a data set*, e.g. "Understand what my competitors are selling". Like *compare*, this mode was found to be a key element of the enterprise search scenarios, and also features in the models of Cool & Belkin [3] and Marchionini [9]. It occurred relatively frequently in site search (50 instances).

6. **Explore**: *To investigate an item or data set for the purpose of knowledge discovery*, e.g. "Find useful stuff on my subject topic". In some ways the boundaries of this mode are somewhat less prescribed than the others, but what the instances share is the characteristic of open ended, opportunistic search and browsing in the spirit of O'Day and Jeffries [11] *exploring a topic in an undirected fashion* and Spencer's [14] e*xploratory*. This mode was the second most common in site search (110 instances).

7. **Analyze**: *To examine an item or data set to identify patterns & relationships*, e.g. Analyze the market so I know where my strengths and weaknesses are". This mode features less prominently in previous models, appearing as a sub-component of the processing stage in Meho & Tibbo's [10] model, and overlapping somewhat with Cool & Belkin's [3] organize. This definition is also consistent with that of Makri et al. [8], who identified analysing as an important aspect of lawyers' interactive information behaviour and defined it as "examining in detail the elements or structure of the content found during information-seeking." (p. 630). Although the most common element of the enterprise search scenarios, it was less prevalent in site search (59 instances).

8. **Evaluate**: *To use judgement to determine the value of an item with respect to a specific goal*, e.g. "I want to know whether my agency is delivering best value". This mode is similar in spirit to *verify*, in that it is concerned with validation of the data. However, while *verify* focuses on simple, objective fact checking, our conception of *evaluate* involves more subjective, knowledge-based judgement, similar to that proposed by Cool & Belkin [3] (61 instances).

9. **Synthesize**: *To create a novel or composite artefact from diverse inputs*, e.g. "I need to create a reading list on celebrity sponsorship". This mode also appears as a sub-component of the *processing* stage in Meho & Tibbo's [10] model, and involves elements of Cool & Belkin's [3] *create* and *use*. Of all the modes, this one is the most commonly associated with information *use* in its broadest sense (as opposed to information *seeking*). It was relatively rare within site search (5 instances).

4. MODE SEQUENCES AND PATTERNS

Applying the modes described above provides a framework for understanding the needs of site search users, and an insight into their likely behaviours. But as with the previous study [13], their real value lies not so much in the individual instance data but in the patterns of co-occurrence they reveals. In most scenarios, modes combine to form distinct chains and patterns, echoing the transitions observed by O'Day and Jeffries [11] and the combinatorial behaviour alluded to by Ellis [5], who suggested that information behaviours can often be nested or displayed in parallel.

Just as new definitions were needed to accommodate the new domain, new patterns of occurrence were identified in the data. Typically these consisted of chains of length two or three, of which the following were most frequent:

1. **Insight-driven search**: (Explore->Analyze-> Comprehend): This patterns represents an exploratory search for insight to resolve an explicit information need, e.g. "*Assess the proper market value for my car*" (45 instances)
2. **Opportunity-driven search**: (Explore-Locate-Evaluate): In contrast to the explicit focus of the pattern above, this sequence represents a less directed exploration in the prospect of serendipitous discovery e.g. "*Find useful stuff on my subject topic"*(31 instances)
3. **Qualified search** (Locate-Verify) This pattern represents a variant of the stereotypical findability task in which some element of immediate verification is required, e.g. "*Find trucks that I am eligible to drive*" (29 instances)

A deeper insight into these patterns can be obtained by presenting them in diagrammatic form, as a network (Figure 1). This diagram illustrates the three sequences outlined above plus other commonly found patterns. It also reflects an outcome of the pervious study, in that certain modes tend to function as "terminal" nodes, i.e. entry points or exit points to a scenario. For example, *Explore* typically functions as an opening, while *Comprehend* and *Evaluate* function in closing a scenario. *Analyze* typically appears as a bridge between an opening and closing mode.

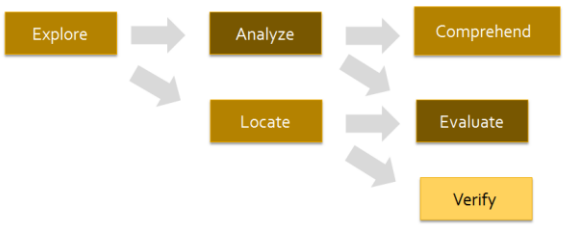


Figure 1. Mode network for site search

4.1 Site search vs. Enterprise Search

The sequences described above also allow us to reflect on some of the differences between the needs of site search users and those of enterprise search. One of the most fundamental differences is an emphasis on simpler "lookup" modes such as *Locate* and *Verify*: these were relatively rare in the enterprise search data, but prominent in site search (120 and 39 instances respectively). Enterprise search, by contrast, emphasised higher-level "investigate" behaviours such as *Analyze* and *Evaluate* (modes which also appeared frequently in site search, but not as prominently: 58 and 61 instances respectively). However, in

neither case was the stereotype of 'search as findability' borne out: even in site search (where it was the most common mode), *Locate* was accountable for no more than a quarter of all instances.

But perhaps the biggest difference was in the composition of the chains: while enterprise search was characterised by a wide variety of heterogeneous chains, site searched focused on a small number of common trigrams and bigrams. Moreover, these chains displayed little evidence of the composite nature observed in enterprise search, in which certain chains were seen to be embedded within others to create larger, more complex sequences of behaviour.

5. DISCUSSION

A key feature of the current model is its emphasis on the combinatorial nature of search modes, and the value this offers as a framework for expressing complex patterns of behaviour. Such an approach is not unique: the second author, for example, has also previously explored the concept of mode chains to describe information seeking behaviours observed in naturalistic settings. However, his approach was based on the analysis of complex tasks observed in real time, and as such was less effective in revealing consistent patterns of atomic behaviour such as those found in the current study.

Conversely, this virtue can also be a shortcoming: the fact that simple repeating patterns can be extracted from the data may be as much an artefact of the medium as it is of the information needs it contains. These scenarios were expressly designed to be a concise, self-contained deliverable in their own right, and applied as a simple but effective tool in the planning and prioritisation of software development activities. This places a limit on the length and sophistication of the information needs they encapsulate, and hence a natural boundary on the scope and extent of the patterns they represent. Their format also allows the analyst to apply perhaps an unrealistic degree of top-down judgement and iteration in aligning the relative granularity of the information needs to existing modes; a benefit that is less readily available to those whose approach involves real-time, observational data.

A further caveat is that in order to progress from understanding an information need to identifying the information behaviors required to satisfy those needs, it is necessary to *speculate* on the behaviours that a user *might* perform when undertaking a task to satisfy the need. It may transpire that users actually perform different behaviours which achieve the same end, or perform the expected behavior but through a combination of other nested behaviours, or may simply satisfy the need in a way that had not been envisaged at all.

Finally, the process of inferring information behaviour from self-reported needs can never be wholly deterministic, regardless of the consistency measures discussed earlier. In this respect, further steps should be taken to operationalize the application of the framework and apply some independent measure of stability or objectivity in its usage.

6. CONCLUSIONS

In this study we have investigated a model of information seeking behaviour derived from the domain of enterprise search, and validated its extensibility to users of consumer-oriented websites and search applications. In so doing, we explored a novel, goal-driven approach to eliciting user needs, and identified some key differences in user behaviour between the two domains.

In addition, we have demonstrated the value of the model as a framework for expressing complex patterns of behaviour, extending the IR concept of information-seeking to embrace a broader range of composite information interaction and use behaviours. Moreover, we propose that our method can be adopted by other researchers who want to take a 'needs first' approach to understanding information behaviour.

7. REFERENCES

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

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

[3] Cool, C. & Belkin, N. 2002. A classification of interactions with information. In H. Bruce (Ed.), *Emerging Frameworks and Methods: CoLIS4: proceedings of the 4th International Conference on Conceptions of Library and Information Science, Seattle, WA, USA, July 21-25, 2002*, (pp. 1-15).

[4] Ellis, D. 1989. A Behavioural Approach to Information Retrieval System Design. *Journal of Documentation*, 45(3).

[5] Ellis, D., Cox, D. & Hall, K. 1993. A Comparison of the Information-seeking Patterns of Researchers in the Physical and Social Sciences. *Journal of Documentation* 49(4).

[6] Ellis, D. & Haugan, M. 1997. Modelling the Information-seeking Patterns of Engineers and Research Scientists in an Industrial Environment. *Journal of Documentation* 53(4), pp. 384-403.

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

[8] Makri, S., Blandford, A. & Cox, A.L. 2008. Investigating the Information-Seeking Behaviour of Academic Lawyers: From Ellis's Model to Design. *Information Processing and Management* 44(2), pp. 613-634.

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

[10] Meho, L. & Tibbo, H. 2003. Modeling the Information-seeking Behavior of Social Scientists: Ellis's Study Revisited. *Journal of the American Society for Information Science and Technology* 54(6), pp. 570-587.

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

[12] 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

[13] Russell-Rose, T., Lamantia, J. and Burrell, M. 2011. A Taxonomy of Enterprise Search and Discovery. *Proceedings of HCIR 2011*, California, USA.

[14] Spencer, D. 2006. Four Modes of Seeking Information and How to Design for Them. Boxes & Arrows: http://www.boxesandarrows.com/view/four_modes_of_seeking_information_and_how_to_design_for_them.

Revisiting User Information Needs in Aggregated Search

Shanu Sushmita
University of California LA
shanusushmita@ucla.edu

Martin Halvey
Glasgow Caledonian University
Martin.Halvey@gcu.ac.uk

Robert Villa
University of Sheffield
r.villa@sheffield.ac.uk

Mounia Lalmas
Yahoo! Labs Barcelona
mounia@acm.org

ABSTRACT

Aggregated search interfaces are a common way to present web search results, mixing different types of results into one single result page. Although numerous efforts have been made to infer users' information needs in "standard" search, we know little about users' information needs within the context of aggregated search. This paper presents the outcomes of a survey of 117 respondents, investigating users' preferences for their type of search result (image, news, video) and their type of information need (informational, navigational and transactional). The survey reveals that users' result preferences differ based on their underlying information needs, suggesting that the taxonomy provided by Broder [1] requires updating to reflect user information needs in the context of aggregated search. For instance, respondents indicated a preference for diverse results (news and reviews about a particular software product) for navigational and transactional queries rather than a single result (the web page to download that software product).

1. INTRODUCTION AND BACKGROUND

Aggregated search is the technique of integrating search results from different verticals (e.g., web, image, video, news) on a single search result page so that users can access the increasingly diverse content available on the web. Aggregated search systems aim to facilitate users' access to "non-standard" web results without having to perform separate searches in the respective verticals, which are source specific sub-collections provided by search engines [13].

Throughout the evolution of web search, users' interaction with search results has been studied by many to improve the quality of the search results and the search experience. Efforts were (and are still being) made to understand users' information seeking process, based upon which several taxonomies describing users' behaviours have been proposed [1, 5, 6, 9, 10, 11, 16].

For instance, in 2002, Broder [1] created a taxonomy of web search, classifying users' information needs into three categories, namely, *informational*, *navigational* and *transactional*. For navigational search, the immediate intent is to reach a particular site (e.g., BBC Homepage); for informational search, the intent is to acquire some information likely to be contained in one or more web pages (e.g., global warming); and finally, for transactional search, the intent is to perform some web-mediated activity (e.g., download, purchase).

Others such as Lindley et al. [16] looked at why people search or go online and identified five main web activities: respite, orienting, opportunistic use, purposeful use and lean-back internet. An example of a respite activity is when people use the web to take a break at work, or through a mobile phone to occupy themselves while waiting. Similarly, Chew et al. [10] explored the contextual and behavioural details of users' interaction with web-based images as they occur in the course of everyday life, showing that users interact with image results as these help creating connections to other people and remote places, or reflecting on the past.

While there is a substantial body of work on understanding users' information needs and browsing activities in "standard" search, far less is known about these within the context of aggregated search. For instance, it is not clear if the existing taxonomies on information needs for "standard" search hold in an aggregated search scenario. In aggregated search, search results may originate from different media (e.g., images, maps) or may be of different genres (e.g., news, blogs). This may have an effect on the way users interact with the results, and affect their preferences for the types of results. A study in [15] investigated the former, but the latter remains largely unexplored. For instance, it is not known whether for navigational queries, users prefer to view a specific website, as would be implied by [1]. A negative answer would mean that a revisit of Broder's three-main-categories of information needs is needed. Also, building an awareness of web activities in aggregated search, which cut across domains, media types and applications, can highlight important details when designing for interactions with the web [16].

The focus of this short paper is, therefore, two-fold: (1) to investigate the preference of search results sought by the users; and (2) to investigate the existing frameworks of web activities within the context of aggregated search. For this purpose, users' preferences for results of several media types and genres are investigated. Furthermore, since Broder's taxonomy has been heavily used (e.g. [3, 7, 9, 15]) we focus on the now classic informational, navigational and transac-

tional categories. We nonetheless aim to extend this work with other taxonomies (e.g., ODP[1]) in future work. This paper makes the following contributions: (1) Investigates users' preference for search results (media and genres) for informational, navigational and transactional search tasks; and (2) Provides empirical evidence to support the need for updating the above three categories within the context of aggregated search.

We present the results of a survey that investigated users' preferences for results of different media types and genres, as answers to informational, navigational and transactional queries.

2. STUDY

A survey containing sixteen questions (4 background questions and 12 search task questions) was distributed on various social networks. The survey allowed us to reach a large and diverse enough number of users, and is a common way to elicit user perceptions and preferences [4, 8]. A total of 117 respondents completed the survey, of which 60 were female and 54 male; the remaining 3 did not disclose their gender. The respondents' age varied between 20-59 years (mean 29). Geographically, respondents were distributed across the US and Canada (3%), Europe (34%), Asia (62%) and Africa (1%). Most respondents were familiar with search engines and used them frequently.

2.1 Task

The aim of the survey was to elicit users' preferences for the types (media, genres) of search results for informational, navigational and transactional search tasks. To this end, we designed four search topics[2] for each of these three categories. The list of topics for each category is listed in Table 1. In total, there were twelve questions for each respondent to answer. The orders of the questions were rotated to minimise ordering bias.

We designed topics that could be understood universally (e.g, global warming, checking emails, buying dvd, software download). Furthermore, the topics were devised to fit the informational, navigational and transactional categories. Therefore, we did not manipulate topics to suit specific media or genre. For instance, for the topic *global warming*, some people may want to read the latest news about global warming, some others may want to view pictures of melting icebergs, while some others may want to watch a documentary on global warming. Therefore this topic does not have an implicit type intent (e.g. image) but requires the gathering of information (informational search task) from many web pages; it is expected that users will look for multiple results to satisfy the corresponding information need. However, it will depend on users which result types (image, news, video, etc) they prefer to view – only news articles, few pictures, or a combination of both.

2.2 Procedure

For each search topic, the respondents were given five choices, namely, web, news, image, video and other results[3]. The respondents were allowed to select as many options as they desired. That is, they were allowed to select just *'one'* or *'all'* options, and therefore were not forced to provide a preference for all the choices listed. This allowed a more natural selection of choices, and hence reduced any design bias. In cases when the respondents selected more than one option, they were asked to rank the choices, by providing "1st", "2nd" .....,"5th" preference for each choice. For instance, if *image, news* and *others* were selected as choices, these had to be ranked in order of preference (e.g., 1st preference - news, 2nd preference – image, 3rd preference – others).

Figure 1, shows the screenshot of an example question with the preference options. Next, the outcomes of the survey are presented.

[1]http://www.dmoz.org/

[2]A search topic describes a search task scenario. The concept of a search task scenario was inspired from [2].

[3]The definitions of these categories were not specified in the instructions and were left open to respondents' interpretation.

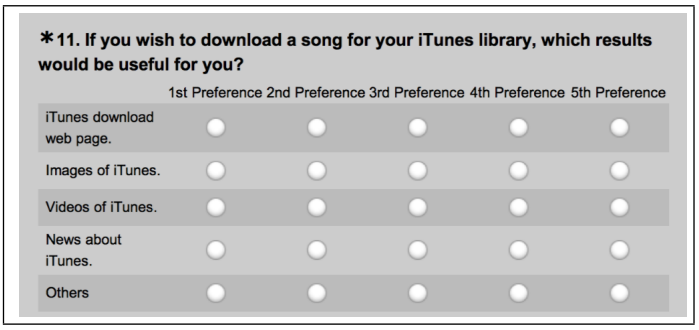


Figure 1: Screenshot showing the preference options provided to the respondents for the selection of search result choices.

Table 2: Median and Interquartile Range for the Preference Rank Score, where Q1and Q3 are 1st and 3rd quartile.

	Navigational	**Informational**	**Transactional**
Result Type	Median (Q1 - Q3)	Median (Q1 - Q3)	Median (Q1 - Q3)
Web	1 (1-1)	1 (1-2)	1 (1-1)
Image	3 (2-4)	3 (2-4)	3 (2-4)
Video	3 (3-4)	2 (1-3)	3 (2-4)
News	2 (2-4)	2 (1-4)	2 (2-4)
Others	4 (2-5)	4 (3-5)	4 (3-5)

3. OUTCOMES

As the data obtained from the survey was non-parametric, we report medians and the interquartile range for the preference scores. The results are reported in Table 2, which shows the median rank of each vertical by information need. Friedman tests were performed to estimate the significance of preference for the results types, among and across the three categories (navigational, informational and transactional). Finally, multiple Wilcoxon-tests were run in the post-hoc analyses while adjusting the p-values using the Bonferroni method. The outcomes from the post-hoc pair wise comparisons for navigational, informational and transactional categories are shown in Tables 3, 4 and 5 respectively. Each row in these tables indicates whether a particular result type was preferred over each of the other result types.

As can be seen in Table 2, most respondents indicated the 'web page' as the most preferred type of results, when

Table 1: List of topics presented to the respondents in the survey. The topics for each category (navigational, informational and transactional) are grouped here, but their order was rotated in the survey to minimise ordering bias.

Navigational Topics
1. When you wish to book tickets with British Airways, which results would be useful for you?
2. When you wish to find an address from yellow pages, which results would be useful for you?
3. When you wish to check courses of a University, which results would be useful for you?
4. When you wish to check your email (e.g, gmail, hotmail, msn, etc), which results would be useful for you?
Informational Topics
5. When you wish to learn about salsa dance, which results would be useful for you?
6. When you wish to gather information about global warming, which results would be useful for you?
7. When you wish to learn on how to make a pancake, which results would be useful for you?
8. When you wish to know about 2011 budget, and how it effected farmers, which results would be useful for you?
Transactional Topics
9. When you wish to download a free software, which results would be useful for you?
10. When you wish to download a song for your iTunes library, which results would be useful for you?
11. When you wish to file a property complaint, which results would be useful for you?
12. When you wish to buy a DVD online, which results would be useful for you?

compared to the other four types (image, video, news and others). The difference was found to be significant for navigational, informational, and transactional cases (rows 1-4 in Tables 3, 4 and 5); thus suggesting that "standard" web results are the prime source of information sought by most users. After web results, news was the second most preferred type of results when compared to image, video and others (6th row in Table 2). For the navigational category, news results were significantly preferred over image, video and others results (rows 6, 8 and 9 in Table 3). However, video was equally preferred to news for informational and transactional categories (row 8 in Tables 4 and 5).

Finally, there is a trend for image and video results to come third in preference from respondents for most categories (4th and 5th rows in Table 2). However, post-hoc analyses suggest a significant difference of preference for video and image over 'other results' for all three categories (rows 7 and 10 in Tables 3, 4 and 5). In addition, video results were significantly preferred to image results for informational and transactional cases (row 5 in Tables 4 and 5), while no significant difference was observed for the navigational case (row 5 in Table 3). Therefore, it is possible that users may prefer image results instead of video results in some cases, and video results in other cases. In addition, image and video being the third preference indicates that providing image and video results for all queries may not be appreciated by users.

In Tables 3 to 5, in only two occasions were the ranking of result types not significantly different: image-video for navigational, and news-video for informational information needs. This indicates that for navigational needs, neither image or video results are judged as important to users, backing up the results in Table 2, where both are ranked bottom. For informational information needs, both news and video were judged equally important to the search tasks, second only to web (Table 2).

Table 3: Results of post-hoc pair wise comparisons for navigational category.

row. no	Pair	Z- Score	p-value
1	Web - Image	-14.09	< 0.0001
2	Web - Video	-13.95	< 0.0001
3	Web - News	-13.62	< 0.0001
4	Web - Others	-13.46	< 0.0001
5	Image - Video	-1.34	0.1814
6	Image - News	5.26	< 0.0001
7	Image - Others	-4.03	< 0.0001
8	News - Video	-7.69	< 0.0001
9	News - Others	-8.38	< 0.0001
10	Video - Others	-3.73	0.0001

4. DISCUSSION

The aim of our study was to investigate, via a survey, users' results preference for navigational, informational, and transactional search topics.

Overall, three key observations can be made from this survey. First, for all query categories, web results continue to be the prime source of information sought by users – 90% for navigational, 54% for informational and 85% for transactional – suggesting that for an aggregated search result page, web results should always be provided. This echoes the findings of [14] where the importance of web results for aggregated result pages was demonstrated through the mining of query logs.

Second, there appears to be a difference between the result preferences for navigational and transactional queries. From Broder [1], the corresponding information needs for these categories were identified to be focused (i.e., specific website, download, etc). In contrast, our study suggests that users also prefer to view other results, and not just one ("to the point") result, or one type of result. More precisely, for the navigational search topics, in addition to web results, respondents also indicated a preference for news and video results. This may be due to the fact that, since an aggregated result page is often provided for most queries by mod-

Table 4: Results of post-hoc pair wise comparisons for informational category.

row no.	Pair	Z- Score	p-value
1	Web - Image	11.94	< 0.0001
2	Web - Video	-7.40	< 0.0001
3	Web - News	-6.62	< 0.0001
4	Web - Others	-13.87	< 0.0001
5	Image - Video	8.55	< 0.0001
6	Image - News	3.96	< 0.0001
7	Image - Others	-9.06	< 0.0001
8	News - Video	0.58	0.5583
9	News - Others	-11.25	< 0.0001
10	Video - Others	-11.80	< 0.0001

Table 5: Results of post-hoc pair wise comparisons for transactional category.

row no.	Pair	Z- Score	p-value
1	Web - Image	-13.40	< 0.0001
2	Web - Video	-12.65	< 0.0001
3	Web - News	-13.17	< 0.0001
4	Web - Others	-13.39	< 0.0001
5	Image - Video	4.64	< 0.0001
6	Image - News	5.33	< 0.0001
7	Image - Others	-4.34	< 0.0001
8	News - Video	-2.30	0.021
9	News - Others	-10.09	< 0.0001
10	Video - Others	-6.77	< 0.0001

ern search engines[4], users are exposed to diverse results and as a consequence, results other than web have now gained prominence. However, whether providing diverse results for informational and transactional information needs facilitates task completion, and/or increases user satisfaction, requires further investigation.

Third, users' preferences for the 'type' of results vary with the query category. For instance, for navigational and transactional search topics, web and news results seem to be preferred. The preference is more mixed for informational search topics, with image results least preferred. In itself, it is not surprising that users' preferences vary with query categories. However, concrete knowledge regarding which 'types' of sought results are preferred would allow for more appropriate aggregation of the different verticals under consideration. Similar investigations were carried out in [12] by Sushmita et al. where, associations between query classifications (e.g., arts, health, etc) and result types were indeed identified. Such knowledge may then be used by search systems, to present particular types of result for different queries, for example, a system may not present (or demote in importance) image results in response to an informational query.

5. CONCLUSION AND FUTURE WORK

We presented the analysis of a survey of 117 respondents' preferences regarding the different types of results for navigational, informational, and transactional information needs. Although small in terms of the number of users and acknowledging the limitation of an online survey, interesting insights emerged from our investigation. The outcomes of the survey support the aggregated search paradigm, showing that users' preferences are for a diverse range of result types. The analysis also indicates a need to revisit the definition of the three categories of information needs [1], within the context of aggregated search. This work initiates two future research questions: (1) What information needs exist within the context of aggregated search? and (2) How to identify suitable results satisfying those information needs?

[4]http://www.slideshare.net/rankabove/com-score-rankabove-final

6. REFERENCES

[1] A. Broder. A taxonomy of web search. *Journal of SIGIR Forum*, 2002.

[2] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *JASIST*, 2003.

[3] L.A. Granka, T. Joachims & G. Gay, Eye-tracking analysis of user behavior in WWW search, *SIGIR*, 2004.

[4] S.A. Grandhi, Q. Jones & S. Karam, Sharing the big apple: a survey study of people, place and locatability, *SIGCHI*, 2005.

[5] M. Kellar, C. Watters & M. Shepherd, A Goal-based Classification of Web Information Tasks, *ASIST* 2006.

[6] H. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang & Y. Li, Detecting online commercial intention, *WWW*, 2006.

[7] B.J. Jansen, D.L. Booth & A. Spink, Determining the user intent of web search engine queries, *WWW*, 2007.

[8] M.R. Morris, A survey of collaborative web search practices, *SIGCHI* , 2008.

[9] B.J. Jansen, D.L. Booth & A. Spink, Determining the informational, navigational, and transactional intent of Web queries, IP&M, 2008.

[10] B. Chew, J.A. Rode and A. Sellen, Understanding the Everyday Use of Images on the Web, *NordiCHI*, 2008

[11] S. Stamou & L. Kozanidis Impact of search results on user queries, *WSDM*, 2009.

[12] S. Sushmita, H. Joho, M. Lalmas & J.M. Jose, Understanding domain "relevance" in web search. *WSSP at WWW*, 2009.

[13] J. Arguello, F. Diaz, J. Callan & J.-F. Crespo, Sources of evidence for vertical selection. *SIGIR*, 2009.

[14] S. Sushmita, B. Piwowarski & M. Lalmas, Dynamics of Domains and Genre Intent, *AIRS*, 2010.

[15] S. Sushmita, H. Joho, M. Lalmas & R. Villa, Factors affecting click through behavior in aggregated interface, *CIKM*, 2010.

[16] S.E. Lindley, S. Meek, A. Sellen & R. Harper, "It's simply integral to what I do" enquiries into how the web is weaved into everyday life, *WWW*, 2012.

Improving search experience on distributed leisure events

Richard Schaller
Computer Science (AI Group)
Uni of Erlangen-Nuremberg
richard.schaller@cs.fau.de

Morgan Harvey
Computer Science (AI Group)
Uni of Erlangen-Nuremberg
morgan.harvey@cs.fau.de

David Elsweiler
I:IMSK
University of Regensburg
david@elsweiler.co.uk

ABSTRACT

This paper examines how simple changes to a search system can influence the user's experience when using the system. In previous work, we evaluated user behaviour with a search tool designed to help people discover events distributed over a city of interest to them personally. We established, contrary to our expectations, that users mostly searched for events they already knew about, made several spelling errors and often achieved poor search performance. Taking these findings as inspiration, we made changes to how the system works. In this paper, we describe and motivate the changes and present a naturalistic log-based study (n=860) to examine the effect on user search behaviour.

1. INTRODUCTION AND MOTIVATION

When studying search behaviour most published work focuses on analysis in the context of work tasks. Such tasks are not necessarily related to work but rather involve people performing a sequence of activities in order to accomplish a goal [5]. A work task has a recognisable start and end, may consist of a series of sub-tasks, and results in a meaningful product [2]. Thus models developed tend to assume that people look for information to close a gap in knowledge [1] which prevents them from completing their current task.

In contrast we want to do search analysis in context of leisure activities where no clear focus on a concrete working task is given. Elsweiler and colleagues [4] proposed a model for what they refer to as casual leisure search, which deviates from standard work-based models. According to their model, in casual-leisure situations users are not focused on accomplishing a task but rather aim to be entertained or to pass time. These needs are influenced by emotional state, physical state or the social context in which they live. Additionally such needs differ from work tasks by weighting the emotions induced by the found content or even the search process itself more than the raw informational content.

Schaller et al. analysed in [7, 8] mobile search behaviour in the context of a distributed event and compared search characteristics and performance of a naturalistic user study to those of mobile web search. A main finding is that the analysed queries were much shorter than those of mobile web search. Also, most queries – in contrast to web search – were for known-items: predominantly events names. Users made a huge amount of spelling mistakes perhaps due to the environmental context (e,g, typing on a bumpy bus) or due to the unfamiliarity with the correct spelling of the known-items. This is probably one of the causes of the poor search performance with over 40% of searches being unsuccessful. A major conclusion of the paper is that people used the system as a tool for filtering and not for searching for new things. The paper gives suggestions for better tailoring the search system towards the observed user behaviour.

In this paper we build upon these results and explore how the proposed changes to the system influenced search characteristics and performance with the overall aim of improving search experience. We first describe the design of a study to answer this question. We report analyses of interaction logs of a variant of the search system including the proposed changes which was tested on a similar event as was used in [8]: The Long Night of Music 2012 opposed to the Long Night of Museums 2011, both located in Munich. We admit possible doubts as to the comparability of these Nights due to the different topics: however we are able to address these doubts by showing that user behaviour is similar enough to make valid comparisons of system changes. We then show differences and similarities in search behaviour between the original and the improved search system and finally draw conclusions as to the effectiveness of the analysed changes.

2. DISTRIBUTED EVENTS

A distributed event is a collection of single events over the same time period and having the same general theme. One such event is the Long Night of Munich Museums[1] (LNMuseums), an annual cultural event organised in the city of Munich[2], which was the context of the study performed in [8]. In addition to a diverse range of small and large museums, other cultural venues, such as the Hofbräuhaus and the botanical garden open their doors for one evening in October. Many venues organise special activities and exhibitions not otherwise available. A similar distributed event is the Long Night of Music (LNMusic) which also takes place in Munich. Aside from pubs, discotheques and clubs also some cultural venues like churches and museums take part which leads to some overlap between both nights regarding the provided events.

Visitors to Long Night include both locals and tourists and represent a broad range of age groups and social backgrounds. In 2011 an estimated 20,000 people visited a total of 176 events at 91 distinct locations at the LNMuseums, including exhibitions, galleries and interactive events. The LNMusic is about the same size with 206 events at 123 locations and approximately 20,000 visitors. Events on both nights take place all over the city, mostly in the city centre, but some, such as the Museum of the MTU Aero Engines

[1]Name in German: Lange Nacht der Münchner Museen
[2]The event is organised by Münchner Kultur GmbH (http://www.muenchner.de/museumsnacht/)

and the Potato Museum, are located in suburbs. Special bus tours are set up to transport visitors between events.

Events can be discovered by means of the booklet that is distributed for free by the organisers and contains descriptions of all events in the order they lie along the bus tours. This booklet is necessarily large (110 A6 pages per Long Night) and can be difficult to navigate.

3. SYSTEM

An Android app was developed in [8] to help visitors of the Long Night find events of interest to them personally. Once they have found and selected the events they would most like to visit, the system can create a time plan for the evening, taking into account constraints such as start and end times of events, time to travel between events and public transport routes and schedules. If the user chooses more events than would fit into the available time then the system tries to maximise the number of scheduled events by leaving out those requiring long travel time. It is also possible for the user to manually customise the plans by adding, removing and re-ordering events to be visited. Based on the created plan, the application can lead the user between chosen events using a map display and textual instructions. Figure 1 provides some screenshots of the app[3] as was used on the LNMuseums and LNMusic.

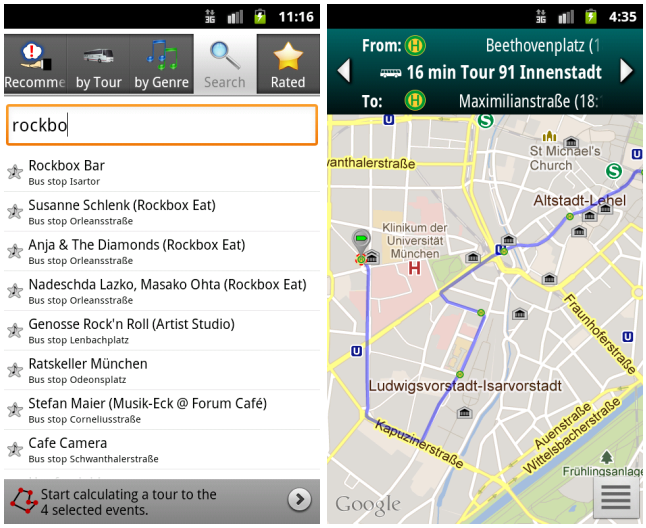


Figure 1: The search screen with a query (left) and the map screen with the planned route (right)

The user has four ways to find events he would like to visit, namely he can: receive recommendations based on a pre-defined profile and collaborative filtering algorithm built into the app; browse events by bus route; browse events by genre or type or submit free-text queries, which search over the names and descriptions of the events.

As described in [8] the search functionality was implemented in Lucene[4] and documents were represented by titles and descriptions from the Long Night booklet. Lucene was extended to perform a search based on topics. Firstly the event descriptions and titles were tokenised and stemmed then to match topically similar words, each token is mapped to one or more topic groups (these groups are taken from [3]). This way terms such as "dinner" and "food" are mapped to the same groups, thus event descriptions containing one of these words could be found by the other. To speed up interaction with the system, queries were submitted after each typed character (search-as-you-type). The presented result list contains the name and nearest bus stop for each of the retrieved events.

[3]a video demo of the application can be found on YouTube (http://www.youtube.com/watch?v=qy1F8fZbowo)
[4]Lucene version 3.1. (http://lucene.apache.org)

4. SEARCH SYSTEM CHANGES

Based on insight gained from user interactions with the original search system, as described in [8], it was determined that the following improvements could be made:

- Grep-Like Search:
  Since users used the system mainly as a filtering tool, the search-as-you-type feature might have led users to give up early: the system tried to match whole (or stemmed) words while the user faces an empty result list after typing in the first few characters but before finishing the word. In our new system – used and evaluated on the LNMusic – we extended the search system with a grep-like feature which would also match parts of words and not just complete words. For example, if a user is looking for the event "Lenbach" it is sufficient to type in just "Lenb". This means that users are not so often presented with an empty list of search results.

- Fuzzy Search:
  It was noticeable from the user interactions that a huge number of spelling errors were being made. This was presumably due to environmental factors, e.g. typing on a bumpy bus or due to the a high number of named entities, the spelling of which people are not familiar. In either case the system was adapted to better support the user by performing a fuzzy search according to [8]. Our system was improved by utilising the Lucene Fuzzy Search mechanism which uses the levenshtein edit distance to match term that differ only by a few characters. If looking for the event "Lenbach" it will be found even if the user by mistake typed "Lembach".

Both changes aim to allow users to more quickly and easily find the events they are interested in and improve the overall search experience.

As the same naturalistic study was used to analyse other parts of the system there were other small changes which are unrelated to the search system analysis: the number of tabs was increased due to the addition of a recommender tab which was beforehand combined with the list of already selected events in one tab. Secondly, the tab position of the recommender tab was tested in an A/B test. Both changes are beyond the scope of this search behaviour paper and should not have any significant influence on it as the layout of the search tab itself wasn't changed.

5. DATA COLLECTION

We examined the search behaviour of users by recording user interactions with our refined app at the LNMusic 2012 in the same manner as with the original app on the LNMuseums2011. Again the app was available for download from Google Play Store and advertised on the official Long Night of Music web page. In total the application

was downloaded approximately 1000 times and 860 users allowed us to record their interaction data (In [8] approx. 500 downloads and 391 users are reported). We recorded all interactions with the application including submitted queries, result click-throughs, all interactions with browsing and recommendation interfaces, tours generated, modifications to tours, as well as all ratings submitted for events. Users interacted on average for 11.79 minutes[5] with the system (median 6.46). 57.3% of users interacted for more than 5; 17.9% for more than 30.

Since queries were submitted after every typed character, it is necessary to pre-process the recorded queries to establish those that the users actually intended to submit. For example, if the user wanted to search for "food", the system logged "f", "fo", "foo", as well as "food". Furthermore, should the user wish to submit a new query, then he must first remove the old search terms from the search box, resulting again in all prefixes but this time in decreasing length.

As in [8] we manually judged queries to be intended or not. 3 assessors separately annotated all of the 12,500 queries logged as being either intended or not-intended. A very high inter-assessor agreement was found (Fleiss' kappa = 0.915, 89.8% of queries which were labeled by at least 1 assessor were also labelled by at least one other). This process resulted in a final list of 1,434 search queries, which is used in the following analyses and compared against the results reported in [8] which are based on 801 search queries.

6. IS A COMPARISON OF NIGHTS FAIR?

Undoubtedly the best way of comparing two version of a system is to run experiments under the same external conditions. Unfortunately this was not possible with our app for the LNMusic as we work together with the organisers to provide a real system for "productive" use and cannot experiment with arbitrary system variations. The alternative would be a lab study, however we consider this to be a less preferable option given that we wish to record interaction with the system in a real-world (i.e. non-simulated) setting. Therefore we looked into whether data obtained from different events could be fairly compared.

We learned from our experience with past Long Nights that user behaviour is to a huge extent independent from the actual type of Long Night. In [7] interviews with visitors of two different Long Nights revealed that beside the topic of an event other characteristics – such as novelty, the time and location of the event or the possibility to take part in the event – play a crucial role. It is precisely this that sets a system for casual leisure activities apart from a system for solving a work task [4].

In this section we want to give more insights into why our study on the LNMusic2012 and the study presented in [8] on the LNMuseums2011 are comparable. First of all both distributed events – although their topics are different – have a lot in common: They take place in the same city, are organised by the same company and hence have the same ads, booklet format, price tag and even the special bus routes provided are partially the same. We noticed that some of the events on the LNMusic were also available on the LNMuseums. We investigated further into how many events are in common between both nights. To do so we looked into how many LNMuseums events had, at the same location, an event on the LNMusic that were organised by the same museum, church, bar, etc.. Surprisingly 21.6% of the LNMuseums2011 events had a matching event on the LNMusic2012. The topic of the matched events might differ but as mentioned above the topic is only one of many relevant aspects for visitors to choose an event.

Secondly we looked into the app usage itself. Upon first start-up of the app we ask our users to fill out a short questionnaire; among others we ask for the age of the user (below 18, 18 to 29, 30 to 39, 40 to 49, 50 to 59 and above 60 years old). Answering of these questions was optional but 246 on the LNMuseums2011 and 495 users on the LNMusic2012 chose to do so. Based on these data (omitting the first and last age groups due to the small sample size) we compared the age distribution of users on both nights with a χ^2-Test revealing a χ^2 of 1.459 and $p = 0.6918$. This result states that this is no significant difference between app users of both nights.

To ascertain if the two studies are comparable it is important that system usage is similar on both nights. As described in Section 3 there were multiple ways (different tabs) of accessing the events. We looked into which of these tabs users were interested in. We therefore define a tab session to start when a user switches to a tab and to end when he switches to another tab. Table 1 shows the number of tab sessions. Based on this data, a χ^2-Test shows a χ^2 of 5.387 and $p = 0.1456$, again indicating no significant difference between users of both nights.

	by Tour	by Genre	Search	Rec.+Rated
LNMuseums	28.0%	14.5%	15.4%	42.0%
LNMusic	26.6%	15.6%	15.1%	42.7%

Table 1: Number of tab sessions

Lastly we considered properties of the search behaviour itself that should be invariant to our changes to the search system. One of the main findings of [8] was the huge number of named-entity searches and we compared the reported numbers to those of our own system. Using the same method described in [8] we instructed 3 human accessors to label all search queries into one of three categories: specific event name, not a specific event name or indeterminate. For 82.0% of all queries at least two of the assessors were able to agree on one of the three categories (Fleiss Kappa of 0.32). 84.5% (LNMuseums2011: 59.4%) of the agreed on queries were marked as clearly named entities and 8.2% (34.6%) that might be named entities. Only 7.2% (6.0%) were labeled as non named-entity searches. It is notable that the low number of non named entity searches is similar to what is described in [8].

In [7] it is reported that the same system used on the LNMuseums2011 was also evaluated on the Long Night of Science in Erlangen-Nuremberg (LNScience2011), which also is a distributed event but dedicated to science. We looked into how search characteristics differ if the same system is used on different nights. We compared query length with respect to the number of characters and the number of terms per query by performing a (non-parametric) Kruskal-Wallis Rank Sum Test. No significant difference between the usage

[5]These figures were calculated by summing the time periods for which a user was active, discounting times where the system reported no interactions for more than 15 seconds. We further discounted any interaction sequence that contains gaps of non-interaction longer than 30 minutes as these are likely due to logging problems caused by running out of power, connection problems, app crashes, etc.

on the LNMuseums2011 and the LNScience2011 could be found (for characters: $p = 0.1169$; for terms: $p = 0.6039$). When performing the same test between queries on the LNMuseums2011 and the LNMusic2011 a highly significant difference can be found with $p \ll 0.01$ for both characters and terms. Thus changes to the search system have an influence on search behaviour but changes to the overall setting of the distributed event have not.

In conclusion we believe a comparison of the app user behaviour on both nights is appropriate, given the circumstances and difficulty of obtaining real-world user data of such apps.

7. INFLUENCES OF THE CHANGES

In [8] many statistics on query characteristics and query performance are given based on analysis of search logs, a common technique in the literature [6]. In this section we recalculate these statistics based on the data logged with the new system and compare behaviour with both app variants to determine what user behaviour changes the search system modifications caused. To do so we consider a number of different indicators of an improved search experience.

The average length of a search query on the LNMusic2012 was 5.6 characters ($\sigma = 3.36$) and 1.14 terms ($\sigma = 0.41$). This is much shorter than what is reported [8] for the LNMuseums2011: 8.9 characters ($\sigma = 5.31$) and 1.21 terms ($\sigma = 0.52$). A Z-test performed between both nights reveals that this difference is highly significant for both metrics (i.e. $p \ll 0.01$ in both cases). It seems that the grep-like searching – which matches also partial words – has influenced people to stop typing much earlier. We assume that there are two main causes that users stopped typing: either they have found what they were looking for (successful search) or they gave up on the search because they couldn't find what they were looking for (unsuccessful search). Users of our system had three options to interact with the entries in the result list: they could view details of an event, mark an event as a candidate for tour inclusion or add the event to an pre-existing tour. We consider any of the three as an indicator for search success and the lack thereof as an indicator of an unsuccessful search. Good abandonment wasn't considered since the result list contains no information beyond the event name and nearest bus stop. The length of successful queries on the LNMusic2012 was 5.39 characters ($\sigma = 3.27$) and 1.12 terms ($\sigma = 0.38$) which is highly significantly ($p \ll 0.01$) shorter than reported in [8]: 9.90 characters ($\sigma = 5.42$) and 1.26 terms ($\sigma = 0.57$). This means users have to type on average 45% less to find the events they are interested in. On the other hand the query length of unsuccessful queries was slightly reduced with 6.39 characters ($\sigma = 3.56$) opposed to 7.47 characters ($\sigma = 4.80$) but slightly longer with regard to terms: 1.20 ($\sigma = 0.49$) as opposed to 1.13 ($\sigma = 0.42$).

Of the 1,434 queries entered on the LNMusic 76.7% resulted in an interaction of the user with an event, meaning they were successful. 23.3% were unsuccessful, a much better conversion rate compared to the 40.3% unsuccessful searches in [8]. This decrease is highly significant ($p \ll 0.01$) and demonstrates that the improved search system was able to assist users in finding events they were looking for.

In [8] a large ratio of 59.75% of unsuccessful search queries had an empty result list. With the improved search system this was only the case in 12.57% of unsuccessful queries. But how successful were those "added" entries? We looked at all 2,157 interactions with events (viewing, rating, selecting) on the result list of the LNMusic2012 app (created by the improved search system). We then ran the corresponding search queries through the old search system and counted whether these events would be on the result list had the original search system been used. Only 938 event interactions would be possible with the previous search system, meaning that 56.5% of the interactions performed by users wouldn't have been possible, simply because the events wouldn't be in the results.

This analysis has revealed indicators of an improved search experience which means that the changes proposed in [8] are useful in the context of distributed events assistance.

8. DISCUSSION AND CONCLUSIONS

In this paper we analysed the changes in query behaviour of users due to modifications of a search system used on distributed events. We first describe two studies performed during two such events, including a description of the system used and what changes to the search system were tested. As the LNMuseums and LNMusic have different topics we then showed that a comparison of user behaviour between both nights is sensible and worthwhile. With this preparatory work we then analysed users' search behaviour by comparing search characteristics and search performance. Overall, users typed much shorter search queries, especially in the case of a successful search. Also comparing query performance revealed a much higher success rate with the ratio of unsuccessful searches being almost halved. Finally we presented a comparison of both search systems running on the same search queries which showed that only half of the interactions with events would have been possible with the old system.

The search system as it is now is designed for users to find events they already know of in advance. But how can users be assisted in finding events that are new to them? How can we better support the discovery of serendipitous events? Since the users seldom used the search system for that purpose, a second tool like a recommender is necessary. The user could then decide if he wants to look for a concrete event he already knows of or if he would rather be inspired by the system. If such a split into two "orthogonal" tools is understood and accepted by users then it is worth investigating and would point the way to vastly better distributed events assistance systems.

Acknowledgments This work was supported by the Embedded Systems Initiative (http://www.esi-anwendungszentrum.de).

9. REFERENCES

[1] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[2] K. Byström. *Task complexity, information types and information sources. Examination of relationships.* PhD thesis, University of Tampere, Dep. of Inf. Studies, 1999.

[3] F. Dornseiff. *Der deutsche Wortschatz nach Sachgruppen.* DeGruyter, Berlin, New York, 2004.

[4] D. Elsweiler, M. L. Wilson, and B. Kirkegaard Lunn. *New Directions in Information Behaviour*, chapter Understanding Casual-leisure Information Behaviour. Emerald Pub., 2011.

[5] P. Hansen. User interface design for IR interaction. a task-oriented approach. In *CoLIS 3*, pages 191–205, 1999.

[6] B. J. Jansen and A. Spink How are we searching the world wide web?: a comparison of nine search engine transaction logs In *IPM*, (1,42) pp. 248–26, 2006.

[7] R. Schaller, M. Harvey, and D. Elsweiler. Entertainment on the go: Finding things to do and see while visiting distributed events. In *Proceedings of IIiX*, 2012.

[8] R. Schaller, M. Harvey, and D. Elsweiler. Out and about on museums night: Investigating mobile search behaviour for leisure events. In *Proc. of Searching4Fun Wksp, ECIR*, 2012.

Opinion Mapping: Information Visualization Approaches for Comparative Sentiment Analysis

William H.Hsu
bhsu@ksu.edu
Kansas State University
+1 785 236 8247

Praveen Koduru
praveen@iqgateway.com
iQGateway, Inc.

Chengxiang Zhai
czhai@cs.uiuc.edu
University of Illinois
at Urbana-Champaign

ABSTRACT

In this position paper, we discuss the problem of extracting information about chronic diseases from the large volume of text written in health blogs, mailing lists, forums, and other electronic venues, then making this information accessible via structured queries, while analyzing it to map out patterns among the opinions and demographics of users. Information retrieval systems exist for spatially-referenced demographic data about diseases such as diabetes and their therapies, but as in the case of communicable diseases, the databases that contain such data are manually populated. For example, in information portals such as *HealthMap.org*, which are searchable by location and disease, the data are user-reported and collaboratively maintained, but not automatically extracted from text. Furthermore, there is as yet no automated means of relating sentiments expressed by users in their text postings to their semistructured profile data. This is because the primary sources for this kind of information have been statistical surveys such as opinion polls, where text responses are often human-interpreted and demographic analysis is done *post hoc*, rather than as part of an information retrieval and extraction task. These limitations indicate a present need for text summarization techniques that integrate quantitative information extraction – which captures symptoms, diseases, and complications of diseases – with opinion summarization.

Categories and Subject Descriptors

H.3.3 **[Information Storage and Retrieval]**: Information search and retrieval – *clustering, relevance feedback, selection process*; H.2.8 [**Database Management**]: Database applications – *data mining, spatial databases and GIS*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

sentiment analysis, social networks, geoinformatics, opinion mining, subjectivity, information extraction, information visualization, human-computer interaction

1. INTRODUCTION

In this paper, we address the problem of information retrieval and information extraction in subjective domains, with applications to visualization of opinions – specifically, thematic mapping of opinions. At present, there is a dearth of methods for integrating user profile data for social networks with blog posts, tweets, and other content from the associated social media. These limitations present an integrative challenge for human-computer interaction (HCI) and information retrieval (IR). Towards this end, the specific aims of the research proposed espoused in this position paper are as follows:

1. **Aim 1.** Extend known **algorithms for named entity recognition and relationship extraction**, to produce basic summaries of diseases and treatments mentioned in texts. The technical objective is to tag where basic entities and opinions are mentioned in freely available text (including both user posts and profiles), then map these tagged elements in space, time, and by topic, to acceptable levels of precision and recall.
2. **Aim 2.** Adapt basic known techniques to the domain of type 2 diabetes – specifically, extracting data from text discussions of diabetes that are archived from health blogs and forums using web crawlers. [1] This entails developing a means of handling entities and quantitative data that have not previously been extracted from text, such as information concerning insulin and oral anti-diabetic drug dosage, HbA_{1c} levels, *etc.* Another functional requirement is some mechanism for entity reference resolution, *e.g.*, abbreviations and synonyms, for known terms. Finally, a **domain-specific ontology** of relevant symptoms, disease attributes, complications, and treatments is proposed. For type 2 diabetes, this includes topics frequently discussed in health blogs and forums: food groups, meal plans, nutritional constraints, and conditions such as obesity that are linked to diabetes. This shall facilitate information retrieval applications such as question answering about meal plans recommended by primary care physicians and specialists.
3. **Aim 3.** Develop methods for sentiment analysis and improve existing ones, to **summarize opinions and discover patterns**. The technical objective is to relate demographic data extracted from text and profiles to qualitative data – namely, the polarity of text at the document, sentence, or aspect level, aggregated across demographic categories such as geographic region of residence. Objects of interest for sentiment analysis include prescribed therapies and specifically side effects, but can extend to disease aspects and complications.

The **overall goal** of this approach is to develop an integrative technology for summarizing online text about chronic diseases, capturing opinions from users' posts and demographic data from a combination of their posts and profiles, and finally using these to discover global patterns indicated by the set of all text documents. The **central hypothesis** of this work is that a combination of entity and relationship extraction, driven by a domain-specific ontology of terms, will result in more precise and accurate summarization of opinions. This will increase the

usefulness of free-form text, written by users of social media, in understanding patterns that are reflected in the opinions and demographics of chronic disease patients.

2. BACKGROUND

2.1 Information Extraction from Health Blogs

The chief potential impact of the research framework and test bed proposed in Section 1 is to provide assistive technologies to public health analysts and health services analysts who are using blogs, microblogs (*e.g., Twitter*), and other social media to explore user opinions about chronic disease issues. As an example, in the application domain of type 2 diabetes, these include dietary treatments such as carbohydrate control, complications such as gastroparesis induced by diabetes that may pose digestive constraints, and recommendations of primary care physicians, therapists, endocrinologists, nutritionists, *etc*.

The availability of mailing lists, blogs, wikis, and other electronic media for content management and dissemination has resulted in rapid growth in the volume of online text data containing voluntarily expressed public opinions about health issues. While general-purpose metadata tools exist for annotating this text, the opinions themselves remain a largely unexplored source of information about how chronic diseases affect populations. Meanwhile, the task of relating content from these various self-publishing media to semi-structured profile data from their users has not yet been effectively automated.

We advocate development of application test beds and experimental systems aimed at improving techniques for information extraction, ontology development and mapping, and text mining to identify opinion patterns. The potential progress in these areas is due in part to the approach of combining information extraction to discover disease mentions with sentiment analysis to establish opinions, and in part to the application of this approach to a new source of data: free-form text describing user demographics, attributes of the chronic disease of interest and its related entities, and opinions and semi-structured profile data.

To help public health researchers tap into these freely available but unexplored sources of opinions, we propose to develop information extraction (IE) and summarization methods geared at health blog postings and similar text. Such postings contain not only opinions, and attribution information that can be used to link them to the users who expressed them, but also factual data about the posters and their opinions. This data can help place opinions in a comparative context [2] with population statistics, such as the reporting frequency of symptoms, side effects, and complications.

2.2 Ontology Development

The research approach centers around using information extraction to obtain structured data in the form of records about chronic disease references in text, which are then linked to users via relational data extracted from their profiles. [3] However, the body of relevant concepts in the healthcare domain and in the clinical domain theory of each chronic disease is much broader. Currently there exist pre-clinical (genomic and proteomic) and clinical translational ontologies [4] that contain information relevant to diabetes, but they do not provide the requisite concepts for mining free-form text written by lay users who are discussing diabetes online. We propose to develop an ontology for text mining in diabetes, and the mappings from extracted entities and relationships into this ontology.

2.3 Opinion Mining (Sentiment Analysis)

This aspect of the proposed work focuses on a basic research problem: sentiment analysis from text, also known as *opinion mining*, whose objective is to determine from analysis of a written document what the author's attitude towards an identifiable topic is. This attitude can be subjective or objective; it can be identified as an evaluation (positive or negative), a declaration of the author's emotional attitude, or a expression intended to evoke an emotional response in the reader. Subjects of interest include chronic diseases, their features or aspects including symptoms, complications, and treatments, and related health services.

2.4 Current State of the Field

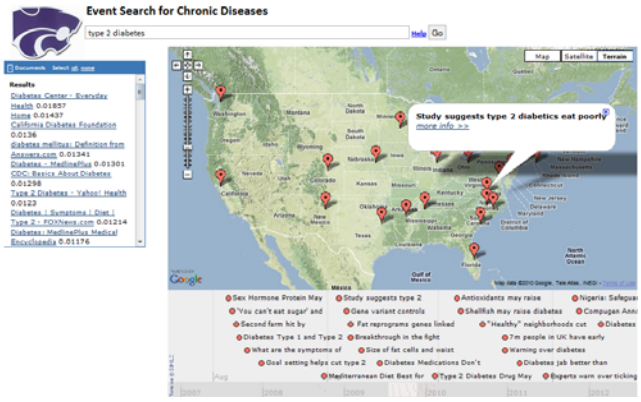


Figure 1. Prototype event search based on a previous IR system for veterinary epidemiology.

Figure 1 depicts a simple search interface for an existing IR system developed by the principal investigator's research group. This system was designed for event extraction in the domain of viral zoonoses, but uses general-purpose software for web crawling and ranking (the latter is developed using *Lucene Java*). One marker is displayed on both the thematic map and the timeline for each returned page, but the only features extracted by this system are the disease name, formatted dates and times given in each article, and locations mentioned in the article.

The thematic map suggests several interactive functions related to opinion mining. One is content-based filtering of articles using the first type of thematic data, demographic and biostatistical attributes; another, collaborative filtering using the second type, polarity scores. Both of these use associations that can be learned from data: a user can search for articles by entering queries that express certain sentiments. In the first case, entities and attributes (*e.g.,* symptoms, complications, and treatments) mentioned in the query may match frequent patterns in the data; in the second, polarity scores themselves can be used to retrieve their "nearest neighbors in opinion space".

3. EXAMPLE: HEALTH BLOGS

The primary value added in adapting the IR and IE workflows described above is an increased capability to explore patterns and trends expressed by an entire collection of health blog posts. As a running example, consider the public health analyst who is interested in charting trends in the use of fast-acting insulin by diabetics. Users often share information about the brands of insulin they use and post opinions about their effectiveness. The following is a post archived on *diabetesforums.com* which is marked up (with a color coding to distinguish entity types and relationship types):

Since I have found out in a previous thread I posted I can use most pen needles with the Novolin 4 pen I got(Haven't used it yet since I have only Humalog for rapids so far), I am here with another question.

I have only used Humalog for a rapid... Does anyone have any insight as to how it compares to Novolog?

In this post, the user is requesting information comparing two drug products (brands of fast-acting insulin, referring to specific delivery mechanisms (pen syringes), eliciting opinions from fellow users, and specifying a requested comparison between named products. Opinions voiced by respondents to this post then discuss how heat-tolerant each brand is, how quickly it acts, and other aspects we refer to as *facets*. [5] Achievement of our primary aims will allow analysts to chart reported biostatistics and opinions, not only about products but about trends, such as the number of units of postprandial insulin taken per gram of CHO.

4. PROPOSED DIRECTIONS

4.1 Mining Social Media (Blogs, Lists, Wikis)

As mentioned above in Sections 1 and 2.1, our approach applies text mining to blogs and social media, a new source of information that is beginning to be studied for opinion and trending topic data, but has not been analyzed for disease-related information that can be related to these data. The novelty of our approach is that it extends named entity recognition and relationship extraction to the domain of understanding free-form text about aspects of chronic diseases (specifically, opinions about type 2 diabetes, its complications, dietary recommendations, and drug treatments). It further develops methods for mapping these new entities and relationships to the terms of an ontology for text mining, and finally leverages the text contained in many online sources to produce integrative summaries of disease mentions and associated opinions.

4.2 New Theory and Methodology

4.2.1 IR (Search Query-Driven) Workflow

In IR applications of automatic text summarization, a user enters a free-form search query and views returned hits that are summarized by topic and aspect – in this case, author opinion. These hits may be organized by space and time. For example, consider the case of a clinical health services analyst, public health analyst, doctor, patient, or other concerned individual who is interested in some aspect of a chronic disease. Such a user typically enters a query into a general-purpose search engine and is either directed to a domain-specialized web portal, also called a *vertical portal*, or browses through documents housed in one.

We seek to advance the state of the field by supporting structured queries, in which a user specifies fields and constraints in addition to traditional search keywords. This is achieved by combining quantitative text summarization (extraction of attribute values) with recognition of entities and relationships. The collection of documents may include some that are dynamically crawled from the web in response to the query. A mixture of labeled and unlabeled data is used to train a semi-supervised topic model. [6] The output consists of structured tuples that are ranked by relevance to the query, filtered to remove hits deemed insufficiently relevant, and finally visualized in a map or timeline view. This view allows the user to more freely explore information by performing interactive manipulations such as online analytical processing or editing the set of constraints.

4.2.2 IE and Summarization (Push) Workflow

IE applications of the proposed summarization technology can be viewed as a more passive variant of the IR application described above, from the user's point of view. [7] No initial query is supplied by the user, but there is an implicit domain of interest from which records should be displayed, corresponding to a combined set of search terms and relevance criteria. When a small set of search terms is known, the IE application can be formalized as a general case of the IR application where "every possible query" is enumerated, multiple crawls are conducted in advance, and the union of all resulting hits is ranked and filtered.

4.2.3 Improved Access through Structured Queries and Opinion Pattern Mining

This workflow is designed to provide analysts with better access to spatiotemporal data. First, it supports approximate range queries, such as: "return records of persons with fasting blood glucose levels close to the non-diabetic range of < 126 mg/dL". Second, it uses measures of semantic relatedness or similarity, *e.g.,* "return posts about adverse effects of Metformin whose expressed sentiments are closest to those in this post". Third, it extracts information in Steps 1 – 3 that in Step 4 can be used to generate *thematic maps*, which portray specific aspects of a geographic region. In this research, the themes fall into two categories: the first, demographic attributes and biostatistics specified by the ontology – some disease-independent, and some disease-specific; the second, quantized measures of opinion polarity (*i.e.*, degree of positive or negative sentiment). The increased support for flexible queries and thematic map generation, compared to IR without relationship extraction and sentiment analysis, will help reveal patterns in the data through interactive investigation.

5. TECHNICAL FOCUS AREAS

5.1 Improvements and Refinements to Theory

The following generic methods are applied in order to meet the functional requirements presented in the preceding section. We refer to them as cross-cutting because they are used in service to all of the technical aims: entity and relationship extraction, ontology development, and sentiment analysis.

5.1.1 Focused Crawling

In previous work on IE, applied to news summarization in the domain of veterinary epidemiology, we used a combination of topical and focused crawling. *Topical crawling* prioritizes pages to be crawled based on user-provided terms (*i.e.*, topics) and seeds (*i.e.,* links to initial pages), while *focused crawling* uses both terms and pages labeled as positive or negative examples of relevant documents. Once tag-formatted web documents (HTML or XML) are crawled, text must be extracted from them.

The on-demand IR system described above functions by passing the user query to a built-in web crawler that fetches hits from a commercial search engine (in this case, *Yahoo*). The results are combined with previously crawled documents, if any, and ranked and indexed as a whole.

5.1.2 Information Extraction

The state of the field in IE for web articles describing disease consists of: payload extraction (of text from HTML), baseline named entity recognition (NER), and extraction of dates, times, and locations in order to localize putative events. In addition to general open natural language processing problems such as co-reference resolution (in particular, pronouns and other anaphora),

word sense disambiguation, and canonicalization of dates, other IE problems that remain unsolved include: resolving alternative abbreviations and synonyms for diseases, disambiguation of place names, associating quantities of persons affected with diseases mentioned, and deduplication of reports. The foundation of our proposed work consists of tasks known to be feasible, but for which general-purpose solutions are still being manually adapted to new domains in current practice: automated named entity recognition and topic categorization. Typically, information extraction is restricted to named entities (Person, Organization, Location, and in our domain, Disease), but attributes such as "causative agent" are not always extracted. Neither are dates, times, quantities, and place names that support the extraction of full tuples of a relationship set. This open problem is of critical significance and is therefore the first of our specific aims.

5.1.3 Web 2.0

The term *Web 2.0* describes an eclectic set of technologies for online interoperability and collaboration. While it includes search, hyperlinking, collaborative authorship and tagging, web services, and syndication, our IE approach focuses on the **authorship**, **tagging**, and **syndication** aspects. Collaborative authorship and editing are mainstays of specialized wikis, but many forums also provide tools for collaboration, from discussion threading and editing history to user profiles, our main source of demographic information besides posts. We will crawl or aggregate profile data, which in some social network and blogging systems (*e.g.*, *LiveJournal*) is published as a publicly available feed. [8] Another source of relational data is the link structure expressed by collaborative tagging, especially annotation by other users *cf. Wikipedia*, social bookmarking *cf. Delicious*, social citation *cf. CiteULike*, and collaborative recommendation *cf. Digg*, *Reddit*, and *StumbleUpon*. We intend to make use of available content management functionality in health wikis and electronic groups. [9] Syndication provides a modern mechanism for refreshing content that is generally more efficient than periodic crawls. We will make use of these three categories of Web 2.0 features and other available content management functionality to assist in the extraction of relational tuples from free text writings online, and in their validation and ranking.

5.1.4 Map and Timeline Visualization

Finally, the generation of views as shown in Figure 1 is a key application of our other primary aims: to build a domain ontology for text mining in diabetes blogs and develop automated mappings from entity recognition systems to this ontology; and to extract the objects and polarity of opinions.

Thematic maps, including opinion maps, help reveal **global patterns and trends** that may have been previously hidden. By visualizing the attributes and related entities of a disease and depicting their variation across space and time, they allow the user to interactively discover these trends. Most previous approaches to construction of thematic maps have been based on electronic medical records and reports compiled by medical providers or observers, such as individual incident reporters for *HealthMap*. [2] The value added by IE operations that automatically populate databases and thematic maps is that they can be applied to the large volume of text that is voluntarily submitted on a daily basis to venues listed at the beginning of this section.

6. ACKNOWLEDGMENTS

Thanks to Tim Weninger, Svitlana Volkova, Surya Teja Kallumadi, Wesam Elshamy, and Andrew Berggren for development work on an early prototype of the information extraction system.

7. REFERENCES

[1] Jiang, J., & Zhai, C. Instance Weighting for Domain Adaptation in NLP. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), (pp. 264-271).

[2] Kim, H. D., & Zhai, C. (2009). Generating Comparative Summaries of Contradictory Opinions in Text. Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM 2009), (pp. 385-394).

[3] Jiang, J., & Zhai, C. (2006). Exploiting domain structure for named entity recognition. Proceedings of the Human Language Technology Conference/the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006), (pp. 74-81).

[4] Craven, M., & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (pp. 77-86). Menlo Park, CA, USA: AAAI Press.

[5] Ling, X., Mei, Q., Zhai, C., & Schatz, B. R. (2008). Mining multi-faceted overviews of arbitrary topics in a text collection. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), (pp. 497-505).

[6] Yue, L., & Zhai, C. (2008). Opinion Integration Through Semi-supervised Topic Modeling. Proceedings of the 17th International World Wide Web Conference (WWW 2008).

[7] Yangarber, R., Steinberger, R., Best, C., von Etter, P., Fuart, F., & Horby, D. (2007). Combining Information Retrieval and Information Extraction for Medical Intelligence. NATO Advanced Study Institute on Mining Massive Data Sets for Security.

[8] Aljandal, W., Hsu, W. H., Bahirwani, V., & Caragea, D. (2009). Ontology-Aware Classification and Association Rule Mining for Interest and Link prediction in Social Networks. Proceedings of the AAAI 2009 Spring Symposium on the Social Semantic Web. Menlo Park, CA, USA: AAAI Press.

[9] Brownstein, J., & Feifeld, C. (2007). HealthMap – Global Disease Alert Mapping System. Retrieved January 25, 2010, from http://www.healthmap.org.

Search System Functions for Supporting Search Modes

Thomas Beckers
Information Engineering
University of Duisburg-Essen
Duisburg, Germany
thomas.beckers@uni-due.de

Norbert Fuhr
Information Engineering
University of Duisburg-Essen
Duisburg, Germany
norbert.fuhr@uni-due.de

ABSTRACT

Tasks in web search are often rather simple, e.g. navigating to an already known web page or looking up a fact. However, tasks in other domains are usually more complex and diverse. Thus, we discuss various search modes of tasks and how they might be supported by functions of a search system. We give examples of the required search functions of different search modes and describe the implications for the design of search systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: General

General Terms

Human Factors

Keywords

system functions, search modes, user interfaces

1. INTRODUCTION AND RELATED WORK

While tasks in web search are often rather simple [4] (e.g. navigating to an already known web page or looking up a fact), tasks in other domains (e.g. searches for scientific literature or patents) are usually more complex and diverse. A set of search system functions that is well-suited for these simple tasks is not appropriate for other more complex task types. In our opinion, each type of task requires a different set of search system functions. Thus, we argue that a "one size fits all" approach (that is, using a search systems with functions e.g. optimized for web search for different tasks in other domains) does not allow the user to search effectively and efficiently. We propose a model of search functions that allows mapping of search activities (search tasks) to necessary system functions comprising the entire search activity.

Hughes-Morgan and Wilson [7] have examined whether improvements of an interactive search system are due to the newly introduced meta-data or to new search functionality. They conclude that users can benefit from improved search features while still using the same meta-data.

Russel-Rose et al. developed a taxonomy for enterprise search and site search by analyzing real-world scenarios [11, 12, 10] based on three top-level categories of search activities originally proposed by Marchionini [8]:

Lookup *a*) Locate *b*) Verify *c*) Monitor

Learn *a*) Compare *b*) Comprehend *c*) Explore

Investigate *a*) Analyze *b*) Evaluate *c*) Synthesize

These categories are orthogonal to each other. Russel-Rose et al. [11] introduce the notion of *search modes*. A search mode is a concrete value of a search activity category. Search modes can be combined to longer sequences or networks. For enterprise search *Locate* is far less common then *Analyze* and *Evaluate*. In the domain of site search the emphasis is on *Locate* and *Explore*.

In the remainder of this paper, we first describe the functional level of search systems. We show how search functions can be mapped to different search modes by giving examples to illustrate how search systems can support each mode and its associated search functions. Subsequently, we describe the implications for designing and developing search systems. Finally, we give an outlook on future work and a conclusion.

2. SEARCH SYSTEM FUNCTIONS

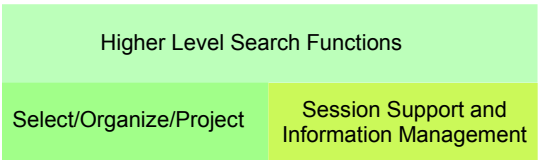


Figure 1: Functional level of IR systems

We divide the functionality of an IR system into three different groups depicted in Fig. 1: *i*) Select/Organize/Project (SOP) *ii*) Session Support and Information Management (SSIM) and *iii*) Higher Level Search Functions (HLSF). In our notion a *search function* is a functionality of the system with which the user can interact or that is fixed by the system designer. A more detailed explanation of the latter two groups and an overall architectural view is given by

The 2nd European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR), Nijmegen, The Netherlands

Beckers and Fuhr [3]. We will concentrate on SOP and (to a lesser degree) on HLSF in the following. In doing so, we will focus on system functions and not discuss their concrete visualizations in the user interface.

Select functions

Select (S) comprises functions for selecting (searching) possibly relevant items.

Ranking method Retrieval functions/ranking methods may be more precision- or more recall-oriented, or they may consider different sources of additional information (like e.g. page-rank). Mutschke et al. [9] showed that search in scientific literature can be improved by considering information about the author, the publication venue or related terms from a thesaurus.

Ranking principle The final ranking might regard each document in isolation, or consider all items above the current one in the output ranking (like e.g. in diversity ranking).

Query language The query structure can be very simple (e.g. a list of terms) or more powerful and expressive, e.g. by supporting simple (boolean) and more complex (wildcards, word distances, etc.) query operators as well as fields and data types.

Formal filter conditions The result set can be filtered by some formal criteria (e.g. by data type, source, date) which is usually done without affecting the RSV.

Query formulation Queries can be formulated a priori as in most systems but also by referring to one or more given items (e.g. query by example, similarity search).

Organize functions

Organize (O) functions deal with the way how the set of result items is structured and organized logically.

Sorting The results can be sorted according to one or more criteria. When searching the best offer for a new smartphone the items may be sorted by price and the trustworthiness or customer ratings. While sorting usually is a one-dimensional organization, also two- or three-dimensional organizations may be helpful, provided that appropriate visualizations are available in the user interface.

Grouping The results can be grouped according to a simple criterion (e.g. grouping by release date, author, source) or according to several facets, as in faceted search [13].

Clustering While grouping is based on some formal criteria, clustering focuses on content aspects based on a some sort of similarity [5]. Although users might have problems interpreting the cluster structure, they might also gain new insights about the result set.

Linking In case there are (explicit or implicit) links between the answer items, the resulting tree or graph structure might be of interest (e.g. Web links, co-author relationships in scientific literature, or friendship connections in social networks).

Project functions

Project (P) comprises functions for the construction of the surrogates to be presented in the results.

Selection Surrogates consists of specific fields of the result items (like e.g. title, author and year in literature search).

Summarization Either unbiased or query-biased summaries (extracts) of the answer documents (or specific fields thereof) can be generated.

Aggregation This function generates a single entry representing several items differing in formal aspects (e.g. mirrors of a web page, various editions of a book) or content (e.g. different reviews of a book in an online store).

Faceting For displaying facets with their existing values and corresponding frequencies, the system must support projection on single facets along with counting values. From the point of view of a relational database, if `F` denotes a facet/attribute, then the system has to process the SQL query "`select F, count(*) from R where ... group by F`" for each facet. Query conditions and restrictions wrt. to the values of a facet then affect the `where` part of the query.

Enrichment By using external data sources the results can be enriched with additional data (e.g. on a product review site, linking to online stores for each product).

Extracting The items can be used to extract new data characterizing the whole result set, e.g. common terms in the documents or frequent authors.

Higher level search functions

According to Bates [1] a system should not only offer basic functionality. It should also provide support for search tactics, stratagems and strategies. In our model a HLSF is a function that uses lower level SOP and/or SSIM functions (called *moves* regarding Bates' terminology) for providing tactical and strategic support. For example, when searching for relevant literature about a certain topic a stratagem consisting of two tactics would be to *i*) search for documents that contain some terms describing the topic and then *ii*) using a function for exploring references and citations of documents to find related documents. An ideal system should also be able to support these kinds of search functions.

3. SUPPORTING SEARCH MODES

We think that the search mode taxonomy is flexible and general enough to be also well-suited for many other domains. We regard a search mode or a sequence of search modes (just called *search mode* in the following for the sake of simplification) as a higher level search function (or task) as defined by Bates. In the following we will give examples which functions are particularly required for supporting certain search modes. Functions from all three groups are required of course but we will focus on those that are the most important and distinctive ones. These requirements are listed in Table 1 and will be explained in more detail in the following.

		Search Functions		
		Select	**Organize**	**Project**
Lookup	**Locate**	Query language, Ranking method		
	Verify			Selection
	Monitor			Selection
Learn	**Compare**		Sorting	Selection
	Comprehend		Grouping	Faceting
	Explore	Formal filter conditions	Grouping, Clustering	Faceting
Investigate	**Analyze**		Sorting, Grouping, Clustering, Linking	
	Evaluate		Sorting	
	Synthesize		Join	

Table 1: Most important and distinctive (groups of) functions for each search mode

Lookup: Locate For supporting this search mode (often a known-item search) it is important to offer appropriate *select* functions that allow the specification of the known attributes. For example, when searching for a scientific publication, the user might know some words from its title as well as the publication venue—so the system must allow for searching in specific fields.

Lookup: Verify/Monitor If the user wants to verify that an item meets some specific and objective criteria or when s/he wants to monitor an item to maintain awareness the system should be able to *project* on the relevant parts or attributes of the result items. For example, when finding out whether a central processing unit (CPU) is compatible with a specific mainboard chipset the system should show the compatibility information of the CPUs in the result items.

Learn: Compare Comparing items in the results to identify similarities or differences requires the system to *organize* the items as a list and to offer a *projection* of all relevant aspects visualized in tabular form. Alternatively, the items may be organized in a multi-dimensional grid. For example, when comparing products, both price and performance of products are relevant criteria.

Learn: Comprehend For supporting comprehension of result items by finding patterns and traits the system should allow the user to *organise* and *project* the results by grouping them according to one ore more facets, in order to gain a understanding of the structure of the result set. For example, a user interested in buying an solid-state drive for his/her computer first has to comprehend the possible values of the relevant attributes (e.g. storage size, host interfaces, operating system requirements, etc.) by faceting.

Learn: Explore Faceted search supports exploration. Besides *selecting* a specific value of a facet as a formal filter condition, the system should offer functions for *organizing* the result items into different groups for each facet. For more content-oriented searches, clustering functionality may help the user in understanding the various aspects of a topic.

Investigate: Analyze Analyzing items to identify patterns and relationships is a very complex task. Thus, the system should offer several versatile and powerful *organization* functions.

There are several functions that may be helpful for the user here, such as *i*) (multi-dimensional) sorting, *ii*) grouping, *iii*) clustering and *iv*) linking of the result items. Sorting result items allows the user to inspect the items by the priority of one or more sorting criteria. The HyperScatter component of the visual information seeking system MedioVis [6] would be a proper visualization and interaction technique. Especially, multi-dimensional sorting might help in understanding the relationship between facets (e.g. when buying a digital camera, the user might want to learn which features have a strong influence on the camera price). Functions for grouping may help the user in gaining new insights or getting an overview of the result items (see preceding search modes). Clustering the result items may be helpful for finding previously unknown similarities by creating groups of items with an unknown meaning. Additionally, a clustered result set may support the user in getting an overview of the found items easier. Functions for linking the items can produce tree or graph structures of the result set. These functions can be used for creating e.g. networks based on some kind of relationship.

Investigate: Evaluate For judging the value of an item concerning a specific goal or purpose the system should be able to let the user *organize* the result items according to the important criterion, e.g. by sorting.

Investigate: Synthesize This search mode occurs when the user is creating new objects from the found result items. We envisage that a system may support this by offering a join function similar to joins in relational databases.

The system does not have to allow the user to perform all functions that are theoretically possible. Instead, the system should perform certain functions automatically and should use suitable preadjustments and defaults (see levels of system involvement by Bates [2]). Which functions the user should interact with depends on the search modes and the domain in which the system is actually used.

4. IMPLICATIONS FOR THE DESIGN OF SEARCH SYSTEMS

In the previous section we provided examples which functions are required for different search modes. An ideal search system should be flexible enough to support a broad variety of search modes. Which set of functions is exactly required

certainly depends on the context the system is used within and the tasks a user typically performs. Adding as many functions as possible may leads to a feature-bloated system. Instead, only the appropriate functions should be offered to the user. Richer functionality requires increased user expertise. Thus, the interaction and visualization techniques have to be chosen carefully to provides an easy-to-use system. Further open research issues concerning rich functionality have been described by Beckers and Fuhr [3].

The discussion in this paper has shown that the ideal search system extends classical IR functionality with typical database functions, as well as more advanced IR functions. Thus, typical IR systems as well as relational database systems are both far away from the ideal system. An XQuery system with full-text search might come closest today, but it lacks all the more advanced IR functions. Whatever the resulting query language might look like, however, it should be clear that it mainly targets at the application developer, who specifies the functionality needed, which is then mapped onto a user-friendly interface.

5. CONCLUSION AND OUTLOOK

We demonstrated how different search modes require different search functions of the system. Thus, an ideal search system suitable for various search modes should not only support classic search functions for ad-hoc retrieval (e.g. ordinary web search) but also more advanced functions described in this paper. Our grouping of search functions allows the identification of functions possibly required for a certain search mode. Previous research in this area can be categorized and integrated.

Further empirical research is necessary to validate our proposed mapping from search modes to search functions. A first step may be to show exemplarily that for a particular search tasks the users can benefit from improved and suitable functionality by controlling the other variables.

6. REFERENCES

[1] M. J. Bates. Information search tactics. *Journal of the American Society for Information Science*, 30(4):205–214, 1979.

[2] M. J. Bates. Where should the person stop and the information search interface start? *Information Processing and Management*, 26(5):575–591, 1990.

[3] T. Beckers and N. Fuhr. Towards the systematic design of IR systems supporting complex search tasks. In *Proceedings of the Task Based and Aggregated Search Workshop @ ECIR 2012*, April 2012.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September 2002.

[5] N. Fuhr, M. Lechtenfeld, B. Stein, and T. Gollub. The optimum clustering framework: Implementing the cluster hypothesis. *Information Retrieval*, 15:93–115, 2012. DOI: 10.1007/s10791-011-9173-9.

[6] M. Heilig, M. Demarmels, W. A. König, J. Gerken, S. Rexhausen, H.-C. Jetter, and H. Reiterer. Mediovis: visual information seeking in digital libraries. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, pages 490–491, New York, NY, USA, 2008. ACM.

[7] K. Hughes-Morgan and M. L. Wilson. Information vs interaction – examining different interaction models over consistent metadata. In *Proceedings of the IIiX conference*, 2012. To be published.

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

[9] P. Mutschke, P. Mayr, P. Schaer, and Y. Sure. Science models as value-added services for scholarly information systems. *Scientometrics*, 89(1):349–364, Oct. 2011.

[10] T. Russell-Rose. A taxonomy of site search. Talk at Enterprise Search Europe, UK, May 2012.

[11] T. Russell-Rose, J. Lamantia, and M. Burrell. A taxonomy of enterprise search. In *Proceedings of euroHCIR*, 2011.

[12] T. Russell-Rose, J. Lamantia, and M. Burrell. A taxonomy of enterprise search and discovery. In *Proceedings of HCIR 2011*, October 2011.

[13] D. Tunkelang. *Faceted Search*. Number 5 in Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

Ingredients for a User Interface to Support Media Studies Researchers in Data Collection

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

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

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

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

ABSTRACT

We describe our efforts to design an interface that supports media studies researchers in collecting data. Based on interviews about their search behavior we arrive at a set of search scenarios and for each we identify IR techniques that provide the required functionality. We end with a discussion about the implementation of such an interface and its re-usability across the humanities.

Categories and Subject Descriptors

H.5.2 [**User interfaces**]: User-centered design

1. INTRODUCTION

Research in the arts and humanities often follows an interpretive, associative method based on historic-cultural materials, including primary sources as well as secondary materials. Humanities researchers increasingly make use of online archives and libraries to collect and compare materials and this is changing the way they work [3]. Although the trend towards e-humanities is being addressed by computer science, current search tools remain ineffective in supporting humanities researchers in data collection [5, 17, 18].

Most support tools for e-humanities research focus either on supporting a single type of search process, are aimed at analysis and organization, or focus on a single collection of primary sources rather than the collection of secondary material across various sources and modalities. Letizia is an example of a user interface that assists a user in browsing the Web by pre-fetching related documents [10]. Flamenco is an interface that supports exploration of image collections through facets [20]. Imagesieve is an exploratory tool for museum archives based on entities [11]. See [15] for an overview of metadata enhanced interfaces for specific digital libraries. Other systems aim to support sensemaking of collected data. CombinFormation, for example, is a creativity support tool for searching, browsing, organizing, and integrating information [9]. Visualization and text analysis tools provide a wider variety of methods to organize and analyze material, e.g., MONK[1] and TaPoR.[2].

In this paper we revisit the search scenarios in which humanities researchers engage. We follow a human centered approach to derive the search scenarios that a tool for data collection in the humanities should support. We focus on a specific group of users, i.e., researchers in the field of media studies. Media Studies concerns the study of production, content and/or reception of various types of media, e.g., social media, film, and television [12]. The search for data in different modalities and across a wide variety of sources make this an interesting group for our analysis. We perform a set of interviews to analyze the information search behavior exhibited by media studies researchers during their research. Our contributions are establishing a set of search scenarios based on these interviews, identifying suitable information retrieval techniques that support these scenarios, and a discussion about the challenges in incorporating these techniques in an interface and its re-usability for other humanities disciplines.

[1] http://monkproject.org/
[2] http://portal.tapor.ca/

2. ESTABLISHING SEARCH SCENARIOS

Most research in the humanities starts out by gathering specific primary sources on a certain topic. When a selection of source materials has been made the search for additional materials starts in order to provide context for the source materials [1, 14]. In terms of information behavior a research project consists of successive information seeking processes each consisting of multiple search processes [19]. Each search process, whether for primary sources or other materials, consists of starting a search, several types of search actions, i.e., browsing, chaining, and monitoring, followed by differentiating, verifying and extracting information [7].

Each of these actions can be observed in the search processes of media studies researchers in various stages of their research cycle [4, 12]. However, the effectiveness of current search tools to support these actions depends on the goal of the search process and the organization of the material. For example, browsing is easily facilitated in an interface by providing facets over the metadata annotations of documents, but metadata is usually unavailable. We will focus on the contextualization stage of the research cycle of media studies, where the primary sources have already been collected and the search for additional material starts. In this stage multiple sources of different modality are searched for and we expect that the analysis of this process will provide search scenarios that facilitate the development of an appropriate search interface.

Interview method and analysis. We interviewed ten media studies researchers from 3 different institutes with varying levels of experience: 2 PhD students, 5 post-doctoral researchers, 1 assistant-professor, and 2 full professors. Several media are being studied: television (10), radio (2), news papers (2), and documentaries (1). The interview was conducted in a semi-structured style and consisted of three parts: (i) identification of a recent research project; (ii) open questions about search processes and research questions during the project; and (iii) an interactive part in which subjects

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	frequency
newspapers	1	4	-	2	-	1	-	1	1	3	7/10
interview	1	-	1	-	-	-	1	2	-	4	5/10
magazine/tvguide	-	1	-	1	-	-	1	2	1	-	5/10
entity homepage	2	-	1	-	1	1	-	-	-	1	5/10
forum/blog	-	-	-	1	3	1	-	-	-	1	4/10
paper archive	1	-	-	-	-	-	4	2	1	-	4/10
Wikipedia	1	-	-	1	-	1	-	-	-	1	4/10
reports	1	-	1	-	-	-	-	2	-	-	3/10
book	1	-	-	1	-	-	-	-	-	-	2/10
specific site	2	1	-	-	-	-	-	-	-	-	2/10
twitter/facebook	-	-	-	-	1	-	-	-	-	-	1/10

Table 1: Number of times a source is mentioned in an interview. Frequency is the fraction of interviews to which a code applies.

wrote down the search processes on index cards. Interviews lasted about 30 minutes, were tape-recorded and later transcribed.

In our analysis of the interviews we focus on those questions that address the process of secondary material collection. We apply open-coding [16], to identify different types of material, the type of information needs the materials satisfy, and search strategies used by media studies researchers to retrieve the material. Given this data we then identify search scenarios on a more general level. Note that when using quotes, square brackets [..] indicate modifications to the original quote to improve understanding or to protect the anonymity of the subject. For identification purposes interviewees are assigned a number, i.e, I1 to I10.

Interview results. We first consider the different types of materials that media studies researchers search for besides their primary source material, e.g., television programs. Table 1 shows the types of material mentioned during the interviews. The source that is most often used are newspapers. They provide relevant context in terms of: (i) reviews about a television program; (ii) information on events during the period in which a program was broadcasted; (iii) to reconstruct which programs were broadcasted during a time period; and (iv) whether a program itself caused some event or controversy. Interviews are used when the required information is not otherwise accessible. Interviews with directors and producers provide context in terms of productions information, e.g., why a certain format was choses for a program. Interviews with people who watched a program that is no longer available provide information on the attitude of viewers in that time. Magazines and tv-guides are interesting mainly for the reviews of programs they contain or for reconstructing a broadcasting schedule. The context of homepages depends on the entity of interest. The homepage of the production company of a program provides production information, e.g., when it was broadcasted. Alternatively, a person mentioned in a program may be of interest and his/her homepage allows the researcher to find out more about that person. Wikipedia is also a popular source for this type of information. The main use for fora and blogs is to get a sense of people's attitudes towards a certain program. Other sources used mostly to get production information are paper archives and internal reports. Books and other sites, e.g., history site, are used to provide historical context for a program. The lack of use of social media as sources is due to our sample of media researchers, who work mostly with television.

Next, we categorized the various types of information need that the materials satisfy into 3 types: (i) general background information on a topic; (ii) information on specific entities; and (iii) identifying conversational information about an event. Table 2 shows the number of times each type of information need occurred in each of the interviews. Searching for background information is mentioned by most of the interviewees. We found two general topics on which media studies researchers require background information: cultural context and media production context. Cultural context is necessary to understand the reception of a program by society: (I7) "[regarding television shows for women in the 70's] I found that as the topics discussed were more taboo, those topics were still taboo then... That the program was broadcasted at a later time slot and there the difficult topics like divorce and birth control were discussed." Another type of context, i.e., media production context, is necessary to understand why programs turned out in a certain way: (I1) "There are different ways to interpret this rebellion. As producers got more freedom in creating television programs and while the television landscape was still very much divided, more artistic programs could emerge." The goal of gathering this type of context is to learn more about the situation in which a program was broadcasted or created.

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	frequency
background	5	-	2	-	4	1	5	4	1	-	7/10
entities	2	-	3	-	1	4	-	-	-	1	5/10
conversational	1	4	1	4	-	-	-	-	-	5	5/10

Table 2: Number of times a type of information need is mentioned in an interview. Frequency is the fraction of interviews to which a code applies.

Regarding context for entities we find that biographic information about people is important: (I6) "For example, if a [person] was mentioned then I would know his name, but not his ethnic background. Part of the analysis was finding out [peoples]' ethnic background" and (I3) "not all journalists are so vain to put themselves online. Especially the ones that are not well known and then I could not find their specialism." For organizations information about the internal culture and policies is important: (I1) "Sometimes you search for policy information. How did the broadcasting company present itself, what does it mean for the broadcasting company... So what role does the program play in the perception of the broadcasting company's own history." These quotes show a particular interest in specific information about entities.

In five cases interviewees engage in a conversational search, i.e., they look for the discussion around certain events. For example three interviewees mention that they look for controversy: (I2) "I searched in newspapers for controversies, you are actually searching in other media for reflection on what happened. And then you find the title of the program. You start with there was a fight, and you need to know what it was about" and (I10) "Of course the programs that are the most controversial, those that fuel public debate and get the most media attention, are the ones that I examined closest." These quotes show how media studies researchers are interested in the reasons that cause a controversy. A similar type of information need is mentioned by two other interviewees interested in multiple views on a topic: (I3) "I was interested in the relation between political issues and whether news programs show multiple views on every topic, for example a government source and an opposition source" and (4) "I chose to organize [political figure]'s story chronologically: her rise, moments of glory, and her fall."

We find that of the information needs described above the search task of finding background information is exploratory, i.e., it is unclear what the actual goal is other than information about a general topic. Regarding the entity information need the search task is very specific, e.g., finding journalists' specialism. The search task may be repeated multiple times for different entities, but the goal remains the same. In the case of conversational search multiple types of search task are required to satisfy the information need. For example, to find multiple views on a political issue, first, an exploratory

Figure 1: Sketch of an exploratory meta search engine, numbers are used to reference components.

search task is required to find the people involved. Second, a more targeted search is necessary to find the attitude of each person towards the issue. We also note that these types of information need occur multiple times during a research project, e.g., (I3) looking for multiple views on political issues and the specialism of journalists reporting on those issues.

Finally, we consider the search strategies used to collect material. The tools used most often are web search engines, for example when searching for controversial documentaries: (I2) "I Google not for documentaries, because I do not know which are controversial. I use keywords of which I know that they are related to controversies such as: conflict." Media studies researchers are trained in searching through archives and so also use various strategies when searching for additional material: (I1) "You have to search in different ways, of course... That was the same in the archive. When you can not find anything on a shelf organized per director, then use decades as a searching criterion. So you are always trying different angles." Another example shows how chaining via web links is used to reconstruct the conversation about an entity: (I4) "right, you end up on a forum with a discussion about her biography. Where one post suggests to look at this and this. Another suggests you should read that. In this way you get a lot of pointers to links in a very organic way, and I collect it in a folder with interesting links." Even for a specific information need such as the nationality of an entity several sources are searched: (I6) "Wikipedia is also a search engine. I needed to know the ethnic background of [people]. You can do this in all sorts of ways, for example fora, but also other sites that provided information about the nationality of [people]."

These quotes show how media studies researchers use multiple strategies and cover multiple sources to get at the information they need. The most popular tool are web search engines that while specialized in navigational search are also used for exploratory and informational type searches.

3. IDENTIFYING IR TECHNIQUES

The types of information need identified in the interviews suggest that an interface that supports multiple data search and collection tasks should support the following search scenarios: (i) general search; (ii) entity information search; (iii) entity relation search; and (iv) information management.

Background search. When the goal of the search is to find general background information on a certain topic, media studies researchers engage in an exploratory search task over multiple sources, e.g., news archives, libraries, and the Web. To support this task we propose to combine features of an exploratory search engine [13] with those of a meta search engine [6], see Figure 1. A search box (1) is available for the user to type in keywords in response to which a ranked list of result snippets is displayed. To support users in finding material across sources the interface aggregates results from multiple sources. A sidebar (2) shows a list of common information sources, e.g., Wikipedia. Checkboxes are available to select or deselect one or more sources from which results for the keywords are retrieved. Facets (3) are available on the right side and support the user in filtering the result set and learning about the topics covered in the results set. A typical issue for meta search engines is the aggregation of results from different sources. We propose to leave control to the user: for each source we display one result with the option to expand (4) a source and display its results. The user is able to drag and reorder the sources in the sources list in order to select the source that will be shown at the top.

Entity information search. Another scenario is when the goal of a search task is to find more information about an entity, e.g., the ethnic background of a person, or an event, e.g., reality shows causing a controversy. To support this kind of tasks we propose an interface that combines techniques for entity resolution [8], list completion [2] and query by example [21], see Figure 2.

Figure 2: Sketch of an entity information search interface, numbers are used to reference components.

Entity resolution is necessary when only an entity's name is known, e.g., the name Michael Jordan usually refers to the basketball player, but the target could be a researcher or anyone with that name. Typing a name in the entity resolution component (1) and selecting a knowledge base against which to resolve the entity, e.g., Freebase,[3] results in a list of possible targets for the entity. These targets, e.g., a Wikipedia page or homepage, provide an identifier for the entity and context information. This is also useful in the case of events, which only in special cases have specific names, i.e., named events.

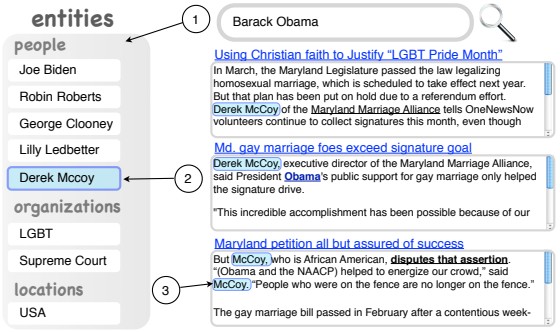
Query by example (2) supports finding information about a topic for which some information (context) is already available. Given a news article about a certain topic, find more documents that describe that same topic. Possible target resources for query by example are video databases and news archives as items from different resources describe an event in different ways.

List completion supports a scenario where a researcher wishes to find a group of entities that all have something in common, e.g., members of the same political party, but he/she has only identified some members of this group. Providing a number of examples the list completion component (3) results in a list of entities that have characteristics in common with the examples, e.g., entering Wikipedia URLs as example entities returns other entities (URLs) from Wikipedia that share characteristics with the examples.

Entity relation search. In some cases multiple types of search processes are required in order to satisfy an information need, i.e., to find multiple views on a topic exploratory and targeted searches alternate. For example, finding who are the opponents and proponents on a political issue and the reasons for their respective views. We propose an interface that facilitates viewing entities in context using dynamic snippets, see Figure 3. Whenever a search query is issued, a sidebar (1) on the left of the interface is populated with entities. To locate candidate entities online named entity recognition

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

Figure 3: Sketch of a search engine with dynamic snippets and entity highlighting, numbers are used to reference components.



is performed (NER). As NER is a costly operation this is done incrementally, i.e., first on the top 10 pages then, if the user paginates, on the next 10 pages, etc. Initially snippets in the result list contain the same text as returned from the source. The snippets, however are dynamic and when hovering over an entity (2) the snippet is updated to show a piece of text from the document in which this entity occurs, highlighting the entity (3). Jumping to an entity's position in result documents and highlighting, allows a user to inspect the context in which an entity occurs without opening each document. To account for the limited amount of space each snippet is made scrollable to enable inspection of other occurrences of the entity.

Information management. In all cases listed above the proposed techniques support finding information, however, this information needs to be stored and organized. Rather than an elaborate information organization environment common to sensemaking tools, we propose to allow the creation and assignment of labels to relevant documents. For example, in the case of the rise and fall of a political figure, documents can be organized according to the start of career, moments of glory and the eventual downfall.

4. DISCUSSION

In this paper we have established several search scenarios in which media researchers engage and recommend IR techniques that provide support in these scenarios. There are however two unresolved issues: (i) will these techniques work for each source or do they have to be tuned towards the characteristics of each data collection; and (ii) is an interface that incorporates these techniques re-usable by other humanities researchers? We believe the first point can be addressed by carefully documenting the characteristics of collections and the dependence of the retrieval performance of IR techniques on these characteristics. Whether linking a video archive with a news archive is a different task from linking a photo archive with a news archive will depend on the agreement of the characteristics of the datasets. To address the second point, we believe it is necessary to separate the functionality and the sources into modules and allow the user to compose the interface required for the search task at hand. These modules also have to be configurable, for example, the facets entities and phrases in a facets search component may be useful in some cases, while others require events and years.

Our main next step is to take these requirements and realize an interface that supports the various information search tasks of media researchers and is re-usable across the humanities.

Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.-061.814, 612.061.815, 640.004.802, and partially by the Center for Creation, Content and Technology (CCCT).

References

[1] D. Altheide. *Qualitative media analysis*. Sage Publications, Inc, 1996.

[2] K. Balog, M. Bron, and M. de Rijke. Query modeling for entity search based on terms, categories, and examples. *ACM Transactions on Information Systems (TOIS)*, 29(4):22, 2011.

[3] C. Borgman. The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, 3(4), 2009.

[4] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *SIGIR'12*, 2012.

[5] E. Collins and J. Michael. How do researchers in the humanities use information resources? *Liber Quarterly*, 21(2), 2012.

[6] D. Dreilinger and A. Howe. Experiences with selecting search engines using metasearch. *ACM TOIS*, 15(3):195–222, 1997.

[7] D. Ellis and M. Haugan. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *J. Doc.*, 53(4):384–403, 1997.

[8] V. Jijkoun, M. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 23–30. ACM, 2008.

[9] A. Kerne, E. Koh, S. M. Smith, A. Webb, and B. Dworaczyk. combinformation: Mixed-initiative composition of image and text surrogates promotes information discovery. *ACM Trans. Inf. Syst.*, 27(1):5:1–5:45, 2008.

[10] H. Lieberman. Interfaces that give and take advice. *In HCI in the New Millennium*, pages 475–484, 2001.

[11] Y. Lin, J. Ahn, P. Brusilovsky, D. He, and W. Real. Imagesieve: Exploratory search of museum archives with named entity-based faceted browsing. *ASIST'10*, 47(1):1–10, 2010.

[12] B. Lunn. User needs in television archive access: Acquiring knowledge necessary for system design. *JoDI*, 10(6), 2009.

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

[14] C. Palmer. Scholarly work and the shaping of digital access. *JASIST*, 56(11):1140–1153, 2005.

[15] A. Shiri. Metadata-enhanced visual interfaces to digital libraries. *JIS*, 34(6):763–775, 2008.

[16] A. Strauss and J. Corbin. Basics of qualitative research: Grounded theory procedures and techniques. *Basics of Qualitative Research Techniques and Procedures for Developing Grounded Theory*, 270, 1990.

[17] E. Toms and H. O'Brien. Understanding the information and communication technology needs of the e-humanist. *J. Doc.*, 64(1):102–130, 2008.

[18] J. Unsworth. Tool-time, or 'haven't we been here already?' ten years in humanities computing. *Transforming Disciplines: The Humanities and Computer Science*, 2003.

[19] T. Wilson. Models in information behaviour research. *J. Doc.*, 55(3):249–270, 1999.

[20] K. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *SIGCHI'03*, pages 401–408, 2003.

[21] M. Zloof. Query-by-example: A data base language. *IBM systems Journal*, 16(4):324–343, 1977.

Exploring Italian Wine: a Case Study of Aesthetics and Interaction in a Generative Information Visualization Method

Luca Buriano
Telecom Italia Lab
Via Reiss Romoli, 274
10148 Torino, Italy
luca.buriano@telecomitalia.it

ABSTRACT
Interactive applications can greatly benefit from Information Visualization (Infovis) methods addressing aesthetic and creative design aspects, to help in effectively conveying the meaning of complex data. We present a novel Infovis design and method, applied to the interactive visual exploration of Italian wines' properties. This work adopts a generative approach, based on automatic creation of the visual layout according to functional as well as aesthetic and perceptual criteria.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Algorithms, Design, Human Factors.

Keywords
Information visualization, graphic design, generative art, generative design, visual recommender systems, interactive data exploration, wine culture.

1. INTRODUCTION
Aesthetic and creative design aspects are essential to the development of a successful Information Visualization (Infovis) project, helping to immediately and effectively convey the meaning of complex data [10]. To this end, we think that inspiration can be found in the approach of generative visual art and generative graphic design, where the artist/designer, after envisioning a set of aesthetic, functional and semantic criteria, models them as a process and lets the resulting system organize into the actual, emergent visual patterns [3][5][6]. This kind of flexibility can greatly benefit interactive applications, where the visualized data are dynamic by nature. Following this framework, we present a demonstration of a novel Infovis visual concept and method, applied to an interactive application scenario. This concept emphasizes, in addition to its specific visual design: a) visual layout creation driven by a fitness function taking into account not only data relationships, but also aesthetic, perceptual and graphic design aspects; b) immediacy and easiness in the interactive exploration of the resulting visualization.

2. RELATED WORK
The aesthetic features of an Infovis project depend primarily on the human designer's creative design insights and skills [10][9][1]. When dealing with dynamic and interactive applications, where the data to visualize are selected on the fly, the designer faces the additional challenge of managing the layout and aesthetics of unpredictable data configurations. In the Infovis field, this is usually addressed in essentially two ways: a) letting the aesthetic features emerge from layout algorithms designed to show data properties and relationships with clarity (see, for example, graph layout algorithms[4][11]); b) directly embedding aesthetic criteria into the visualization algorithm (see, for example, circle packing layout algorithms [2]). However, the combined use of a) and b) and the embedding of multiple, potentially complex aesthetic factors remain challenging and relatively unexplored areas. Our work aims to integrate the approaches a) and b) in a single generative framework. The generative process is applied to a visual concept with many degrees of freedom, related to the family of grid visualizations (see for example [8]). The process is driven by a fitness function, able to take into account multiple data-driven, aesthetic and graphic design visual factors. The flexibility and modularity of the fitness function allows the designer to experiment with different aesthetic criteria and styles. Moreover, the generative approach naturally leads to the creation of a diverse set of visual solutions for a given data set, enriching the user's experience with a source of visual novelty.

3. APPLICATION SCENARIO
Our work's example application scenario is a visual recommender system on the domain of Italian wines from the Piemonte region, based on a database of these wines and their properties related to smell, taste, grapes and production locations. This application scenario has been envisioned in the context of the "PIEMONTE Project"[1], a research project whose core concept is that "smart things" can play the role of gateways for enhancing the interaction between people and a territory with its cultural heritage.

When the user chooses a wine from the list, the system extracts from the database a set of wines with similar properties (according

[1] http://www.piemonte.di.unito.it/index.html

to a given similarity function) and creates an interactive visualization of this set, allowing visual exploration of wines' properties.

4. VISUALIZATION

Our visualization method works on a set of n items, where each item, identified by a name label, is linked to a list of properties taken from a domain-specific dictionary and categorized in property groups. In the wine recommendation scenario, items and item properties become wines and wine properties, categorized in the following property groups: smell, taste, primary grape, secondary grape and production location.

The algorithm is composed by the following steps:

1) An initial visualization layout is built as a n x m matrix, where each row corresponds to a distinct item and each column corresponds to a distinct property (m is the number of distinct properties appearing in the given item set). Each element in this matrix, corresponding to a item-property combination, is assigned a color, determined by the property group the element's property belongs to, according to a given color table. For the wine domain, we chose the following color table association: taste = red, smell = indigo, primary grape = green, secondary grape = cyan, production location = yellow. Elements corresponding to item-property combinations not appearing in the visualized item set (called in the following "empty elements") are assigned the background color. A special column in the matrix will contain item's name labels.

2) A large number of candidate visualization layouts is generated, where each candidate is obtained by randomly rearranging rows and columns in the matrix (keeping individual rows and columns intact, for data coherence). Candidate layouts are assigned a score according to a fitness function; the actual visualization layout is randomly selected among the candidates with the highest scores. The fitness function takes into account a set of factors related to the visualization's functional, aesthetic and graphic design factors, including:

- minimization of the distance between rows corresponding to similar items (the similarity between two items is calculated as a function of the properties they share);
- preference for visual grouping of columns linked to the same property group (i.e., of the same color);
- preference for specific shape properties of local or global visual emergent patterns (e.g., preference for spatially compact patterns, minimizing the number of gaps between colored matrix elements);
- preference for specific color pairings of matrix elements, according to a given color theory (e.g., preference for complementary colors of adjacent elements).

The fitness function is calculated as a weighted sum of components, where each component calculates the score assigned to a given layout with respect to a different design factor. This modular structure allows for seamless modelization and inclusion of other design factors when needed, for example when the designer wants to experiment with different visual and interaction styles.

During this step, groups of two or more columns can be optionally fused together into one column, provided that non-empty elements from these columns won't overlap in the resulting column. These column fusion operations generate new candidate layouts.

3) Finally, the chosen visualization layout is displayed (Figure 1). Each matrix's element is drawn as a borderless tile painted with the color associated to the element. Item names are displayed in their column, with the tile belonging to the chosen item highlighted by a thin border. Optionally, the brightness and width of adjacent columns of the same color is slightly modified, in order to increase visual diversity and allow the user to tell these columns apart more easily.

5. INTERACTION

Interactive exploration of the resulting visualization can be performed in several ways, including:

- selecting a wine in the name column: all the wine's properties in the corresponding row will be highlighted, by showing their labels (e.g. in Figure 1, the 5th wine from the top is selected) ;
- selecting a property column: the property's label will be shown at the interaction point and all the wines sharing that property will be highlighted, emphasizing their name labels' font and/or coloring their tiles with the same color as the property tiles (e.g. in Figure 2, the “amarognolo” (slighty bitter) taste property is selected). It should be noted that, by default, properties' labels are intentionally not shown until they (or related wine names) are selected, in order to encourage visual data exploration and reduce information overload. At any moment the user can, of course, ask for permanent display of a given label, or all labels;
- asking for alternate layouts (from the best candidates found in the step 2) of the visualization algorithm);
- making a new recommendation/visualization query, choosing one of the visualized wines, or a wine from the global list;
- changing the visualization's settings, e.g. setting the number of similar wines shown as the result of a query, or setting the minimum number of wines a property must belong to, in order for the property to be shown;
- manually rearranging the layout of rows and columns with drag-and-drop actions.

When the user makes a new query, a smooth visual transition between visualization layouts is provided by animations “dissolving” the current layout (by moving tiles outwards in random directions and fading them out) and “assembling” the new one (by moving tiles from random outscreen positions to their actual position in the layout, and fading them in).

6. IMPLEMENTATION

The Infovis method described in this paper has been implemented as a prototype application for the wine recommendation scenario using Processing[2], an open source programming language and environment for creating images, animations, and interactions [7]. The availability of a powerful language for building visual structures, together with the compatibility with the Java programming language, made Processing an excellent rapid

[2] http://processing.org

development environment for both the visual design and the algorithmic aspects of our work.

So far, we ran the prototype on standard PCs and on an interactive whiteboard, the latter's large touchscreen providing a natural environment for our work's visual and interactive features.

7. PUBLIC DEMONSTRATION

The prototype application was exhibited at the CHEESE 2011 festival[3]. We were located in the headquarters of Slow Food[4], the event's organizer and PIEMONTE Project's partner. At this place, one of the event's focal points, people attending the event could see the application, interact with it and comment on it. We received a strong positive feedback from the event's visitors, about both the usefulness and the aesthetic value of the presented application.

8. ACKNOWLEDGMENTS

We wish to thank Slow Food for providing the database of wines' properties used in our work.

9. REFERENCES

[1] Bateman, S., Mandryk, R.L., Gutwin, C., Genest, A., McDine, D., Brooks, C. Useful Junk?: the Effects of Visual Embellishment on Comprehension and Memorability of Charts. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI '10),* 2573-2582. DOI=http://doi.acm.org/10.1145/1753326.1753716

[2] Collins, C.R., Stephenson, K. A Circle Packing Algorithm. Compuntational Geometry 25 (2003), 233-256

[3] Galanter, P. What is Generative Art? Complexity theory as a context for art theory. In *GA2003–6th Generative Art Conference* (2003)

[4] Heer, J., Bostock, M., Ogievetsky, V. A Tour Through the Visualization Zoo. *Communications of the ACM,* Volume 53, Number 6 (2010), Pages 59-67. DOI=http://doi.acm.org/10.1145/1743546.1743567

[5] McCormack, J., Dorin, A. and Innocent, T. Generative Design: a paradigm for design research. In Redmond, J. et. al. (eds) *Proceedings of Futureground, Design Research Society*, Melbourne (2004).

[6] Mueller, B. VisualPoetry - generative graphic design for poetry on the road. *SIGGRAPH '09*. DOI=http://dx.doi.org/10.1145/1667265.1667310

[7] Reas, C., Fry, B. *Processing, a Programming Handbook for Visual Designers and Artists.* The MIT Press (2007).

[8] Shneiderman, B, Feldman, D, Rose, A., Ferré Grau, X. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries (DL '00),* 57-66. DOI=http://doi.acm.org/10.1145/336597.336637

[9] Tufte, E.R., Goeler, N.H., Benson, R. *Envisioning information.* Graphics Press (1999).

[10] Vande Moere, A. Infosthetics: the beauty of data visualization (interview for PingMag magazine). http://pingmag.jp/2007/03/23/infosthetics-form-follows-data/

[11] Visual Complexity. http://www.visualcomplexity.com/vc/

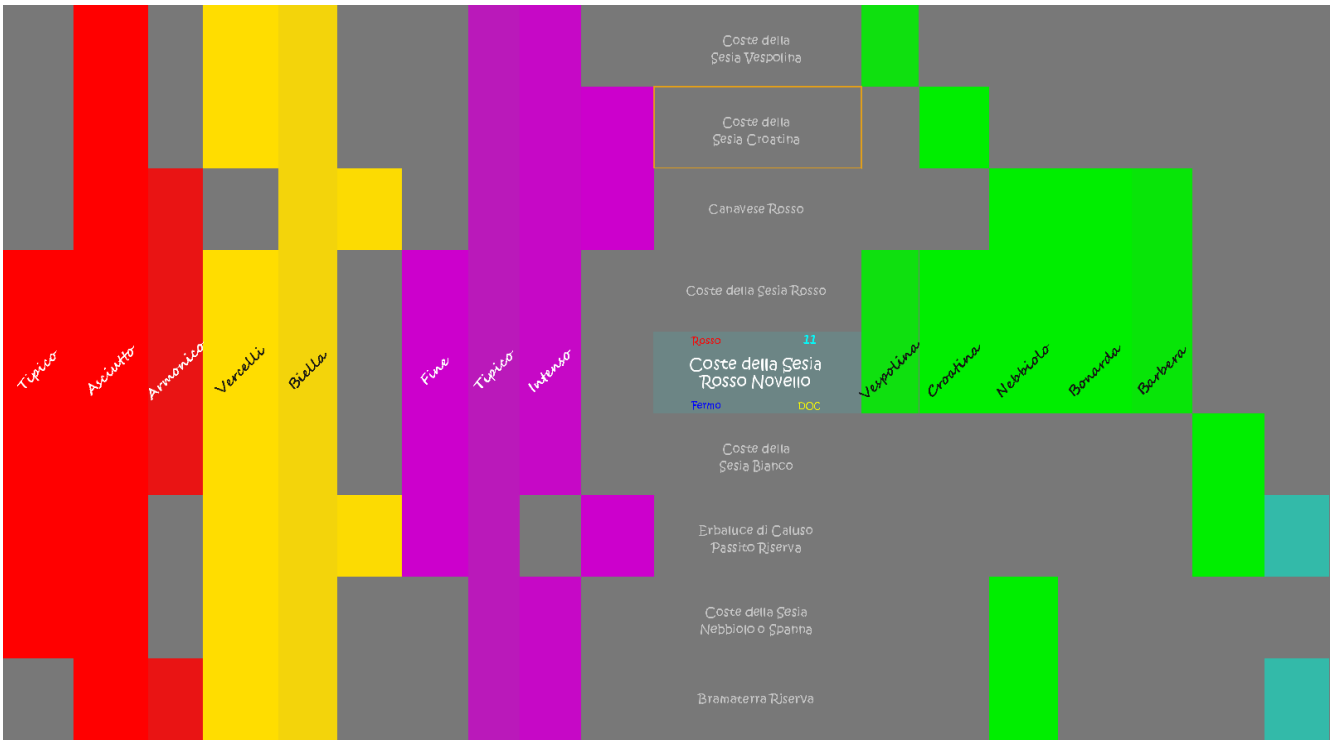


Figure 1.

[3] http://cheese.slowfood.it/

[4] http://www.slowfood.it/

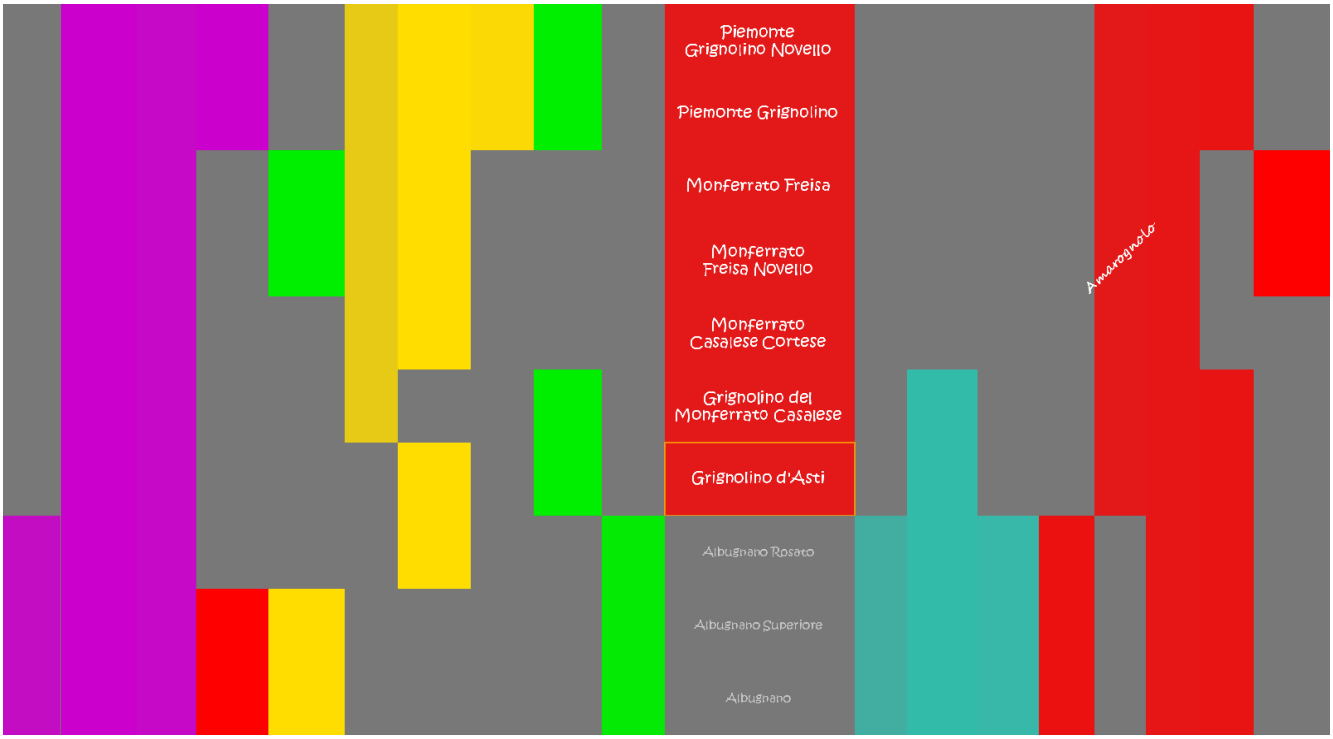

Piemonte
Grignolino Novello
Piemonte Grignolino
Monferrato Freisa
Monferrato
Freisa Novello
Monferrato
Casalese Cortese
Grignolino del
Monferrato Casalese
Grignolino d'Asti
Albugnano Rosato
Albugnano Superiore
Albugnano
Amaragnolo

Figure 2.

From Task-based Evaluation to Feature-based Evaluation in Personal Search

Sargol Sadeghi
School of Computer Science & Information Technology
RMIT University
Melbourne, Australia
seyedeh.sadeghi@rmit.edu.au

Mark Sanderson
School of Computer Science & Information Technology
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Falk Scholer
School of Computer Science & Information Technology
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

ABSTRACT
Task-based evaluation has been suggested as a solution for comparing search systems in the personal context. However, as personal search tasks are broad, dependent on users, and have different levels of specificity [3], focusing on the building blocks (or characteristics) of these tasks could provide a more reliable and maintainable alternative for evaluation. Moreover, the characteristics can be used to determine to what extent evaluation results are generalizable and comparable across different users and tasks.

In this position paper, a *characteristic reference model* for personal search tasks will be introduced. Based on this model, different search systems can be compared not only in relation to task types, but also in terms of the characteristics that are most influential in search tasks, increasing the level of detail at which comparisons can be made.

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

General Terms
Performance, Design, Experimentation, Human Factors.

Keywords
Personal Search, Task-based Evaluation, Task, Search Characteristic.

1. INTRODUCTION
Providing search solutions to retrieve information that has been seen previously is the main focus in the *personal search* context [8]. To compare the effectiveness of search systems in the personal context, identifying common search tasks is of key importance. For example, Kelly and Teevan [3] proposed building a shared collection of common tasks instead of studying tasks in separate research groups. Common tasks for evaluation purposes have also been suggested in other disciplines such as HCI (Human Computer Interaction). For an instance, Whittaker et al. [7] introduced *reference tasks* with the goal of comparing interaction techniques.

However, it is challenging to identify common search tasks, particularly in the personal context, due to the variety of search needs among different users. Controlling the variety of tasks under a set of task *types* was proposed as an approach for evaluating personal search systems by Elseweiler and Ruthven [1]. In this study, three task types were identified based on a search *characteristic* to control the evaluation experiments; and a *task-based* evaluation conducted where the search systems are compared in relation to the search tasks. However, as the task-based evaluation focuses on specific task scenarios, there is a disadvantage that the acquired results cannot be generalized [5]. This is while solving task-based evaluation problems and developing a new type of evaluation has been highlighted [9].

To overcome this problem, we propose to incorporate the underlying *characteristics* of tasks. These characteristics, being more general in nature, can support the identification of commonalities across different tasks in terms of their components. For this purpose, we introduce a characteristic reference model in the next section.

2. CHARACTERISTIC REFERENCE MODEL
With the focus on search characteristics to compare personal search systems, first we must acquire knowledge about the range of characteristics that can affect the retrieval process. Based on these characteristics, we can then identify *similar tasks*, which have common search characteristics. This notion of *explicit* similarity supports a fair comparison of search systems in relation to the user tasks.

However, it is also possible to define *implicit similarity* between tasks. Here, tasks do not necessarily share the same set of characteristics, but their characteristics have been demonstrated to have the same effect on the retrieval process. Consider the following simple example of the implicit similarity concept.

From pilot user studies that we have conducted with the aim of identifying different types of personal search tasks, the user's *level of knowledge* in relation to the target information and task has been observed as a search characteristic influential in retrieval results. Based on this characteristic, we proposed a hierarchy of personal task types for level of knowledge, as shown in Figure 1.

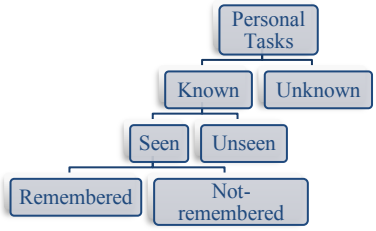


Figure 1. Personal task types and level of knowledge

In the proposed task hierarchy, for example, the user's state of knowledge might be that the target information is *unknown*, where the user does not know whether the required information item exists. Another possibility is that the user is searching for an information item that they know exists and have seen before, but is currently *not-remembered*.

In our observations of users, there are situations where user search behavior for not-remembered tasks is the same as for unknown tasks. For example, one of these situations is when the last access time to the information is prior to last month; here, the user does not know how to get to the information.

In the literature, the time of last access to required information has been called the task *temperature*. For this search characteristic, three values of *hot* (accessed within the last week), *warm* (accessed within the last month), and *cold* (accessed prior to the last month) have been suggested [1]. Based on this observation and from the gathered characteristics and values, it is possible to derive a simple rule as an example of implicit task similarity, illustrated in Figure 2.

If:
Task A= {<Level of knowledge: Unknown>}
Task B= {<Level of knowledge: Not-remembered>,
<Temperature: Cold>}
Then:
Task A similar to Task B.

Figure 2: Implicit similar tasks

From Figure 2, it can be seen that if there are two task scenarios identified under two different types (e.g. unknown and not-remembered), in some situations (e.g. cold temperature) they could have a similar effect on the retrieval process. In other words, it is possible that tasks which are in fact highly similar can occur under different task types. Such relationships have not been considered in task-based evaluations, where the focus is on specific task scenarios.

The previous scenario is a simple example; more realistically, it is likely that many different characteristics affect search tasks, in terms of: user, search need, search strategy, search context, information, and the collection of information. Deriving comprehensive rules for task similarities requires extensive user studies in both qualitative and quantitative aspects. We intend to extrapolate a set of rules composed of *Characteristic: Value* settings, as a reference model for identifying similar tasks.

In building this reference model, we need to further explore:

- the key characteristics that are influential in a search task
- interdependencies between characteristics
- the importance of characteristics in affecting retrieval results

Such a model will incorporate the characteristics proposed when studying tasks in different search applications (such as the goal of the user, task complexity, and topic familiarity [2, 4, 6], in both work task and search task aspects), as these are potentially applicable in the personal context. Characteristic settings will be derived by observing real task scenarios and mapping how search characteristics affect search tasks. In this mapping, we consider the interactions of characteristics.

Based on this characteristic reference model, similar tasks can be either created from scratch, or selected from the recorded tasks in current studies where characteristic details are available. Search systems can then be compared in relation to explicitly or implicitly similar tasks. The advantage of using this model is not only limited to enriching the comparability of personal search systems, and the generalizability of comparison results, but it can also lead to a complementary evaluation approach, where assessing the effect of one characteristic on the performance of search systems is important.

3. CONCLUSION

In this paper, we proposed a characteristic reference model for evaluating personal search systems. As there are a variety of tasks in the personal context, this model is based on identifying building blocks, and how they affect search tasks. This approach will enable better control and comparability across different users and tasks, rather than focusing on specific instances of tasks as is currently done in task-based evaluation. Focusing on these characteristics not only facilitates the evaluation of search systems based on search tasks through detailed comparisons, but also provides evaluations on characteristics in affecting the effectiveness of search systems.

4. REFERENCES

[1] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 23–30. ACM, 2007.

[2] P. Ingwersen. Selected variables for ir interaction in context: Introduction to irix sigir 2005 workshop. In *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, pages 6–9. Citeseer, 2005.

[3] D. Kelly and J. Teevan. 11 understanding what works: Evaluating pim tools. *Personal information management*, page 190, 2007.

[4] S. Kim and D. Soergel. Selecting and measuring task characteristics as independent variables. *Proceedings of the American Society for Information Science and Technology*, 42(1):n–a, 2005.

[5] W. Kraaij and W. Post. Task based evaluation of exploratory search systems. In *Proc. of SIGIR 2006 Workshop, Evaluation Exploratory Search Systems, Seattle, USA*, pages 24–27, 2006.

[6] Y. Li and N. J. Belkin. An exploration of the relationships between work task and interactive information search behavior. *JASIST*, 61(9):1771–1789, 2010.

[7] S. Whittaker, L. Terveen, and B. Nardi. Let's stop pushing the envelope and start addressing it: a reference task agenda for hci. *Human-Computer Interaction*, 15(2):75–106, 2000.

[8] D. Elsweiler, D. E. Losada, J. C. Toucedo, and R. T. Fernandez. Seeding simulated queries with user-study data forpersonal search evaluation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 25–34. ACM, 2011.

[9] K. J¨arvelin. Ir research: systems, interaction, evaluation and theories. In *ACM SIGIR Forum*, volume 45, pages 17–31. ACM, 2012.

Visualization of Clandestine Labs from Seizure Reports: Thematic Mapping and Data Mining Research Directions

William H. Hsu[1] bhsu@ksu.edu
Mohammed Abduljabbar[1] xec@ksu.edu
Ryuichi Osuga[1] ryusuga@ksu.edu
Max Lu[2] maxlu@ksu.edu
Wesam Elshamy[1] welshamy@ksu.edu

[1] Department of Computing and Information Sciences
[2] Department of Geography
Kansas State University
Manhattan, KS 66506
+1 785 236 8247

ABSTRACT
The problem of spatiotemporal event visualization based on reports entails subtasks ranging from named entity recognition to relationship extraction and mapping of events. We present an approach to event extraction that is driven by data mining and visualization goals, particularly thematic mapping and trend analysis. This paper focuses on bridging the information extraction and visualization tasks and investigates topic modeling approaches. We develop a static, finite topic model and examine the potential benefits and feasibility of extending this to dynamic topic modeling with a large number of topics and continuous tome. We describe an experimental test bed for event mapping that uses this end-to-end information retrieval system, and report preliminary results on a geoinformatics problem: tracking of methamphetamine lab seizure events across time and space.

Categories and Subject Descriptors
H.3.3 **[Information Storage and Retrieval]**: Information search and retrieval – *clustering, relevance feedback, selection process*; H.2.8 [**Database Management**]: Database applications – *data mining, spatial databases and GIS*

General Terms
Algorithms, Experimentation, Human Factors

Keywords
information extraction, information visualization, event extraction, topic modeling, geoinformatics, spatiotemporal information retrieval, data mining, machine learning, time series

1. INTRODUCTION
In this paper, we address the problem of event visualization based on structured data, in the form of time-referenced and georeferenced relational tuples, and on unstructured data, in the form of free text. Information extraction systems based on named entity recognition (NER) and relationship extraction have enabled detection of events mentioned in free text and extraction of structured tuples describing the location, time, along with other attributes of an event. Identifying hotspots and trends, however, remains an open problem. One limitation is the absence of ground truth for high event activity. In some cases this is due to a lack of well-defined criteria for activity and relevance, while in some it is due to limitations in existing annotation interfaces.

We first present a basic approach to event visualization. Our general framework makes use of mapping tools such as *Google Maps* [1], the Google web toolkit, and timeline visualization tools such as *MIT SIMILE* [2]. It also builds upon previous work on gazetteer-based event recognition and syntactic patterns for semantic relationship detection. Next, we show how a system developed originally for visualization of animal disease outbreaks reported in online news documents can be adapted to display reports of methamphetamine lab seizures compiled by regional law enforcement. We briefly outline the development of a domain-specific data description language for increased portability and ease of information integration. We then discuss the role of topic modeling and information retrieval approaches in filtering and ranking events.

A key technical contribution of this work is the application of topic modeling algorithms in order to compute the posterior probability of a particular spatial location, time unit, or combination given the type of event, which is treated as a topic. This allows the data to be interrogated systematically in order to display geographic regions that are more prone to events of interest. A potential application of this is to construct a time composite map of administrative divisions within a state or province, or a spatial composite time series by month or year, showing active regions. These can be visualized using a *choropleth map*: a map in which regions (geographic regions in this case) are coded by colors or grayscale intensity levels. These represent a variable of interest – in this case, event frequency. Finally, the ability to estimate marginal likelihoods over locations and times given the event type parameters can also be used to filter events, to display only those that fall within a specified frequency range. For example, the system can be configured to search for seizures of methaphetamine production labs in counties or districts where they are common or rare.

2. EVENT VISUALIZATION TASKS
2.1 Spatiotemporal Event Extraction
The goal of event extraction is to identify phenomena related to specific actions, occurrences, relationships, or entities. For example, a positive test for a contagious animal disease on a farm is an event that may be tied to an epidemic and identified *post hoc* as indicating an outbreak. The seizure of equipment from a methamphetamine production facility, or of waste products from a dump site, is an example of an event in the domain of drug

enforcement. Events that can be localized in space and time form the basis of spatiotemporal event extraction.

In the domains of veterinary epidemiology and drug enforcement, decision support systems are typically based on spatiotemporal event extraction and visualization. When events are already available in structured form, they are usually compiled manually from investigative reports by local or national authorities: animal health agencies in the case of veterinary epidemiology and state of national bureaus of investigation in the case of drug seizures. By contrast, unstructured data often comes into the decision support system as the result of a web crawl based on domain-specific resource identifiers: seeds (URLs) and search terms. Federated displays and user interfaces for these decision support systems often combine event data from structured data repositories with data extracted from free text. This entails data integration challenges such as disambiguation, deduplication and identity uncertainty for entities and events; expansion of existing named entity sets from gazetteers (lists of known entities); and inference of attributes for relationships representing events and entities representing actors and objects. These topics are beyond the scope of this paper; we refer the interested reader to existing literature on the current state of the field in domain-focused relationship extraction.

Instead, we suppose here that some preliminary classification step has already taken place to identify the entity that serves as anchor point for an event, and that further classification or inference has identified a putative time and location for the event. Whether this is accomplished through supervised inductive learning from text corpora or as a result of basic pattern matching, our starting point is a candidate tuple to be analyzed, considered for presentation to a decision-maker or search user, and if selected, visualized in the context of events of interest.

2.2 Georeferencing and Map Visualization

Mapping out spatially-referenced events, even using structured data sources, entails a straighforward but data-intensive georeferencing task: looking up the coordinates (latitude and longitude) of street addresses and postal codes where events are reported to have occurred.

The resulting coordinates are placed into a spatial database management system (SDBMS) for visualization using software libraries and services such as *Google* Maps, as shown in Figure 1. For this purpose, we developed two alternative access layers with a unified representation and geographic information system (GIS) data model. The first layer is based on Google's Keyhole Markup Language (KML) and a file-based application programmer interface (API), while the second layer is based on a PHP interface to a MySQL database implementing the KML schema. Our front-end application, *TimeMap*, can be configured to use either layer.

2.3 Timeline Visualization

Figure 1 depicts the data integration between the map and timeline visualization subsystems. The seizure event in April, 2010 is represented on the map by a pop-up note, on the monthly scale timeline (upper right) by a circled dot, and on the yearly scale timeline (lower right) by a circled point.

2.4 Thematic Mapping

The object of thematic mapping is to depict phenomena and trends in a geospatial context. Toward this end, we have added a mapping overlay to the *TimeMap* framework that allows map transformation such as superimposition of data and transparency layers to be applied using the Google Maps API. This includes choropleth maps with dynamically computable color palettes.

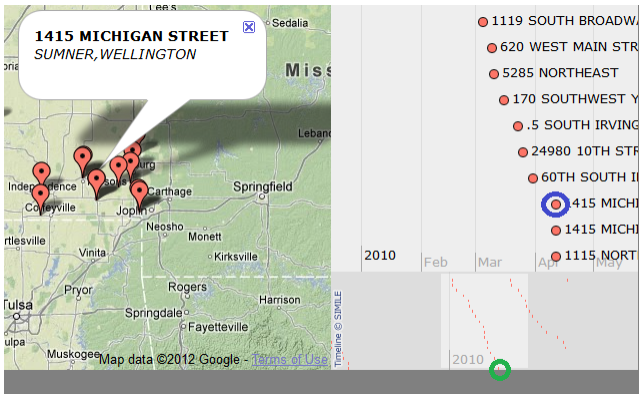


Figure 1. Map and timeline visualization of meth lab seizure events (2004-2011) using *Google Maps* and *MIT SIMILE*. Seizures from the first half of 2010 are depicted, with one event selected.

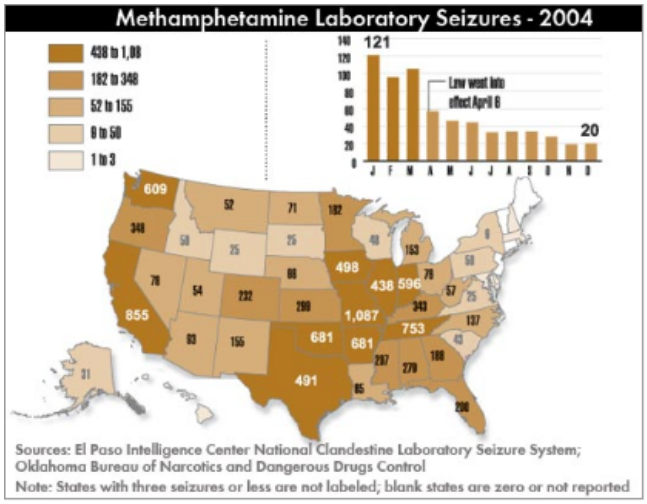


Figure 2. Choropleth map of 2004 meth lab seizures in the USA. (Associated Press, 2005)

Figure 2 is a choropleth map depicting meth lab seizures by state in 2004. [3] This map, which was published by the Associated Press and featured on the ABC News web site in 2005, does not take state population or intrastate population distribution into account. More importantly for information retrieval applications, it does not provide any drill-down interface *cf. HealthMap* [4] or similar event visualization services. One of the reasons for the development of the geospatial visualization components of *TimeMap* was to facilitate information retrieval and multimodal information access using well-established visualization techniques such as thematic mapping and small multiples.

2.5 Spatial Time Series Prediction

A final rationale for the *TimeMap* visualization framework arises from domain-specific data mining objectives in epidemiology and criminology. Governmental agencies devoted to agriculture, public health, and law enforcement often encounter a need for predictive analytics tools to assist with decision-making in both public policy and intervention, and with civic outreach. In the domain of public health, tools such as *HealthMap* [4] have begun to do for individual citizens what more general crime-mapping systems are intended do for search users: provide relevance filters

based on criteria related to incident frequency, corroborative reporting, and significance.

3. TOPIC MODELING

As mentioned in Section 2.1, named entity recognition combined with date and location can provide a means of extracting a stream of events and updates from news stories. This also holds for microblogs and other social media. In order to classify new events and detect the emergence or **revival** of new event-related topics, however, a mechanism for monitoring update streams is needed. This requires a more flexible topic model than the fixed or expandable sets of named entities used for structured information extraction. Furthermore, the frequency and semantic heterogeneity of event reporting from multiple media outlets, even on the web alone, may require enrichment of the parameters beyond those used in classical generative models for information retrieval. We now examine possible extensions to these document clustering models.

3.1 Static Topic Models

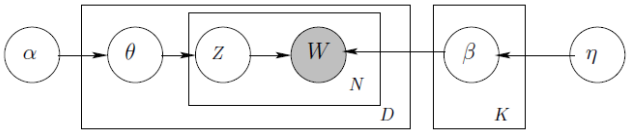


Figure 3. Plate model for Latent Dirichlet Allocation (LDA) in a system with an *N*-word lexicon, *D* documents, and *K* topics.

Figure 3 illustrates the kind of generative Bayesian topic model widely used to cluster static collections of documents. Here, θ is a topic distribution for a document, while Z is the topic sampled from θ for word W. β is a Markov matrix giving the word distribution per topic, and η is the Dirichlet prior parameter used in generating that matrix.

3.2 Dynamic Topic Models

As our preliminary experiments with historical data on both epizootic disease outbreaks and meth lab seizures showed, news flashes do not admit the kind of stationarity assumed in Figure 3. Specifically, the latent variables of our topic model change over time as a result of concept drift and the arrival of new topics, which we can think of as a birth-death process tied to observable events. Blei and Lafferty (2006) proposed a dynamic topic model with fixed topic count K in which each topic's word distribution and popularity are linked over time. [5] Meanwhile, older topic modeling algorithms such as Latent Semantic Analysis (LSA) [6] that permit K to vary suffer from problems such as proximity of different senses of a polysemous word, while variants Probabilistic Latent Semantic Analysis (PLSA) [7] exhibit parameter growth linear in the number of documents D.

3.2.1 Discrete Time, Infinite Topic

Ahmed and Xing (2010) proposed a partial solution to this problem by introducing an infnite Dynamic Topic Model (iDTM) that allows for an unbounded number of topics and an evolving representation of topics according to a Markovian dynamics. [8] They analyzed the birth and evolution of topics in the neural computation community based on the *Neural Information Processing Systems (NIPS)* conference proceedings. Their model evolved topics over discrete epochs (time units). All proceedings of a conference meeting fall into the same epoch. This model does not suit less "bursty topic" applications such as meth lab seizures or disease outbreaks, which are asynchronous whether reported in the news or in local law enforcement records.

For topic modeling applications such as event visualization, such discrete time model may be too brittle. An extension to continuous time will give it the needed flexibility to account for variability and change in temporal granularity.

3.2.2 Continuous Time, Finite Topic

Meanwhile, Wang *et al.* (2008) proposed a continuous time dynamic topic model that uses Brownian motion to simulate the evolution of topics over time. [9] Although this model uses a novel, sparse variational Kalman filtering algorithm for fast inference, the number of topics it samples from is bounded, which severely limits its application in news feed storyline creation and article aggregation. When the number of topics covered by the news feed is fewer than the pre-tuned number of topics K specified in the model, similar stories will appear under different headlines. On the other hand, if the number of topics covered becomes greater than the preset number of topics, topics and headlines will get conflated.

3.2.3 Proposed: Continuous Time, Infinite Topic

To accommodate the needs of non-bursty news updates in domains such as our event visualization domains, we propose a hybridization of the infinite topic model and the continuous time model that combines a hierarchical Dirichlet process (DP) for dynamic topic abstraction and refinement with Brownian motion to capture stochastic topic drift. [10]

We can use a variational inference algorithm such as variational Kalman filtering to factorize the variational distribution over latent variables:

$$q(\beta_{1:T,z_{1:T},1:N}, \theta_{1:T} | \hat{\beta}, \phi, \gamma) = \prod_{k=1}^{K} q(\beta_{1,k}, \ldots, \beta_{T,k} | \hat{\beta}_{1,k}, \ldots, \hat{\beta}_{T,k}) \times \prod_{t=1}^{T} \left(q(\theta_t | \gamma_t) \prod_{n=1}^{N_t} q(z_{t,n} | \phi_{t,n}) \right)$$

where β is the word distribution over topics and $\beta_{1:T,z_{1:T},1:N}$ is the word distribution over topics for time 1:T, topic $z_{1:T}$ and word index 1:N, where N is the size of the document lexicon.

4. APPLICATION TEST BED

4.1 Prior Work: Veterinary Epidemiology

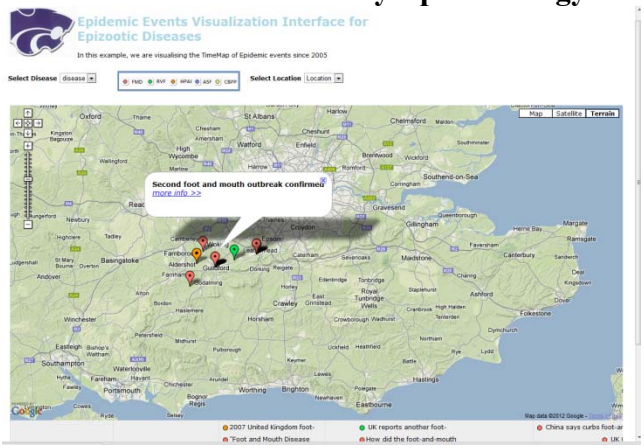


Figure 4. Kansas Information Integration and Analysis Center (KIIAC) for epizootic diseases.

Volkova & Hsu (2010) describes earlier work on computational information and knowledge management (CIKM) and information extraction. [11] This research, motivated by a need to visualize digests of news articles on animal disease outbreaks as shown in Figure 4, led to the earliest prototype of our event visualization system, implemented using *Google Maps* and *MIT SIMILE*. This system used syntactic detectors for semantic equivalence assertions (Volkova *et al.*, 2010). [12]

4.2 Kansas Meth Lab Seizures

A more recent version of the event visualization system is represented in the meth lab application described in Sections 1 and 2. This system forms the test bed for both the visualization techniques described in Section 2 and the topic modeling techniques intended to raise the precision of the federated system by improving relevance filtering and ranking.

5. EXPERIMENTAL EVALUATION

Figure 5 shows some baseline descriptive statistics for meth lab seizures from the test bed discussed in Sections 2 and 4.2.

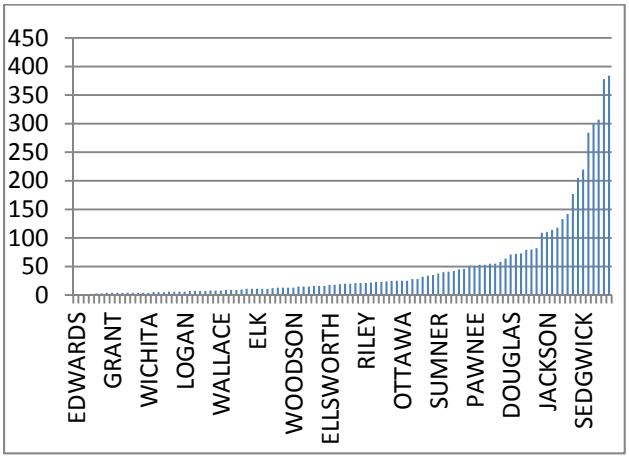


Figure 5. Column graph of the 4942 total meth lab seizures in Kansas, 2000 - 2011, by county (104 with seizures).

In our topic model, topics are event types, which are 50 different types of methamphetamine lab seizures. Given the text of a lab seizure report or a news story of the event with its date and location as a prior, we can use the topic model shown in Figure 3 to evaluate the likelihood of the event given the prior. Events with likelihood value above a threshold are considered highly likely to occur given the location and time of the event.

Table 1. Topic "Abandoned dump site" proportion per seizure reports for four Kansas counties.

	2000	2002	2004	2006	2008	2010
Cowley	**0.0345**	0.0188	0.0188	0.0182	0.0001	0.0175
Crawford	0.0185	0.0172	0.0188	0.0185	0.0182	0.0188
Cherokee	0.0185	0.0175	0.0175	0.0178	0.0181	**0.0333**
Reno	**0.0350**	0.0172	**0.0344**	0.0166	**0.0527**	0.0172

Given a collection of police lab seizure reports with date and location, we ran the topic model on it and evaluated the topic composition of each report in the collection. Table 1 shows the topic "Abandoned dump site" proportion per seizure reports for four Kansas counties over six years. If we set a likelihood threshold value of 0.02, then for year 2000 the event will be marked on the map for Cowley and Reno counties. The event will not get marked on the map for year 2002 for any of these four counties, and for year 2004 will be marked only in Reno County, and so on.

6. CONTINUING AND FUTURE WORK

In continuing work, we are validating the baseline LDA output by using it to filter and rank search results the seizure database represented in Figure 5 and Table 1. Relevance feedback from multiple subject matter experts is being used to evaluate both LDA and DP-based topic models. [11]

7. ACKNOWLEDGMENTS

Thanks to Loretta Wyrick-Severin (Kansas Bureau of Investigation) for assistance with public information requests and to Surya Teja Kallumadi, Tim Weninger, Svitlana Volkova, John Drouhard, Landon Fowles, and Andrew Berggren for development work on *TimeMap*.

8. REFERENCES

[1] Google. (2010). *Google Maps*. Retrieved June 29, 2012, from https://maps.google.com

[2] Massachusetts Institute of Technology. (2008). Retrieved June 29, 2012, from SIMILE: Semantic Interoperability of Metadata and Information in unLike Environments: http://simile.mit.edu/wiki/Main_Page

[3] Associated Press. (2005, November 1). *Meth Stats*. Retrieved June 29, 2012, from ABC News: http://abcn.ws/NYggv5

[4] Brownstein, J., & Feifeld, C. (2007). *HealthMap – Global Disease Alert Mapping System*. Retrieved January 25, 2010, from http://www.healthmap.org

[5] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of ICML 2006*, (pp. 113-120). ACM Press.

[6] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

[7] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1-2), 177-196.

[8] Ahmed, A., & Xing, E. P. (2010). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *Proceedings of UAI 2010*, (pp. 20-29). AUAI Press.

[9] Wang, C., Blei, D. M., & Heckerman, D. (2008). Continuous time dynamic topic models. *Proceedings of UAI 2008*, (pp. 579-586). AUAI Press.

[10] Blei, D. M. (2012, April). Introduction to probabilistic topic models. *Communications of the ACM, 55*(4), pp. 77-84.

[11] Volkova, S., & Hsu, W. H. (2010). Computational knowledge and information management in veterinary epidemiology. *Proceedings of ISI 2010*, (pp. 120-125). IEEE Press.

[12] Volkova, S., Caragea, D., Hsu, W. H., Drouhard, J., & Fowles, L. (2010). Boosting Biomedical Entity Extraction by using Syntactic Patterns for Semantic Relation Discovery. *Proceedings of WI-IAT 2010* (pp. 272 - 278). IEEE Press.

Towards Detecting Wikipedia Task Contexts

Hanna Knaeusl
Chair for Information Science
University Regensburg
Germany
hanna.knaeusl@ur.de

David Elsweiler
Chair for Information Science
University Regensburg
Germany
david.elsweiler@ur.de

Bernd Ludwig
Chair for Information Science
University Regensburg
Germany
bernd.ludwig@ur.de

ABSTRACT

Wikipedia is a resource used by many people for many different purposes. We posit that it might be beneficial to alter the content or the way content is presented depending on the task context. Here we describe a small pilot lab study to investigate features of interaction that might help to infer the contextual situation surrounding wikipedia search tasks. We describe our effort to collect data and analyse relationships between the features and the assigned task context.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Preference Elicitation, Info Seeking Behaviour

Keywords

Eyetracking, Wikipedia

1. INTRODUCTION

Information portals such as Wikipedia represent rich sources of information covering an incredibly broad range of topics. Many Wikipedia entries are also long and can cover aspects ranging from overviews and introductions to more detailed descriptions of advanced aspects that are perhaps only suitable for topic experts. Single pages can also contain not only text, but images, info-graphics, lists and navigational information. Previous research suggests that these resources will have several different contexts of use. For example, Marchionini [11] identifies three main types of search tasks, all of which are applicable to Wikipedia: *Lookup* tasks include finding answers to specific questions, known-item searches or navigating to specific pages. These tasks are contrasted with *exploratory search* tasks, which include *learn* tasks, where the aim is to acquire larger amounts of knowledge and achieve an enhanced understanding of a given topic, and *investigate* tasks, where the user makes use of found information and continues to contribute to or generate knowledge in some way. Elsweiler et al. [4] provide an additional task dimension, distinguishing between work-oriented tasks where information is required to complete some job and casual-leisure tasks, where the aim is more pleasure-focused, e.g. to pass time, to relax, to be entertained etc.

Wikipedia contributors are encouraged to create pages in a way that meets the needs of as many users as possible by including information on a topic with sufficient quantity, quality and completeness and structuring the content in a way that makes sense generally. Nevertheless, one could imagine that different content or different presentations of the same content might be more suitable in specific contexts. For example, lookup tasks may be best supported when facts in an article are presented as a list that can be scanned easily. In such scenarios, content such as images may be less helpful and perhaps even distracting. Contrastingly, in casual-leisure situations, users may want to focus on multimedia content or have information presented in a way that encourages browsing and information discovery.

We believe examples like this suggest there may be benefit in moving away from static pages, which try to cater for all usage situations, to dynamic pages that are generated appropriately based on the context of use. As a first step towards exploring this hypothesis, in this paper, we investigate how the context of use – the task type being performed – might be detected automatically from user-interactions with the system. We want to establish if the way the user interacts with the system, e.g. his mouse and keyboard interactions, eye movements, and click behaviour can provide implicit feedback regarding the usage scenario and user goals.

With this aim in mind, we present a small pilot study that allows us to evaluate a methodology for detecting the features of interaction that might help us infer the contextual situation surrounding a user's search task. We collect interaction data in the context of a controlled laboratory study and analyse relationships between the features of interaction and the assigned task context. The data show that for the small number of users in our study, the behaviour exhibited when completing tasks of different types is very different; users interact with different types of content in different ways. Further, we provide evidence that it is possible, at least for some users, to predict these behaviours based purely on mouse and keyboard interactions.

2. RELATED WORK

In the IR community a large amount of work has been performed to establish if interaction data can be used as a surrogate for explicit relevance judgements. This is known as implicit relevance feedback. Early research in this area demonstrated a correlation between the time spent reading a

action	label	description
Read	RE	User is reading text
Scan	SC	User scans content e.g. headlines, lists or whole page
Examine	EX	User examines element
Navigate	NV	User navigates

element	label	element	label
Headline	HD	Text passage	TX
List	LI	Introduction	IN
Picture	PI	Info Box	IB
Charts, tables etc.	IG	Links in Wikipedia	WI
Other navigation	ON		

Figure 1: Annotation labels for the user actions during Wikipedia search and for the gazed elements

document and explicit relevance judgements [12]. Although this has been disputed in naturalistic situations [10], White and Kelly show that when task type is taken into account clear signals can be found[16]. Other studies have shown that the amount of scrolling on a Web page [3], click-through for documents in a browser [9], bookmarking behaviour [7] and eye movements during the search [2]can all be used as implicit feedback to improve retrieval performance.

Interaction data can also be used as a means to predict user emotions. For example, Fox et al., show that query log features can be used to predict searcher satisfaction [6] and Feild et al.[5] used interaction data and physical sensors to predict levels of user frustration with high accuracy.

A third group of studies show correlations between different styles of interactions e.g. for some users visual attention on the screen can be predicted via mouse coordinates [15] We believe that the interaction style, the emotional state of the user and the motivating task context will be intrinsically related and that the work done previously suggests it may be possible to predict the task based on interaction data. We explore this in a small pilot study below.

3. DATA COLLECTION

In this section we provide details of the data collected and explain the motivation behind recording the data.

3.1 Study Design

Data was collected via a laboratory based user study with 4 users. The participants were information science students (1 male, 3 female) aged between 20 and 30. All of the participants were experienced wikipedia users and were comfortable using the wikipedia search facilities. Although this user population is not large or diverse enough to provide generalisable results, it is sufficient for our aims, which were to evaluate and improve the methodology and get a sense for the feasibility of our ideas.

Each participant performed 6 Wikipedia search tasks (2 of each of the 3 types of interest - lookup, learn and casual-leisure). The tasks were presented in the form of a simulated scenario and were ordered randomly to minimise learning effects. Example tasks for each type are shown in Figure 2.

After initially greeting the participant, the experimental procedure was explained in person. Then, to prevent biases, the participant was led automatically through the experiment on screen, with task descriptions, questionnaires and a web-browser window appearing when appropriate. The experimenters observed the tasks remotely in an adjoining room, where the participant's screen was mirrored.

lookup

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	0	0	0	42	0	0	0
NV	0	0	0	0	0	0	46	0	0
RE	0	0	0	23	0	23	0	27	0
SC	53	0	0	24	18	0	0	59	12

learn

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	89	0	0	93	0	0	0
NV	0	2	0	0	0	0	52	0	0
RE	1872	0	0	72	0	0	0	93	0
SC	172	0	6	2	0	0	0	62	285

casual-leisure

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	137	0	2	85	0	0	0
NV	0	11	0	0	0	0	105	0	0
RE	1876	0	6	274	1	0	0	90	32
SC	177	0	2	8	6	0	0	60	134

Table 1: Absolute frequencies of content elements for actions for the investigated task types

3.2 Data Collected

We collected a large amount of data from each participant before, during and after the study.

Questionnaires: A pre-study questionnaire collected demographics, search experience, and experience with wikipedia of the participants. Pre-and post-task questionnaires elicited perceptions of the task and domain knowledge, of success and the experience including emotional aspects, and finally a post-study questionnaire provided general impressions of the experiment.

Eyetracking Data: We recorded participant gaze patterns using an SMI RED eye-tracker. The associated BeGaze software recorded videos files of screen interactions with an additional layer indicating the area of the screen where the user is focusing his gaze. We manually annotated these complete overlaid video sequences with two labels. The first describes what the user is doing ("action"). This is a simple coding scheme but aligns with reading psychology research [14, 13]. It was the annotator who decided which action to code at what moment by following the focus displayed in the layer on top of the recorded screen. The second label describes the content ("element") being focused on and is derived from the elements available in Wikipedia pages. The label was assigned when the focussed on an area on the screen so long that the annotator could assume the element in the area was perceived.The full set of labels for actions and elements is presented in Fig. 1. The intuition behind the labels was that the style of reading for different task types and the content elements used will be very different. By labelling videos in this way we could test this intuition empirically.

Browser Logs: We instrumented the firefox web-browser to log all user interactions during the search process.

Timestamp information was used to align interaction data from different sensors.

Lookup: Last night you watched a documentary about the sinking of the Titanic. Suddenly you wonder how many passengers were on board when the catastrophe happened. Search in Wikipedia for this information.

Learn: Friends from abroad are visiting Germany and you plan to travel together to visit the small but beautiful city of Regensburg. As preparation for the trip you want to know more about the city and its history. Use Wikipedia to do this.

Casual-leisure: You have a few minutes before your class starts but you are already sitting in the lecture hall. Kill this time using wikipedia using the next six minutes to look at whatever topic(s) take your fancy.

Figure 2: Examples of the kinds of tasks assigned to study participants.

4. EVALUATION OF THE DATA

We analyse the data in two stages. First, in Section 4.1, we examine the distribution of video labels for different types of task to determine if users behave differently or focus their attention on different kinds of topics when completing different task types. Second, in Section 4.2, we show how these labels can, in turn, be predicted using interaction data from the eyetracker and browser. The first stage provides evidence that the user's preferences for content elements depends on the search task, endorsing our suggestion to customise web pages at run time. The second stage provides some evidence for our hypothesis that the interactions a user performs in a browser may be used to predict which actions he trying to complete and which content elements he is preferring at that moment.

	LO vs RE		**LO vs CA**		**RE vs CA**	
action	χ^2	p-value	χ^2	p-value	χ^2	p-value
EX	9	0.011	9	0.029	18	0.006
NV	9	0.011	9	0.011	18	0.001
RE	13	0.043	6	0.301	27	0.079
SC	36.563	0.064	45	0.039	45	0.039

Table 2: χ^2-tests for Different Distributions of Content Elements per Task Type (LO: lookup, RE: learn, CA: casual-leisure)

4.1 Reading Style and Content for Task-types

Technical difficulties meant we were only able to work with data for 6 casual-leisure, 4 lookup tasks and 4 learn tasks. We first divided the data into 500ms frames, allowing us to normalise the counts by task length, and counted relative frequencies of frames for which label combinations occur for each task type (see Table 1). Visually inspecting the distribution of content for actions, suggests the reading style and the elements of content interacted with were very different in different task contexts. This is confirmed by pair-wise comparisons using chi-squared tests for the distributions content elements for each possible pair of task types (see Table 2).

Examining the results in Table 2, we observe that all but one combination of action type shows highly significant differences in the distribution of content elements examined. The exception is the distribution of elements for lookup and casual-leisure tasks, which initially seems counterintuitive, as one would expect these two tasks to be very different. Below we summarise the main similarities and differences between the task-types and attempt to explain what these mean in the context of our work.

When completing lookup tasks, the participants do not typically read content, the exception being page introductions. Instead they scan large portions of the page very quickly, looking for the snippets of information that will satisfy their specific information need. They tend to scan a number of different kinds of content elements during tasks. This can be seen from Table 1 with counts being spread over text passages, introduction, info boxes, lists and headers. Images are noticeably missing from lookup tasks. It seems as if the participants have decided that for the tasks assigned, images will not useful and are able to avoid them.

Learn and casual-leisure differ from lookup tasks in that they both tend to be longer in time and have more interactions. They also both involve reading actions, which were rare for lookup. By this we mean that the user focuses attention on whole passages of text and attends the text from left to right and line by line. Another similarity between learn and casual-leisure tasks is the way that text passages are consumed, with the counts for these tasks being very similar. There are differences between learn and casual-leisure tasks, particularly in terms of the elements used other than text passages. During learn tasks the focus tended to be on headers, while for casual leisure, the focus was on elements such as introductions and info boxes, which allow the user to gain an overview of what a page is about and allow them to judge whether it is interesting or not. We assume that headers are useful for learn tasks because here there is a concrete information need i.e. users do not just need to find something that is interesting or not, but need specific informational content. In this sense headers will help the user determine whether a paragraph is worth reading or not.

4.2 Predicting Style and Content Preferences

To determine if the manually assigned labels can be predicted from interaction data alone, we calculated statistics for counts of the synchronous occurrences of video labels and input events for the 500ms frames introduced above. As we were searching for the simplest features possible (so they could eventually be computed easily during a browser session at runtime) we used the frequencies of the most common mouse events and the average saccade distance (i.e. eye movement) per frame as features. More precisely, for each frame we descretised these features into two levels: *low* and *high* based on the mean value over all frames.

Table 3 (left) gives an example for the information we computed from the raw log data. In order to understand whether the knowledge of the `mousemove` frequency is relevant for predicting user actions and content elements, we performed a series of χ^2-squared tests for all six search tasks for one of the test participants chosen at random (in total about 30 minutes of interaction). The results are reported in Table 3(right). With the exception of the rare `click` events, all features are highly significant. We interpret this as a positive indication that for individual users – depending on their personal interaction style (see [1, 8]) – it is feasible that the reading behaviour label could be predicted during a brows-

action	mousemove high	mousemove low	element	mousemove high	mousemove low
NV	5	6	IN	30	12
RE	18	5	IB	8	10
SC	41	18	WI	5	6
			LI	21	1

Task	scroll action	scroll el.	click act.	click el.	mousemove act.	mousemove el.	avg.sacc.dist act.	avg.sacc.dist el.
1	***	***	***	***	***	***	***	**
2	***	***				*	***	***
3	*	**		*	***	***	* **	
4	*	***	*			**		***
5	***	***			***	***	*	
6	***	***	**		***	***	**	***

Table 3: Frequency counts of user actions and `mousemove` events and of content elements and `mousemove` events occurring simultaneously (left). The table on the right shows the significance results for χ^2-squared tests.

ing session. The results of the χ^2-squared tests indicate that knowing at run-time whether the observed input events occur below or above average at any point of time increases the accuracy of predicting the video labels as annotated for that moment as the distribution $P(\text{action}|event = \text{low})$ differs significantly from the distribution $P(\text{action}|event = \text{high})$ for any annotated action and for any annotated element type. This oberservation opens the way for runtime prediction of the user action and preferred elements. From that information, the system can predict the current task type and use this information for generating content dynamically.

5. CONCLUSIONS

The preliminary data analysis we have presented provides clues that, firstly, reading behaviour and preferences for content elements depend on the surrounding task context and, secondly, both behaviour and preferences may be predicted for individual users based on their interaction style.

There are several limitations to this work. That we only have data from four participants from a relatively homogenous group means we cannot generalise. However, we claim that the presented methodology is well suited to address our long term research questions outlined in the introduction and the pilot has provided us with insight into how to improve a full study. In addition to resolving several technical challenges, we have learned that the great care will need to be taken when simulating tasks. For example, were few images looked at in lookup tasks, simply because of the tasks we chose? We also plan to look at more complicated prediction features and account for the fact that individual differences in participants (cognitive, reading style [14]) will exists and that users interact in different ways (people who follow eye movements with their mouse, people who don't) [15]. At EuroHCIR, we look forward to engaging with the broader HCI and IR communities to discuss the ideas in this paper; we are particularly eager to receive feedback on the next steps along this research path, including brainstorming solutions to some of the empirical design challenges of running such experiments and identifying and dealing with the many factors which should be incorporated in the full study.

6. REFERENCES

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

[2] G. Buscher, A. Dengel, and L. Van Elst. Eye movements as implicit relevance feedback. In *CHI'08: Extended Abstracts on Human Factors in Computing Systems*, page 2991–2996, 2008.

[3] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the IUI*, page 33–40, 2001.

[4] D. Elsweiler, M. L. Wilson, and B. Kirkegaard Lunn. *New Directions in Information Behaviour*, chapter Understanding Casual-leisure Information Behaviour. Emerald Publishing, 2011.

[5] H. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proc of SIGIR 2010,*, 2010.

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

[7] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceedings of SIGIR*, pages 130–137, 2010.

[8] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *Proceedings of CHI*, CHI '12, pages 1341–1350, New York, NY, USA, 2012. ACM.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinki, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inform. Syst.*, 25(2), 2007.

[10] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of SIGIR*, page 408–409, 2001.

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

[12] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieva. In *Proceedings of SIGIR*, pages 272—281, 1994.

[13] J. Nielsen. *Designing Web Usability*. New Riders, Berkeley, Calif., 2006.

[14] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psych. Bull*, 124(3):372—422, 1998.

[15] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. In *SIGIR Workshop on Web Information Seeking and Interaction*, pages 29–32, 2007.

[16] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM 2006*, page 297–306, 2006.

CUES: Cognitive Usability Evaluation System

Matthew Pike
Department of Computer Science
Swansea University, UK
matpike@gmail.com

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

Anna Divoli & Alyona Medelyan
Pingar Research
anna.divoli@pingar.com
alyona.medelyan@pingar.com

ABSTRACT
A Cognitive Usability Evaluation System, CUES, was constructed to allow the simple integration of cognitive data from a commercialized EEG brain scanner, with other common usability measures, such as interaction logs, screen capture, and think aloud. CUES was iteratively evaluated with a small number of participants to understand whether and how the visualisation of EEG data alongside other measures, provided value for usability evaluation. Results indicate that although there are a lot of objective measurements available from the brain scanner, the largest value came from qualitatively identifying EEG patterns, and correlating them with think aloud data. Recommendations for using CUES and for future developments are both provided.

Categories and Subject Descriptors
H5.2. Information interfaces and presentation (User Interfaces): evaluation/methodology, screen design.

Keywords
Information Seeking, Cognitive Load Theory, EEG, Usability

1. INTRODUCTION
Evaluation of user interfaces is typically restricted by what can be observed in specifically designed experimental environments or through fieldwork. Aside from objective measures like time to complete a task, researchers use questionnaires, interviews, think-aloud protocols, and subjective observations to determine how satisfied or frustrated the users are with a particular interface. In some way, mouse movements, eye-tracking patterns, or differences in heart rate can indicate emotional state of the subject, but arguably, looking at the brain activity directly would be more effective and accurate. Different neuro-imaging devices and electromagnetic brain scanners have been recently introduced as tools that can assist interface evaluation [1, 5] and they were found to be accurate [2, 3]. In 2011, Wilson argued that brain scanning devices might be useful for evaluating search user interfaces and their impact on a user's cognitive load [6].

In this paper, we introduce the Cognitive Usability Evaluation System, or CUES, as a universal tool to integrate cognitive EEG data with other standard usability measurements. CUES can be used to run studies with multiple participants and capture various data that may assist researchers in performing the evaluation. CUES is designed to capture brain activity, as returned by an off-the shelf EEG-device Emotiv EPOC[1]. In addition, CUES visualizes the captured outputs as shown in Figure 1, such as mouse movements (callout #2), audio (#3), and EEG data (#5).

To our knowledge, CUES is the first usability evaluation system that features a brain scanning device as an integral part. However, others have reported experimental results of using alternative devices for measuring brain activity as users perform specific tasks. Kitamura *et al* used fMRI outputs to show that after repeating the task of learning how to use chopsticks the neural activity patterns indeed indicate learning [2]. Cernea *et al* [1] used the same EEG-device as the one used in CUES, the EPOC[1]. Their goal was to evaluate EPOC's accuracy as it predicts users' facial expressions (smiling, blinking) and their emotional state (calmness, excitement, engagement, frustration). Cernea *et al* found that EPOC's predictions are accurate in 70% to 100% cases, with the exception of excitement but concluded that it may be a hard to define excitement as an intrinsically mental activity. Vi and Subramanian [5] were able to accurately detect confusion created by user interface design using the EPOC.

Despite providing difficult to use, and often noisy data, these overall positive experiences of using brain scanners as research tools, as well as the lack of functioning systems for running user studies, have motivated us to build CUES. The following sections describe CUES and provide a formative study of the value provided by the EEG data.

Figure 1: A screenshot of the CUES Visualiser.

2. CUES
CUES is a collection of applications that allows researchers to manage, automate and visualise user studies, as described below.

2.1 Study Setup and Recording
CUES provides an intuitive interface for managing study related data, including: participant details, study tasks, study conditions and the data sources to capture during a study. CUES' study setup component provides a range of settings for managing participants in different study conditions. Having configured the study, it can then be "run" within CUES. CUES is designed to capture interactions between participants and web pages. To facilitate this, CUES provides a simple customised web browser, which to a participant appears indifferent from their everyday browser. In the background, however, the browser is capturing: audio, brain data, screenshots, mouse trails, and JavaScript based web events.

Audio is captured from the machines input device (e.g. Microphone). Brain data is acquired from the Emotiv EPOC

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

device. As well as providing the raw Electroencephalography (EEG) signal, CUES collects the EPOC's pre-classified emotions (Engagement, Excitement, Frustration, *etc.*) and facial features (Smile Extent, Frown Extent, *etc.*). Screenshots of the webpage, as seen by the user, are captured at specified intervals, with a full-page capture occurring upon each page load. Finally, JavaScript web events are captured via a custom JavaScript library that is injected into each page by the browser. These events allow CUES to capture user interactions with the web page such as button clicks, highlighted text, data entry etc. All captured data is stored in a suitable, open format (Audio: wav, Screenshots: JPEG, Other: XML) allowing the data to be analysed using other software as well as the Visualiser (described below).

2.2 Visualising the Study

The Visualiser, shown in Figure 1, provides a way of correlating various types of data in a time series. CUES also offers options to customize of the visualisations, such as choosing which brain data and/or web events to include on the timelines. Further, and perhaps most importantly, the evaluator can stack multiple records on top of each other for comparison, allowing them to compare, for example, one participant's performance on multiple tasks, or several participants' performance on a certain task.

A hierarchical tree (#1) is provided for browsing the available recordings, which is ordered by study tasks, conditions and participants. Once selected, each recording is opened within its own tab (#2). Every visualisation within a single recording is linked to the audio waveform display (#3). The waveform visualises the audio captured during the study, and optionally has the ability to be played with sound or muted (useful when comparing many recordings at once).

Brain data are plotted on a 2D graph (#4), and emotions can be selectively added to the graph via the emotion selector (#5). Additionally, web events (such as page loads, mouse overs and mouse clicks) can be selectively added to the graph (#4) through the event selector (#6). Each event is added at the bottom of the chart at the point in time that the event occurred. Hovering over the event's box on the chart gives additional event details (#7).

Finally, there are additional visualisations that utilise the captured screenshots and mouse data. A researcher may select their desired visualisation from the tab component (#8). In Figure 1, we see that the participant's view of the web page at time X is overlaid with their recent mouse trail (#9). Other visualisations include a heat map of the cursor position, trail location on the entire page, the entire page by itself, and the visible region view.

3. FORMATIVE EVALUATION OF CUES

To study the utility of the brain data we adapted the RITE method [4] to iteratively make changes to the methodology as we learned about CUES' capabilities. This process involved reflecting on the utility and value of the data captured after each participant, and trialling alternative configurations, such as: capturing facial expressions with the camera, turning off features, separating or joining the recording of subtasks to find the right level for analysis, and so on. This process allowed us to examine and contrast recommendations for using and improving CUES.

3.1 Procedure and Participants

To create a scenario within which to trial CUES, tasks were designed to evaluate the design of 4 very different taxonomy interfaces: Yippy, CarrotSearch, MeSH, and ESD. Taxonomies like these are a common form of Search User Interface feature. The first two of these systems present automatically generated categorisations of web search results, yet Carrot provides users with alternative visualizations. The last two allow users to browse carefully designed taxonomies aiming at more expert audience. MeSH in particular, is highly specialized and is used mostly for automatic indexing tasks. This variation ensured different reactions from the participants. We chose this particular scenario, as it aligned with our other interests. Our findings about these taxonomy interfaces will be presented in a separate future paper.

Six digital economy graduate students with different backgrounds, including graphical design, geography, and economics, were recruited to take part in the study. Gender was balanced, and age ranged between 22 and 45. Participation involved: 1) consent form (approved by the institution's ethics committee) and setup of the EPOC Emotive EEG scanner, 2) Phase 1: non-interactive brain response to systems' designs, 3) Phase 2: content-agnostic exploration of the systems, 4) Phase 3: applied exploration of the systems, and 5) a final debriefing interview. The applied Phase-3 involved participants searching for content relating to their current research, whereas Phase 2 always began with the initial query: 'Schools'. Participation took 1 hour, where participants were allowed to take breaks from wearing the Emotiv if needed. Participants were given an Amazon voucher for their time.

3.2 Quantitative Analysis

In analysing the system, we first found that certain outputs from the EPOC had more value than others. Frustration, Short Term Excitement (STE), and Engagement were the three emotions that showed most variance during interaction. While Meditation showed almost no variation at any point in the study, Long Term Excitement (LTE) showed some usable variation for recordings of 10 minutes or longer. These were infrequent in the study, and so our analysis focused on Frustration, STE and Engagement.

Although apparently a form of objective measurement, analysing EEG data does not lend itself comfortably to summarisation or statistical comparison. As can be seen in Figure 1, the data varies dramatically throughout a task phase. One may hypothesise that average emotive values would help find the "most exciting" or the "most frustrating" system. However, as can be seen in Figure 3, participation averages tend to approximate with each other as they go through a number of peaks and troughs. Further, from the very first interaction with the system, participation diverges. This divergence in behaviour means that the data at t=20s for one user is based on a completely different interaction for another participant. Consequently, to make a standard comparison, we must take a common event and examine the corresponding data. In our study, this was most obviously represented by Phase 1, in which all participants were shown every UI one at a time, creating data that could be compared both within and between participants.

To further investigate the types of analyses that the CUES Visualiser could support in future developments, we performed some manual analyses of the example data, shown in Figures 2-4.

Statistical Analyses. Figure 2 summarizes the average responses for frustration and STE for three of the participants (p3, p4 and p6). The comparison shows that different systems create varying initial and delayed emotions. MeSH and ESD, for example, create initial peaks of frustration, but drop lower after 20s, while Yippy creates a form of frustration that peaks later. It is possible to take some statistics, with the peak of STE for MeSH being almost significantly highest at t=11s ($F(2)=6.47$, $p=0.056$).

Summarising Data. Figure 3 shows graphs from Phase 2 that compare results from different participants for the same system. We should note that there is some data capture issues in places. The engagement data for participants p4 and p5, for example, are

almost identical and appear to represent missing flat-lined data. Notably, however, general averages across the 3-5 minute tasks were quite even, indicating that averaging the data will not be especially valuable for analysis. Similarly, Figure 4 shows the participants' average emotions throughout Phases 2 and 3 while evaluating the 4 systems. Although we were hoping we'd see relationships between other forms of usability data, such as subjective preferences captured in interviews, we were unable to find any obvious relationships.. In the future, we will investigate other quantitative approaches that might be relatable to other forms of usability measures, such counting the number of EEG graphs' peaks and troughs above and below given thresholds, as well as their scale, and allowing summarisations during certain events or time-periods, rather than for entire tasks.

3.3 Qualitative Analysis

Of all the data comparisons above, it is very difficult to draw any conclusions about 'average data' having much value, because average data across an entire task means very little. Even averaging across participants at a given time is difficult, when interaction diverges. In our experience, however, the most valuable insights gathered from the brain data were in watching for patterns in the signal curves and investigating the correlated subjective data, such as the think-aloud data and the mouse trails, for additional insights. This combination was much more valuable than the other combinations we tried, such as recording the facial expressions with a camera. This valuable qualitative process involved two approaches, described below.

Approach 1: Validating Think Aloud. This first approach involved playing back the brain, think-aloud, and mouse trail data in real time, which allowed us to qualify utterances in the think-aloud approach. For example, using think-aloud alone, there were many occasions where participants would utter a comment indicating that they did not understand something. Using levels of frustration and engagement, we could clearly see which of these occasions was creating a significant barrier to use, and which were unimportant. Further, we could identify possible reasons for silence during the verbal-protocol, with some peaking in frustration when, for example, pages were not loading. Other occasions were silences during peaks of STE and -engagement.

Approach 2: Event Detection. This approach involved a more predictive style. After determining common patterns, described below, we were focusing on these patterns as we analysed each participant's brain data. As content was playing back in real time, we could 'see ahead' which parts of the system the user would find confusing or when the user was about to figure something out. Beyond giving us these specific insights, the patterns also allowed us to examine the times of high confusion; or to examine the times of effective progress.

Common EEG patterns:

- High frustration and low excitement
  - often indicating confusion
- A peak of frustration followed by a peak of excitement
  - often indicating comprehension
- Low excitement and frustration, with high engagement
  - often indicating effective progress
- Low frustration and high excitement
  - often indicating (good) discovery

4. DISCUSSION

Overall, we experimented with both quantitative and qualitative data captured by CUES, as well as approaches to analysing them using CUES. Overwhelmingly, we found that the best value provided by the brain scanner was in qualitative analysis, where the data allowed us to a) augment the verbal protocol, b) see ahead of the verbal protocol, and c) explore and examine specific parts of the verbal protocol. In each of these cases, we found it extremely helpful to also see the user's view, mouse trail, and logged interactions.

Despite appearing as a quantitative source, the qualitative value gained from augmenting other more common usability metrics. The specific added value came in two areas. First, the brain data provided additional insight and context into the content of the verbal protocol, which is otherwise often ambiguous and open to the interpretation of the investigator. Second, the brain data added a visual dimension to the verbal protocol, which is what allowed us to both see ahead and specifically explore the data.

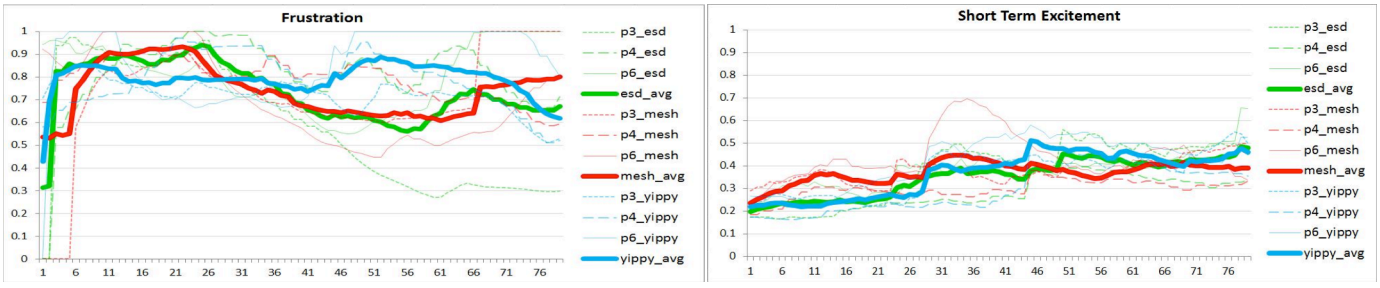


Figure 2: Initial response time-curves in the first 45s of seeing a UI

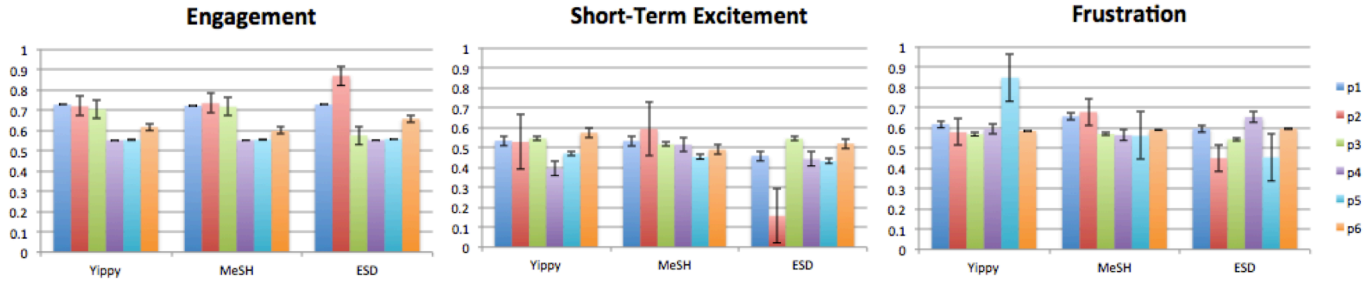


Figure 3: Consistency between users in Phase 2

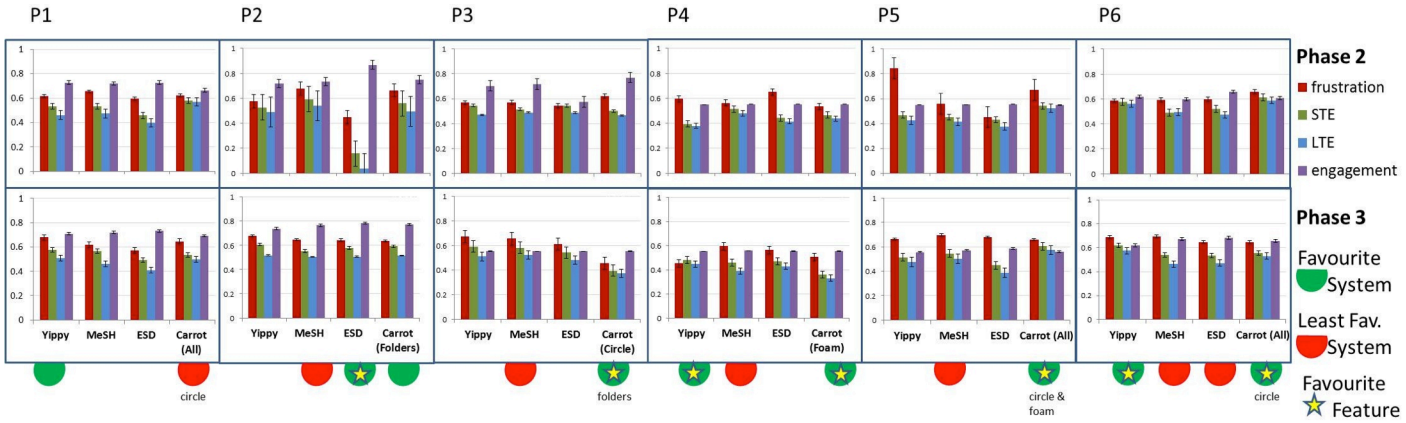


Figure 4: Internal Consistency between users in Phases 2 and 3. Favourite and least favourite system and systems with favourite features are also shown (based on the interview questions at the end of the study).

4.1 Limitations

Despite finding a lot of value in analyzing the EPOC data qualitatively, there are still some well-known limitations to using EEG data. First and foremost, EEG data is easily confounded by body movement. The motor control of fingers, hands, and arms, for example, can create noisy data and arbitrary peaks. In CUES, however, the cross-validation in the think-aloud and brain data allows for some of this noise to be ignored. So far, however, we have not specifically measured body movement.

Further, we frequently saw, especially during the interviews, frustration correlate with speaking. Although it seems like a verbal protocol may, therefore, completely mask the data, we found it was times when participants had to think and explain what was happening. In this case, the verbal protocol often made frustration and lack of understanding more visible in the system.

There are many other limitations to the study, which was only a formative investigation into the utility of CUES , using a scenario focused on evaluating a single form Search User Interface feature. We plan to run a much larger hypothesis-driven evaluation of CUES in the future.

4.2 Recommendations for using CUES

Good data. Despite concerns, we were able to get good data regardless of hair length, etc. However, one must watch out for flat-lined data from one or 2 bad sensors, which leads to data loss.

Waiting for data. We discovered that there is a 10s lead time as certain pre-classified measures begin to show. Short tasks, such as visual exposure, need to be extended to include this lead time. LTE required tasks must be 10+ minutes long to have value.

Comfort. We learnt that participants could wear the device for sustained periods of time. Some participants experienced mild discomfort after wearing the device for more than 40 minutes.

Task Chunking. Correctly separating out tasks is important. If you want to compare a person's response to System A versus System B, they must be in separate recordings to facilitate easy comparison and analysis.

4.3 CUES Improvements

Conducting the study allowed us to identify several areas for improving CUES. One feature of the system captures an entire website, rather than just the page view, but this created an unusual page load event that, in turn, created artificial levels of frustration in our first participant. To be useful, this element needs an alternative implementation to have no visual effect on the user. Further, we also wished to separate the viewport capture frame rate from the mouse data, as the current mouse trail was limited to the frame rate chosen for screen capture. In order to avoid data loss, it was suggested that a warning appear during tasks when any of the EPOC sensors lost its signal. In regards to the Visualiser, greater control was desired to easily see all the elements when stacking several records on top of each other. In this paper, we also explored alternative visualisation and analyses, which we hope to integrate in the future. Feedback also indicated that global controls, rather than per record, were desirable, to avoid constant reconfiguration from the default. Finally, the motion and control over the viewport and playback is currently tied; future versions will allow independent control.

5. CONCLUSIONS

This paper described CUES, a prototype system designed to utilise cheap off the shelf EEG brain scanners to help run usability studies. A formative evaluation provided many insights into the value of different features. Despite being primarily objective in nature, we found that the EEG data was most effective when analysed qualitatively in parallel with think-aloud data. The EEG data a) helped to validate or qualify ambiguous think aloud comments, and b) added a visual dimension to the verbal protocol allowing us to look ahead at their experience and explore the data for certain events. Ultimately, we conclude that a lot of value can be gained from using CUES to investigate EEG brain measurements in parallel with other usability measures such as logs, screen captures, and think-aloud protocols.

6. REFERENCES

[1] Cernea, D., Olech, P.-S., Ebert, A. and Kerren, A., EEG-Based Measurement of Subjective Parameters in Evaluations. In *HCII'11 - Posters*, 279-283. 2011

[2] Kitamura, Y., Yamaguchi, Y., Hiroshi, I., Kishino, F. and Kawato, M., Things happening in the brain while humans learn to use new tools. In *CHI'03*, 417-424. 2003

[3] Liu, Y., Sourina, O. and Nguyen, M.K., Real-Time EEG-Based Human Emotion Recognition and Visualization. In *CW'10*, 262-269. 2010

[4] Medlock, M., Wixon, D., Terrano, M., Romero, R. and Fulton, B., Using the RITE method to improve products; a definition and a case study. In *Usability Professionals Association*. 2002

[5] Vi, C. and Subramanian, S., Detecting error-related negativity for interaction design. In *CHI'12*, 493-502. 2012

[6] Wilson, M.L., Evaluating the Cognitive Impact of Search User Interface Design Decisions. In *EuroHCIR 2011*, 27-30. 2011

Collaborative environment of the PROMISE infrastructure: an "ELEGantt" approach

Marco Angelini
Sapienza University of Rome
Italy
angelini@dis.uniroma1.it

Claudio Bartolini
HP Labs
USA
claudio.bartolini@hp.com

Gregorio Convertino
Xerox Research
Centre,Europe
convertino@xrce.xerox.com

Guido Granato
Sapienza University of Rome
Italy
granato@dis.uniroma1.it

Preben Hansen
SICS
Sweden
preben@sics.se

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

ABSTRACT

This paper focuses on developing lightweight tools for knowledge sharing and collaboration by communities of practice operating in the field of information retrieval. The paper contributes a motivating scenario, a characterization of these communities, a list of requirements for collaboration, and then a system design proposed as a proof-of-concept implementation that is being evaluated.

1. INTRODUCTION

This paper focuses on the problem of supporting knowledge sharing and collaboration in communities of practice that operate in the field of information retrieval (IR). These communities include developers, researchers, and stakeholders who periodically collect and use scientific data produced by the experimental evaluation of IR systems. Specifically, the communities considered include those involved in three specific IR domains: Patent, Cultural Heritage, and Radiology.

The research context of the work reported in this paper is the PROMISE NoE. This project aims at advancing the current tools for IR communities to perform experimental evaluation of complex multimedia and multilingual information systems. The ultimate goal of the project is to develop a unified infrastructure for the community to efficiently collect and reuse data, knowledge, tools, methodologies, and communities of end users. In this context, providing adequate support for collaboration is crucial. Herefrom the specific goal of the work reported in this paper: designing and evaluating lightweight support for knowledge sharing and collaboration.

Currently, the following problems result from lack of suitable collaboration tools:

1) Greater **effort** is required by individual members, who contribute as volunteers, for sharing knowledge and collaborating. In the long term, this discourages broader participation.

2) Poor **reuse of content** and process information across the multiple instantiations of similar experimental evaluation processes. Over time, this leads to inefficient processes: e.g., content is always recreated from scratch, successful processes (best practices) cannot be reused, novices cannot be easily trained based on shared experience.

3) The overall community cannot easily reflect on (and thus re-engineer) its own **workflow** around specific TRECs.

2. MOTIVATING SCENARIO

The starting point of our analysis is a typical IR evaluation campaign (lab). In a typical scenario, Adam (lab organizer) is preparing an IR experiment and evaluation task and spends time and resources for coordinating, communicating and assembling people and resources in order to proceed with the overall evaluation task, e.g. recruiting people that will be responsible for different evaluation task(s). Communication and sharing of information may be different within different across sub-tasks. Furthermore, they may be different between labs without any awareness among actors of the similarities/differences in the evaluation task processes. Thus, it is important to identify the stages in the evaluation task process as well as how collaborative and information sharing activities are manifested.

3. CHARACTERIZING IR COMMUNITIES

The CLEF experimental platform involves a series of CLEF Labs and one or more tracks within each Lab. Each Lab as well as each track involves a certain set of tasks that could be considered as a task process or workflow. In order to define and describe these tasks we have investigated the lab and track organizers of a CLEF experiment, how they performed their work and what steps they went through during their work. Furthermore, we have extracted requirements for collaborative information handling and information sharing activities specifically [3, 5]. An evaluation campaign is an activity intended to support IR researchers providing a large test collection and uniform scoring procedures. An evaluation campaign is organized within an evaluation framework like TREC or CLEF and can involve different domains (cultural heritage, patent, radiology and so on). Within an evaluation campaign there are many tracks, such as multimedia, multilingual, text, music, images, etc. A track can be organized differently according to a specific domain and include, in turn, several tasks. A task is used to define the structure of the experiment, specifying a set of documents, a set of topics, and a relevance assessment. For each task the set of documents can be structured defining, for example, a title, keywords, images and so on. A topic represents an information need. Documents can be assessed as being relevant or not (or more or less relevant) for a given information need (topic).

Some of the most common tasks that we observed as part of a

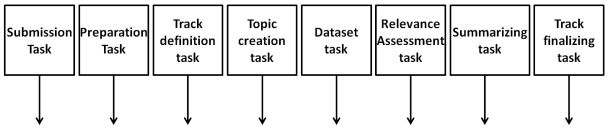


Figure 1: Task stages of organizing an experimental IR track in CLEF.

typical evaluation campaign include the followings: submission, preparation, track definition, topic creation, data set, relevance assessment, summarizing, and finalizing.

3.1 Observed roles in IR Communities

Within an evaluation campaign many people are involved in different tasks, such as organizing, creating topics, managing collections, handling participants and submission, choosing measures, and running the final evaluations.
The set of actors involved in PROMISE activities is not homogeneous and depends on the domain which is taken into account. Looking at three domains (patent, medical, and cultural heritage), we defined the following actors:
- **organizers**: people who are in charge of preparing a campaign; it is possible to distinguish domain organizers and track organizers;
- **participants**: people who run their algorithm(s) according to the actual tasks;
- **relevance assessors**: people who make the relevance assessment;
- **topic creators**: people who define topics for a given task;
- **site administrators**: (e.g. system administrators);
- **other researchers**
- **annotators**: people who annotate resources to highlight some hidden information.
Each of these actors is described along with a set of activities or tasks. Moreover, a user can have more than one role.

3.2 Observed tools of IR Communities

The collaborative work may be performed in a non-structured manner using basic tools for collaboration in an *ad hoc* fashion.
The following are the most commonly used tools for collaboration:
- **E-mail**: the most common way to organize the work and spread information;
- **Face-to-face meetings**: useful to discuss more effectively about problems and solve them; it is complex if people don't work in the same building;
- **Video conference tools** (e.g. Skype): used instead of face-to-face communication;
- **Shared workspaces** (e.g. shared document editor, desktop sharing): useful to share documents. However an issue may arise where many people work on the same document.

4. COLLABORATION REQUIREMENTS

As mentioned in the introduction, a more suitable collaborative tool is needed to help researchers to accomplish their tasks. To realize it we have to overcome some limitations.
The first one is the impossibility to define a common detailed workflow due to the presence of different domains, each of them with specific needs. This makes it difficult to realize a collaborative environment completely specified. Despite this limit, it is possible to individuate some common needs such as: communication with other actors, access to data of previous campaign, sharing of task flow of actual evaluation campaign, and sharing workspace with actors involved in the same tasks.

Another aspect that characterizes the work of people involved in a lab is that there is an alternation between individual and collaborative work, which is in contrast to a too rigid environment.
The basic idea of our system is to improve the actual tools used in IR community without defining the collaborative environment in a too rigid way. Our purposes in this paper are:
1) Collecting the tools actually used (e.g. Skype, Googledocs, etc.) in a structured environment. 2) Making available to users other collaborative tools (polls, news). 3) **T@GZ**: a social software system for organizational information sharing.

4.1 Requirements by role

In order to get information on the actual tasks users performed, we made requirements elicitation through a number of questionnaires to track organizers within the three predefined domains of Patent, Cultural Heritage, and Radiology in the CLEF platform. Working through these questions we identified: a) different **roles** of actors involved in the CLEF experiments, b) **requirements** for collaborative information handling activities and information sharing and c) **links** between roles and collaborative events.
Furthermore we describe each task stage regarding subtask involved (fig.1):
- **Submission task**: preparing a lab proposition including sub tracks, acceptance of the lab or not.
- **Preparation task**: preparation of a CLEF lab flyer including details of each lab, obtaining databases, preparing a copyright agreement, preparation of the web page.
- **Track definition task**: definition of broad tasks, start of registration for the participants.
- **Topic creation task**: preparation of detailed topics for each task, release of topics, checks of the copyright forms.
- **Data set**: data access for all registered participants with signed copyright forms for their tracks.
- **Relevance judgment task**: preparation of the judgement system; finding qualified judges; submission of all runs by participants; pooling of the runs to create documents to judge; judgement of the documents for relevance; evaluation of all submitted runs.
- **Summarizing task**: release of ground truth; submission of participants' papers with results and technical descriptions; analysis of the results and submission of overview paper.
- **Finalizing task**: CLEF workshop and labs with discussion; feedbacks on CLEF; preparation of next year of CLEF; distribution of responsibilities.

4.2 Implicit requirements

Support the community in **managing the process**:
- **Tasks and roles**. Various roles are involved in communities work on experimental evaluation of IR systems. Multiple tracks and tasks are part of a process occurring in an evaluation experiment such as CLEF. Within a track, some of the tasks are interdepend. Specific roles perform specific tasks. One member can play multiple roles. The set of the role and tasks changes across different IR domains.
- Assume both individual and collaborative works. Many tasks are conducted individually and the individual work is interleaved with collaborative work. Support both individual and collaborative works and faster transitions.
- Only some of the steps in the work-flow are **fully specified** before or at the outset of the process, others are defined during the process.
Needs related to **existing work tools**:
- As for other communities of knowledge workers [1], the members of these communities use email as their primary communication tool. It would be helpful for any new tool for knowledge sharing and collaboration to build on the central role of email.

- Clear added value. The user should be required to log onto a new system only if such new system supports additional functions that are useful and are not already supported in email or other general-purpose media already in use (e.g. VoIP).
- Difficult tracking and reuse. Perhaps a useful role can be played in facilitating a smoother integration across the multiple existing tools.

Needs related to **new collaborative functions** that might support collaboration:
- **Groups**. Group creation is tied to the task.
- **Polls**. The polls component should be integrated with the other components (this is an example of function not available in email).
- **Collaborative workspace**. It is desirable for the members of the community to have easy access to the shared resources and typical steps, which, ideally, should be made available all in one place.
- **Process visualization**. We observed visible differences in the different instantiations of the work-flows elicited from different sub-communities. It allows people to become aware of differences and similarities in the way sub-communities go about performing the same process.

5. CONTEXT AND RELATED WORK

Recent research has pointed to the importance of investigating and supporting collaboration in the field of Information Retrieval (IR). For example, [4] reviews the studies of collaboration relative to this field and concludes that the IR field needs to better understand and improve the systems that support both direct and indirect collaboration during information tasks. This is supported by studies in various IR settings. [6] investigated patent engineers, which is a specific community conducting IR processes, and found that they were involved in various collaborative activities. Overall, existing studies have observed that collaboration is indeed endemic to the broader activities of individuals who perform information seeking, searching, and retrieval tasks. However, with our work we aim to address two specific limitations of the existing literature. First, the prior studies of collaborative practices in IR have focused on describing the practices of teams or groups of users in different settings and user communities (e.g., academia, industry, medicine, patent offices; see [4] or have developed new tools to support these practices [8], but have not yet systematically investigated how to support collaboration at the level of a large communities of practice. Second, while there has been research on how to support communities of practices of professionals (e.g., [10]) or scientists (e.g., collaboratories, see [6], we focus specifically on how to support lightweight knowledge sharing and collaboration in the community of IR researchers and professionals who develop and evaluate IR tools. This is a community with unique needs for collaboration and types of workflows: recurrently, specific sub-communities of volunteers need to agree on, build, and refine evaluation campaigns for testing IR systems.

A key distinctive property of the IR communities is that their workflows cannot be fully specified a priori. That is, if we consider a continuum from highly specified to highly unspecified processes, then we could classify the instances of workflows of the IR communities as intermediate cases along this continuum. [2], who named this continuum the Specificity Frontier, observed a gap between two existing approaches for supporting collaboration: most collaborative systems have focused on either automating fixed work processes (e.g., Enterprise Resource Planning tools) or simply supporting communication in ad-hoc processes (e.g., email). To adequately support the collaboration in IR communities we need to bridge the gap between these two approaches using lightweight tools that are compatible semi-structured workflows. While the specific instances of these workflows share several of the tasks and roles, the specific instances will inevitably vary across IR domains (and data types), evaluation campaigns, and over time (because the process is refined by the IR community in a collaborative manner as it is repeated over the years). Interestingly, recent research on communities of professionals pointed to the same need for collaborative tools that are able to support flexible realizations of the processes rather than forcing the community into hard-coded processes [9].

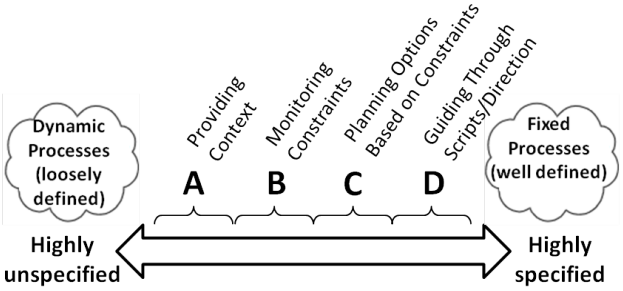


Figure 2: Specificity Frontier

Articulating further the design requirements for supporting semi-structured workflows, [2] divides the specificity frontier into four sub-spectra: providing context for enactment, monitoring constraints about the task, providing/planning options to reach a goal, and guiding through a given script. Building on this classification, in this paper we focus on providing support to the IR communities in the first two sub-spectra.

6. DESIGN AND ARCHITECTURE

To fulfill the requirements described on Section 4, we devised the architecture shown on Figure 3. It refers to the classical CLEF experiments organization, that is arranged in terms of different domains (e.g., Medical, Patents, etc.). For each domain one or more tracks are available. As described on Section 3, the organization of a track is a complex collaborative process and encompasses several tasks that exhibit some precedence relationships. The task flow of a track is formalized using an **E**xtended **L**ight liv**E** **Gantt** (EleGantt) and is shown within the CUI to the user, acting as the main entry point for collaborative activities (Figure 4). A suitable administrative interface allows for adapting the task flow of a track to procedural changes. According to [2] (Figure 2), EleGantt is in charge of providing a part of the process context, i.e., a structured to-do list. The second part of the context, i.e., a shared common space, is provided by T@GZ [7]. EleGantt is an extension of the traditional Gantt chart. It allows for:
- **attaching a rich set of meta-data** to tasks with the goal of supporting collaborative activities: involved people, involved roles, associated tags, kind of collaboration activities needed to accomplish the task, and the list of other processes that share the same activity;
- **expressing the non overlapping constraint** between tasks that must be executed in sequence;
- **specifying temporal uncertainties** (e.g., minimum and maximum duration of an activity), and degree of freedom for milestones and deliverable releases. Moreover, the EleGantt visualization is both a visualization of the task flow and an interactive interface that allows for exploring and accessing task flow associated information, like roles, people, similarities with other task flows, etc.

T@GZ is a social software system for organizational information sharing. In T@GZ the user can share by simply sending an

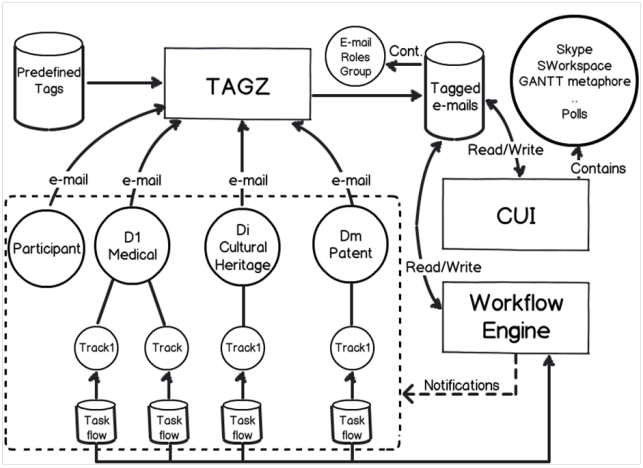


Figure 3: The architecture of the Promise collaborative system

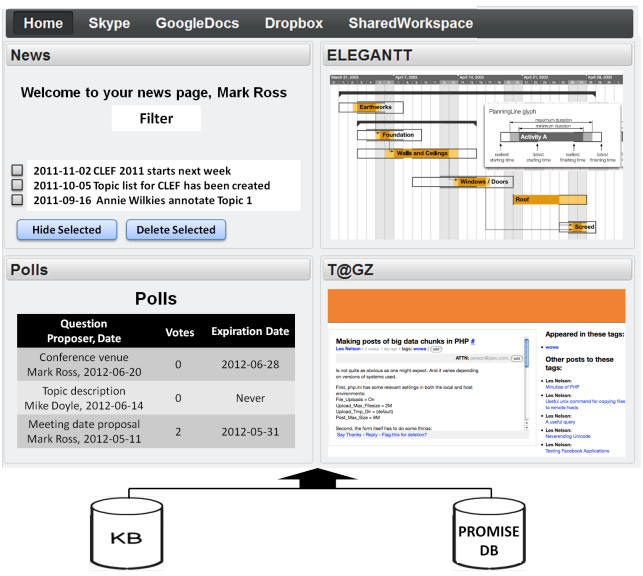


Figure 4: The Promise Collaborative User Interface (CUI)

email message with the content to be shared, and addressing the message to one or more topic specific keywords. For example, one might use the address, bizdev@share.X.com, for referring to information related to "business development" topic (see Figure 4, top left). Thus, the content of that email is 'tagged' by the keyword 'bizdev'. Any mail may have multiple tags attached in this manner, in the 'To' or 'CC' fields, using any client. While enabling easy publishing and re-finding of this information, the system does not induce people to send additional emails other than those that they are already sharing. Focusing now on the implementation of T@GZ in the whole system, using a set of predefined tags (i.e., the tags associated to the elegantt' s tasks), T@GZ provides a means for indexing the emails that are exchanged among the organizers of the tracks, including links to smart attachments. The work-flow engine is aware of the elegantts and using time information and inspecting the KB sends through email different kinds of notifications (e.g., a deadline is approaching, it is time to move to the next step, etc.) to people involved in the tracks organization.

The Collaborative User Interface (CUI) is the Web based access point to all the collaborative activities (see figure 4). It is basically split in two subcomponents. The first one allows for managing personal user collaborative information (left part of the picture), e.g., messages, polls, etc. The second one refers to the whole process and allows for both exploring it using EleGantt, discovering people, roles, and tags and browsing the whole set of tagged emails. Moreover, the CUI contains a set of tabs that allows for accessing the collaborative tools that have been specified in the EleGantt. In order to provide the user with an unique access point to the process resources, the CUI sends tagged emails to T@GZ, containing a link to the collaborative resources (e.g., link to dropbox folder or to Google Docs document)

7. CONCLUSION AND FUTURE WORK

As result of our investigation we identified some general challenges or open issues for this domain: 1) find a good balance between the need for flexibility to fit various, partially-defined processes and the need for enough specification in order to allow automation 2) identify the set of predefined tags (some common and some domain specific) 3) semi-automate the tagging process (e.g., an intelligent assistant).

Acknowledgements

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

8. REFERENCES

[1] V. Bellotti, N. Ducheneaut, M. Howard, I. Smith, and R. Grinter. Quality versus quantity: e-mail-centric task management and its relation with overload. *Human-Computer Interaction archive*, 20(1), June 2005.

[2] A. Bernstein. How can cooperative work tools support dynamic group process? bridging the specificity frontier. . In *ACM conference on Computer supported cooperative work (CSCW '00)*, 2000.

[3] M. Croce, E. Di Reto, G. Granato, P. Hansen, A. Sabetta, G. Santucci, and F. Veltri. Collaborative User Interface Requirements. In *PROMISE deliverable D5.1*, 2011.

[4] J. . Foster. Collaborative information seeking and retrieval. In *Annual Review of Information Science and Technology, Volume 40, 2006, 329Ű356*, 2006.

[5] P. Hansen, G. L. Granato, and G. Santucci. Collecting and Assessing collaborative requirements. In *Collaborative Information Seeking (CIS 2011) workshop*, 2011.

[6] P. Hansen and K. Jarvelin. *Collaborative information retrieval in an information- intensive domain. Information Processing and Management (IPM).* 2005.

[7] P. M. Joshi, C. Bartolini, and S. Graupner. T@gz: intuitive and effortless categorization and sharing of email conversations. In A. Mille and F. L. e. a. Gandon, editors, *WWW(Companion Volume)*, page 365. ACM, 2012.

[8] M. Morris and E. Horvitz. SearchTogether: an interface for collaborative web search. In *20th annual ACM symposium on User interface software and technology (UIST '07).*, 2007.

[9] H. R. Motahari-Nezhad, C. Bartolini, S. Graupner, S. Singhal, and S. Spence. IT Support Conversation Manager: A Conversation-Centered Approach and Tool for Managing Best Practice IT Processes. In *Hewlett Packard Laboratories Palo Alto,USA,2010*, 2010.

[10] R. S. W. . Wenger, E.; McDermott. *Cultivating Communities of Practice* . Harvard Business Press, 2002.

Search User Interface Design for Children: Challenges and Solutions

Tatiana Gossen
Faculty of Computer Science,
Otto-von-Guericke-University,
Germany
tatiana.gossen@ovgu.de

Marcus Nitsche
Faculty of Computer Science,
Otto-von-Guericke-University,
Germany
marcus.nitsche@ovgu.de

Andreas Nürnberger
Faculty of Computer Science,
Otto-von-Guericke-University,
Germany
andreas.nuernberger@ovgu.de

ABSTRACT

In this paper we describe the main challenges in designing search user interfaces for children. Young users require emotional support, language support, memory and cognitive support, interaction support and support to judge document relevance. We discuss possible solutions for each challenge. We also present a working prototype of a web search interface whose main target group are users of primary school age. Our interface is colourful and voice supported, contains possibilities for both searching through text input and browsing in menu categories, has a guidance avatar for emotional support and a result storage functionality to support children's cognitive recall.

Keywords

Web Search Engine, Children, Search User Interface.

Categories and Subject Descriptors

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

General Terms

Design, Human Factors, Management.

1. INTRODUCTION

In times of digital natives more and more children are going online. According to a recent report [8], children of ages five to nine spend about 28 minutes online daily and this time continuously grows. The German 2010 KIM[1] study [17] reports that about 60% of the German children of ages six to thirteen use the Internet and 70% of those use search engines. Children are using the Internet for different purposes, especially for entertainment like online games or watching videos on *Youtube*, for communication and for information search, e.g. related to their school activities [17].

In the modern society, finding information in the Internet is an important skill that a child needs to develop. If a child succeeds in finding the information, it feels competent and develops self-confidence. In contrast, if it is not able to find good results, a child may develop a feeling of incompetence. That could even lead to a feeling of inferiority, especially in the "industry versus inferiority" period of child's psychosocial development (age 6–12) [5]. Children's immaturity in the emotional domain is not the only aspect that is different from adult users. Children's cognitive abilities are also not fully formed [21]. Thus, children do not have the same abilities and knowledge as adults and constitute a separate user group. The special characteristics of children are challenging and should be considered by the development of web search engines, including the design of web search user interfaces (UIs).

In order to support children in their search, special search engines for children, have been launched, e.g. *kidrex.org*, *onekey.com*, *askkids.com*, *kidsclick.org*, *dipty.com*, *blinde-kuh.de*, *fragfinn.de*, *helles-koepfchen.de*, *quinturakids.com* etc. Currently, their main purpose is helping children to find *only* child appropriate content in the WWW. Another important aspect is the usability of those search engines. It is of importance that search engines for children match the particular skills of children in order to increase their usability for children. Unfortunately, current search engines for children not always match the skills and abilities of children [6].

The aim of this work is to develop a novel web search UI which meets the needs of children, i.e. fits their cognitive abilities, knowledge and provides the necessary emotional support. This interface should support children in their search in a web document collection. Our primary focus is textual information retrieval, as web documents mostly have a textual form and are written in natural languages. When designing tools for children, there is a need to target very narrow age groups [19]. Cognitive abilities and knowledge of a fourteen years old and a seven years old child strongly differ. In this paper we concentrate on primary school age children as in our opinion this user group is the most challenging one. In the following we underline challenges in the design of web search user interfaces for young users and present possible design solutions.

2. DESIGN CHALLENGES & SOLUTIONS

Emotional Support: Based on Erickson's theory of psychosocial development [5] children require emotional support and a feeling of success. This can be achieved by proper guidance. The idea here is to provide children with enough help to support their search process in order to avoid frustration. We propose building a guidance avatar that captures children's failures, e.g. getting no results or spelling mistakes, and explain how to do better.

[1]KIM is a German acronym for Children and Media ("Kinder + Medien, Computer + Internet"). It is a German user study which is regularly conducted in the form of interviews.

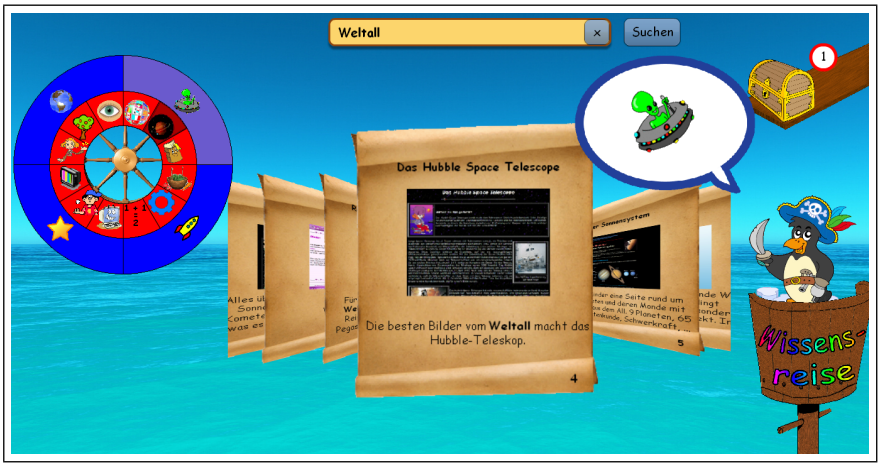


Figure 1: Screenshot of the *Knowledge Journey* user interface: a guidance figure and a treasure chest on the right hand side, query input elements on the top, a menu with many categories on the left hand side and a coverflow with search results in the middle.

Language Support: Children, especially in the primary school age, read slowly and are still learning to write [23]. In addition, children have a limited domain knowledge [11] and difficulties with typing using a keyboard [22]. This results in problems with query formulation and spelling errors [2, 7]. Therefore, a search UI for children should provide different possibilities for children to formulate their information need. We suggest using a browsing menu with many categories which meet children's information needs. This menu should be image based and audio supported in order to navigate ergonomically and fast within it. Besides the browsing, we also suggest to provide the opportunity of keyword-oriented search supported by spelling correction mechanisms. Children can choose the way they want to start searching. With an increasing domain knowledge (possibly gained from browsing in categories) children can employ keyword-oriented search more efficient.

Cognitive Support: According to theories of human cognitive development, human development occurs in a sequential order in which later knowledge, abilities and skills build upon the previous ones [20]. Piaget [21] describes four development stages. Children in primary school age are in the concrete operational stage of their development which is characterized as a stage where children learn to reason logically and have difficulties with thinking abstractly. Their understanding is limited to concrete and physical concepts. Therefore, categories used in the menu should not be abstract and browsing menu should have a flat hierarchical structure. Metaphors used in the user interface should be familiar to children and have a connection to the physical world (this is also advised in [3]).

Memory Support: According to the information processing theory [13], information processing of children differs from the adults' in terms of how they apply information and what memory limits they have, i.e. children can represent and process less information than adults. Information retrieval processes may cause children's memory to overload. This explains children's "looping" behaviour during the information seeking process. Children click, repeat searches and revisit the same result web page more often than adults do [2, 7]. To support children's cognitive recall we can provide a result storage functionality. It is also important to show a clear back-button or just present the search result in the same window (e.g. using frames) in order for children not to get lost.

Interaction Support: The information processing rate influences the fine motor skills of children [4, 10]. Young children's performance in pointing movements, e.g. using a mouse, are lower than that of adults. Therefore, the search user interface should prefer simple point-and-click interactions and clickable interface elements should be large enough to be easily hit [3].

Relevance Support: Children also have difficulties to judge the relevance of the retrieved documents to their information need [12]. Children are frustrated by too many results and do not have the ability to determine the most relevant and "best" documents [14]. A child-suitable form of results presentation can support children's judgement of results' relevance and provide relevance clues. Each result item should have a website image and its description. Akkersdijk et al. [1] also suggest displaying the results using a *Coverflow* technique where the user navigates horizontally. Coverflow allows users to concentrate on one item at a time. It also does not require complex interactions like scrolling as a vertical results list used in common search engines.

3. SEARCH INTERFACE

We considered the requirements for user interface design and developed a search user interface for children called *Knowledge Journey (KJ)*. We used multimedia elements in the UI design to make the appearance attractive for children. We also took into account that all clickable items are of appropriate size. We used font sizes larger or equal to 14 pt as advised in [3].

Our search user interface KJ uses the metaphor of a treasure hunt where a user takes a journey to gather relevant search results. The interface of KJ is shown in Fig. 1. It consists of five groups of elements: a guidance avatar (here a penguin pirate), a treasure chest, a coverflow, elements for keyword search and a pie-menu for browsing. In the following we are going to describe each element group.

3.1 Guidance Avatar

In order to start a "Knowledge Journey" a child selects a guidance avatar (see Fig. 2a). The avatar concept is familiar to children from computer games. It allows individual user personalization, e.g. girls can select a female pirate or penguin, there are also figures for younger and older users. The guidance avatar supports children's search process in order to avoid frustration: in the current version it supports children by providing a spelling correction after a misspelled query is submitted (see Fig. 2b) and enlarges images of menu categories providing animations (Fig. 1). A further possible function of the guidance avatar is an explanation how to search and what to do in case of finding no results.

(a)

(b)

Figure 2: Screenshot of the user interface: select which pirate accompanies you by the *Knowledge Journey* (a) and guidance avatar makes a suggestion by a misspelled query (b).

3.2 Browsing Menu

In order to support children who have difficulties to formulate a query, a browsing menu with many categories is designed. There exist different types of menus. We used a pie menu as it can be operated with simple point and click interactions and presents a good overview of categories. The pie menu is placed on a steering wheel. We use the metaphor that a steering wheel is used to define the search coordinates to provide a search direction. Initially top categories of the menu are shown (see Fig. 3, middle). We choose menu categories like entertainment, sports and hobbies, history, universe, geography, nature, persons etc., as they meet the information needs of children described in [16]. Each category has a number of subcategories. Children are comfortable to use a two-level hierarchical organized menu for browsing [11]. Corresponding subcategories are opened when a child clicks on a top category.

Mousing over the category triggers an action of a guidance avatar, i.e. it shows a large animation to explain the category. Icons and animations are used to indicate categories because images better match the cognitive skills of children than written words [9]. They also make the user interface more attractive for children as they prefer colourful designs with multimedia content [18, 15, 3]. In addition, we provide voice support. By placing a mouse long enough on the pie menu item, a voice explanation is played telling what category is selected. Users can also hide the menu by clicking in the middle of it. Then, only the wheel is shown (see Fig. 3, left). The menu can be opened again by clicking on the wheel. If a child clicks a category it receives results visualized as a coverflow. The category name is also placed as a text in the search input field.

3.3 Results Presentation

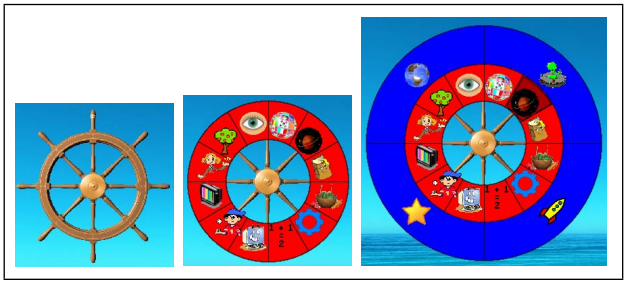


Figure 3: Screenshot of the user interface: browsing menu on a steering wheel in three different levels (closed, opened, opened with 2nd hierarchy level).

The result presentation is shown in Fig. 1. We use a coverflow where each item is presented on a papyrus roll that contains the webpage's title on top, its thumbnail (preview) in the middle, a textual summary and a result number according to the relevance on the bottom. A child can interact with our coverflow using simple point and click operations. It can open a webpage by clicking on the result item that is in focus or switch to the next or previous page by clicking on an item that is not in focus. The whole papyrus roll area is clickable and thus it is easy to hit.

When designing a search UI for children, search results and links should not be opened in a new window or tab as this inhibits backtracking with the browsers' back button and thus provokes "looping" behaviour. Users can easily get confused or lost and start searching for the way back. We decided to open a webpage in the same window using a frame (see Fig. 4). In order to return to the search a child clicks on the "X"-Button. It can also store a webpage using a "+"-Button.

3.4 Results Storage

A child can store relevant results in the "treasure chest". This form of storage aims to support children's memory to prevent cognitive overload. The number of stored results is shown near the chest. Furthermore, we use physical concepts like the size of the chest to show the amount of "treasure", i.e. a chest icon becomes larger with each additional stored result (compare Fig. 1 and 5). By clicking on the chest, a journey journal opens (Fig. 5). We use a book metaphor, where each reversal of the book contains information about a stored webpage: its thumbnail, a textual summary and a title. A child can add notes to each website. It can also open the website again by clicking on its picture in the book. If a child does not like a website anymore, it can delete it by clicking on the

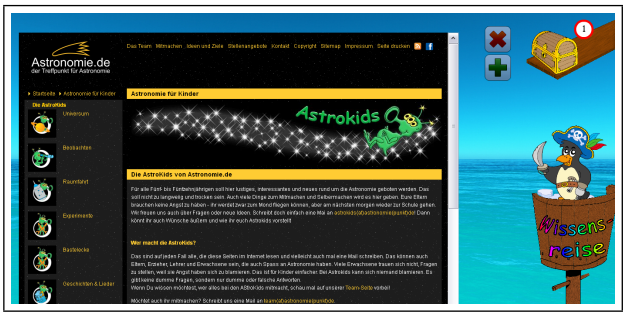


Figure 4: Screenshot of the UI: website opens in a frame.

Figure 5: Screenshot of the user interface: journey journal with favourite web pages.

"-"-Button. Tiles in the form of small website thumbnail (below the journal) are used to navigate within the book.

4. DISCUSSION & OUTLOOK

In this paper we described the challenges when designing search user interfaces for children. We demonstrated possible solutions based on which a novel user interface, called *Knowledge Journey*, was designed. We presented the user interface elements of KJ, i.e. a guidance avatar for emotional support, a treasure chest for memory support, a pie menu for language support and a coverflow to support the judgement of results relevancy. The interface also uses simple interactions to support children's fine motor skills. A comparative user study with 28 young users of age seven to twelve (average 9.5 years) was conducted where we compared our user interface with a *Google-like* UI. We evaluated what features of both interfaces children like most or do not like and the results are promising, i.e. 17 participants preferred KJ interface and five liked both. In the future we are going to do a deep analysis of the study results in order to improve the interface.

5. ACKNOWLEDGEMENTS

We are grateful to Ina Bosse for her support in development. The work presented here was partly supported by the German Ministry of Education and Science (BMBF) within the ViERforES II project, contract no. 01IM10002B.

6. REFERENCES

[1] S. Akkersdijk, M. Brandon, H. Jochmann-Mannak, D. Hiemstra, and T. Huibers. ImagePile: an Alternative for Vertical Results Lists of IR-Systems. *Technical Report TR-CTIT-11-11, Centre for Telematics and Information Technology, University of Twente*, (ISSN 1381-3625), 2011.

[2] D. Bilal and J. Kirby. Differences and similarities in information seeking: children and adults as Web users. *Information Processing & Management*, 38(5):649–670, 2002.

[3] R. Budiu and J. Nielsen. *Usability of Websites for Children: Design Guidelines for Targeting Users Aged 3–12 Years, 2nd edition*. Nielsen Norman Group Report, 2010.

[4] S. Card, T. Moran, and A. Newell. The model human processor- an engineering model of human performance. *Handbook of perception and human performance.*, 2:45–1, 1986.

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

[6] T. Gossen, J. Hempel, and A. Nürnberger. Find it if you can: usability case study of search engines for young users. *Personal and Ubiquitous Computing*, 2012.

[7] T. Gossen, T. Low, and A. Nürnberger. What are the real differences of children's and adults' web search. In *Proc. of the 34th international ACM SIGIR conf. on Research and development in Information*, pages 1115–1116. ACM, 2011.

[8] A. Gutnick, M. Robb, L. Takeuchi, J. Kotler, L. Bernstein, and M. Levine. Always connected: The new digital media habits of young children. The Joan Ganz Cooney Center at Sesame Workshop, 2011.

[9] D. Hackfort. *Studientext Entwicklungspsychologie 1: Theoretisches Bezugssystem, Funktionsbereiche, Interventionsmöglichkeiten*. Vandenhoeck & Ruprecht, 2003.

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

[11] H. Hutchinson, A. Druin, B. B. Bederson, K. Reuter, A. Rose, and A. C. Weeks. How do I find blue books about dogs? The errors and frustrations of young digital library users. In *Proc. of the 11th International Conf. on Human-Computer Interaction (HCII 2005). Mahwah, NJ: Lawrence Erlbaum Associates*, 2005.

[12] H. Jochmann-Mannak, T. Huibers, L. Lentz, and T. Sanders. Children searching information on the Internet: Performance on children's interfaces compared to Google. *SIGIR'10 Workshop on accessible search systems*, pages 27–35, July 2010.

[13] R. Kail. *Children and their development*. Prentice Hall Upper Saddle River, NJ, 2001.

[14] A. Large and J. Beheshti. The Web as a classroom resource: Reactions from the users. *J. of the American Society for Information Science*, 51(12):1069–1080, 2000.

[15] A. Large, J. Beheshti, and T. Rahman. Design criteria for children's Web portals: The users speak out. *J. of the American Society for Information Science and Technology*, 53(2):79–94, 2002.

[16] S. Livingstone. Children's use of the internet: Reflections on the emerging research agenda. *New media & society*, 5(2):147, 2003.

[17] Medienpädagogischer Forschungsverbund Südwest. KIM-Studie 2010. Kinder+ Medien. *Computer+ Internet. Stuttgart*, 2011.

[18] S. Naidu. Evaluating the usability of educational websites for children. *Usability News*, 7(2), 2005.

[19] J. Nielsen. Children's websites: Usability issues in designing for kids. *Jakob Nielsen's Alertbox*, 2010.

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

[21] J. Piaget, B. Inhelder, and B. Inhelder. *The psychology of the child*, volume 5001. Basic Books, 1969.

[22] P. Solomon. Children's information retrieval behavior: A case analysis of an OPAC. *J. of the American Society for Information Science*, 44(5):245–264, 1993.

[23] A. Stuart. When should kids learn to read, write, and do math? WebMD, 2007. Online at `http://children.webmd.com/features/when-should-kids-learn-read-write-math`, accessed 18.07.2012.

EyeGrab: A Gaze-based Game with a Purpose to Enrich Image Context Information

Tina Walber
University of Koblenz-Landau
Institute for Web Science and Technologies
Koblenz, Germany
walber@uni-koblenz.de

Chantal Neuhaus
University of Koblenz-Landau
Institute for Web Science and Technologies
Koblenz, Germany
cneuhaus@uni-koblenz.de

Ansgar Scherp
University of Koblenz-Landau
Institute for Web Science and Technologies
Koblenz, Germany
scherp@unikoblenz.de

ABSTRACT

We present *EyeGrab*, a game for image classification that is controlled by the users' gaze. The players classify images according to their relevance for a given tag. Besides entertaining the players, the aim is to enrich the image context information to improve the image search in the future. During the game, information about the shown images is collected. It includes the classification concerning the tag, a rating of the given images by the user ("like" or "not like") and the eye tracking information recorded when viewing the images. In this work, we present the design of the game and compare two design variants – one with and one without visual aid – concerning the suitability of the game for image annotation. The variants of the game are evaluated in a study with 24 participants. We measured the user satisfaction, efficiency and effectiveness of the game. Overall, 83% of the users enjoyed playing the game. The results show that the visual aid is not helping the users in our application; it even increases the error rate. The best classification precision we achieve is 92% for the game variant without visual aid.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Input devices and strategies

General Terms

Human Factors

Keywords

Eye tracking, Game with purpose

1. INTRODUCTION

The search for digital images is still a challenging task. It is often performed based on context information, e.g. tags describing the image content. Tags assigned to specific image regions instead of the image can improve the search results [9]. Also the ratings of the images can be used to deliver good search results, as it is done on some image stocking pages like Photo.net.

The game *EyeGrab* has been developed as a game with a purpose (GWAP) to improve or collect these information: the description by tags, a personal rating and information about image regions. The game is controlled by eye movements, which on the one hand enhances the user satisfaction and on the other hand allows for collecting gaze information that will be analyzed to gain information about the image content. The overall goal of EyeGrab is to enrich the images with contextual information in order to improve future search tasks.

The players look at images falling down the screen. The task is to classify the images as relevant or irrelevant to a given tag. Relevant images are rated by the participants into "I like it" or "I do not like it". We have compared two different interaction designs of the game in a study with 24 subjects and measured effectiveness, efficiency, and satisfaction. While the first variant provides visual aids in form of highlighting the interactive regions and a gaze-cursor, which visualizes the subjects' fixations on the screen, the second variant does not. Overall, we can state that the vast majority of the participants enjoyed playing the game and that the gaze-based control of the game was experienced as an improvement on the entertainment. The players which received the visual aids had the impression to be supported by it. However, the results show that the visual aids lead to significantly more incorrect and missing classifications. In fact, given a ground truth image data set, we have achieved the best classification precision with 92% for the game variant without visual aid.

The satisfaction and the precision of the results gained in our experiment are very satisfactory. Based on this outcome, we will continue with the evaluation and conduct information extraction from the gaze paths in a next step. Based on a prior experiment [8], we can use this gaze information to add region-based annotations to the images.

2. RELATED WORK

A large number of applications were introduced in the past that use eye movements as input medium, often for people with disabilities, e.g., for a drawing system [1]. Also the use of gaze information as relevance feedback in image retrieval was investigated with promising results, e.g. [6]. Walber et al. [8] showed that specific image regions can be identified, using gaze information. The development of sensor hardware like cameras in computers is continuously progressing. Already now eye tracking can be performed with a commodity web camera. San Agustin et al. [7] compare a commercial eye tracker and a webcam system. The results for the webcam system are satisfactory and comparable to the commercial system, although still with limitations concerning the comfort of use. Based on this development, one can assume that eye tracking could be performed for more users in the future and it will be possible to use the technology also in playing games.

Data obtained from eye tracking is less accurate than, e.g., from a computer mouse, due to natural movements of the eyes. It can be

difficult for the users to focus the gaze on a specific region to select a button. One possibility for supporting the users in controlling an application by gaze is to visualize the gaze as a cursor. Some related work indicates the problem of distraction from this kind of visualization [2], others see the chance of such a natural "pointer" [5].

Some years ago, a new class of applications appeared, the so-called games with a purpose (GWAPs) [3, 4]. The goal of GWAPs is to gain information from humans in an entertaining way. One example is the game Peekaboom [4] where two users play together for labeling image regions. Another is the ESP-Game [3] with two randomly assigned players, each tagging one image and trying to provide the same tags as the team mate. Tobii recently introduced the game EyeAsteroids and claims to be the first purely eye-controlled arcade game. It is entertaining, but does not have the goal to benefit from the users' activities. Eye tracking fascinates users as an unusual kind of input device. One can benefit from this curiosity by offering entertaining applications that also gain some information from the users. Despite the variety of eye tracking applications and games, *EyeGrab* combines – to the best of our knowledge – for the first time both, the aspects of leveraging from user activities like in GWAPs and controlling the application by the use of eye movements.

3. DESCRIPTION OF THE GAME

EyeGrab is a single-player game that takes place in a galactic dimension. The task is to clean up the aliens' universe by categorizing and rating images. Before the game starts, the user is asked to enter his or her nickname using the keyboard (Figure 1). The rest of the game is then played exclusively by the use of the eye tracker. Every gaze-based selection takes place after a dwell time of 450 milliseconds to avoid random choices. For example, the selection of the gender is done by focusing a male or female character as shown in Figure 1. Subsequently, the player is shown a small introduction to the game's rules (no screenshot).

The game has three rounds, with three categories ("car", "house", and "mouse"). First, the category is shown (see Figure 2), next the round starts. 30 images fall down the screen as depicted in Figure 3. The player categorizes the falling images into one of three categories. He or she can select an image by fixating it for more than 450 milliseconds. When an image is selected, it is highlighted by a thin, red frame. Next the image is classified into "like it", "don't like it", or "not relevant", where the first two imply that the image is described by the named tag and the third specifies that the image does not belong to that tag. To classify an image, the user looks at the area of the intended classification on the screen as shown in Figure 3 (same dwell time as above). The player receives points for each correctly categorized image, negative points for each false one and no points for images that fell off the screen without classifying them. To further challenge the user, the speed is increased with higher levels. A high score list is presented to the user at the end of the game.

Figure 1. Start screen with gaze-selected "male".

Figure 2. Presentation of the category.

For each category, the images were chosen from the 100 most relevant Flickr-pictures. 20 of them were randomly selected and combined with 10 pictures of a different category. An inter-rater agreement with 3 neutral persons was used to confirm the categorization and to create the ground truth.

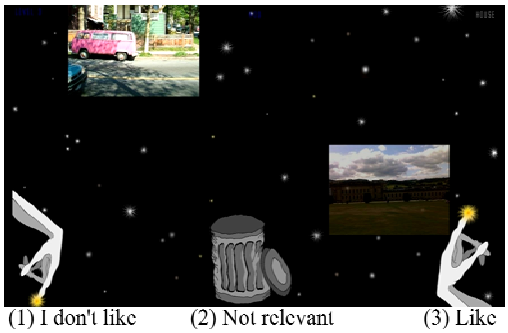


Figure 3. Gaze-based image classification.

Two versions of the game have been implemented, one offering visual aids (see Figure 4a) to the user and the other one without such help (see Figure 4b). The visual aids include a highlighting of the "action areas", i.e., areas which perform an action when being fixated, and the visualization of the gaze point on the screen (gaze-cursor). Examples are the classification buttons as shown in Figure 3 (details in Figures 4a and 4b).

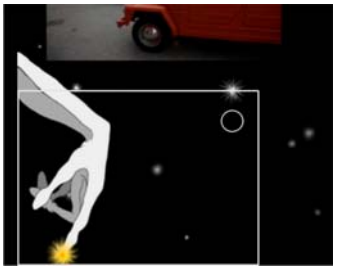


Figure 4a. Visual aids (rectangle: action area, here: "not like", circle: gaze cursor).

Figure 4b. No visual aids.

4. EVALUATION DESIGN

In order to evaluate *EyeGrab*, 24 subjects (7 female) played the game. The subjects' age was between 15 and 32 years (mean: 24, SD: 3.9). 19 subjects were students, 2 research assistant, one pupil and 2 had other professions. Most of the players had experience in gaming (mean: 3.5, SD: 1.31). Only a few were familiar with eye tracking, this is indicated by 19 subjects rating the question concerning their eye tracking experience with one (mean: 1.63, SD: 1.38). The subjects were randomly divided into two groups A and B. Group A had no visual aids during the game, whereas group B did. 8 users were wearer of glasses or contact lenses (4 in each group). There were no problems using the eye tracker for those subjects.

To avoid distractions, the game was played in our eye-tracking lab providing a chair, a desk with an eye tracker, and a standard monitor. The first step was a calibration of the eye tracker. After this was done, the game was started without further instructions and was played with 30 images in each of the three rounds. The data from the first round is not used in the later analysis, because it has only served for getting the subjects acquainted with the usage of the eye tracker as input device. At the end of the experiment, every user filled out a questionaire, including personal information and questions about the performance of the game. The answers were given on a 5-point Likert scale.

5. EVALUATION RESULTS

5.1 Satisfaction

The questionnaires show that the subjects enjoyed playing the game. On average, the statement "It was fun playing the game." is rated 3.46 (SD: 0.93) considering all 24 subjects. 20 of the 24 users agreed to this statement. One of the following questions was if the participants felt like the interaction with the eye tracker increases the fun of the game. 14 subjects agreed or strongly agreed to this statement (mean: 3.5, SD: 1.25). Also, most of the subjects did not feel disturbed by the eye tracker (mean: 2.25, SD: 1.5).

5.2 Effectiveness and Efficiency

One round of the game comprises 30 images and it takes about two minutes including the introduction and the input form. Each level has a different pace at which the images fall down the screen. Thus, the classification per image takes between 2.6 and 4 seconds.

In total, 1440 pictures were shown to the subjects within the game (30 pictures per category "house" and "mouse" times 24 subjects). Only in 42 cases, the image passed without classification, resulting in a total of 1398 classified images. 1162 images were correctly classified (83%). Thus, only 236 images were incorrectly classified. Overall we had 897 true-positive classifications, 128 false-negative and 108 false-positive classification, which leads to a precision of 89% and a recall of 88% over all users. For the group with the better results (the group without visual aid, see next section) we obtain a precision of 92%.

5.3 Visual aid

The subjective perception of the users in group B (the group that was provided with visual aids) was that the visual aids supported them in the classification tasks. The question regarding the visual highlighting of the active areas was rated as very helpful with an average of 4.67 (SD: 0.49). The subjects also answered that displaying the gaze point was very helpful and scored this question on average with 4.5 (SD: 0.67). However, to our surprise, the following statistical analysis of the data shows that group B with the visual aids misclassified significantly more images than group A did.

Group A correctly classified 296 images for category "house" whereas group B correctly assigned 264 images for this category. Regarding the category "mouse", in total 317 correct assignments were made by group A whereas group B correctly assigned 287 images. Regarding the misclassified images, group A misclassified 59 images for category "house", whereas group B wrongly assigned the image category in 81 cases. For the category "mouse", the number of incorrect assignments is 37 for group A and 57 for group B. We compared the values for correct and incorrect assignments for group A and B in a 2x2 Chi-square test for both categories. The differences are significant regarding a significance level of $\alpha = 0.05$ with $\chi 2$ (1, N = 700) = 5.14, p = 0.023 for the category "house" and $\chi 2$ (1, N = 698) = 5.6, p = 0.018 for "mouse". In group B 31, images passed without classification, in group A only 11 images were not classified.

These results indicate that the visual support is not improving the classification. Despite the good impression of the visual support that group B expressed, the following question might be an indicator that this group felt less comfortable with the eye tracker-based interaction than group A did: we asked the subjects to state if they preferred a mouse-based interaction instead of the eye tracker-based one. On average, subjects of group A scored this question with 2.17 (SD: 1.47) and group B with 3.25 (SD: 1.48). Using a two-tailed Mann-Whitney U-Test, a weakly significant difference was determined stating a preference of group B over group A to use the mouse to play *EyeGrab* (U = 43, z = -1.719, $n_1 = n_2 = 12$, p = .085).

6. FUTURE WORK

For the current version of our *EyeGrab* game, we have used pre-classified images in order to verify the classification performance of the subjects. We plan to use images without annotations in future extensions of the game.

Also the detailed analysis of the gaze information will be performed in a next step. In a small sample of 5 images classified by one user, we received 231 gaze points on the images. An example of a gaze path visualization is shown in Figure 5. We

expect a sufficient number of fixations and correct classification to allow a detailed analysis.

We received 897 ratings for the shown images. 556 of them were positive. The quality of these ratings has to be investigated in a future experiment, e.g., by repeating the ratings in another context with the same users or by using a ground truth set with images, often liked by a big number of other users. However it has to be clear, that a subjective rating can never be "correct" or not. These investigations can only provide an indication of the worth of the rating. Overall, this detailed analysis will allow us to identify the regions that correspond to the category given in the EyeGrab game. Such region-based annotations will allow for a better retrieval of the images in the future.

Figure 5. Visualization of fixations on a classified image.

7. SUMMARY

We have introduced the gaze-based game with a purpose *EyeGrab* to classify images using an eye tracker. We have shown that the game has the potential to entertain the players and that the classification results are good enough to advance beyond the gaze analysis. This analysis is the first step in the direction of extending image context information with information gained in an eye tracking game. The next step will be the analysis and evaluation of the gained information and to use it for improving image search tasks.

8. REFERENCES

[1] Hornof, A.J. and Cavender, A. 2005. EyeDraw: enabling children with severe motor impairments to draw with their eyes. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 170.

[2] Jacob, R.J.K. 1993. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in human-computer interaction*, 151–190.

[3] Von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326.

[4] Von Ahn, L., Liu, R. and Blum, M. 2006. Peekaboom : A Game for Locating Objects in Images. *SIGCHI conference on Human Factors in computing systems*, 55–64.

[5] Zhang, X. and Ren, X., and Zha, H. 2008. Improving eye cursor's stability for eye pointing tasks. *SIGCHI conference on Human factors in computing systems*, 525–534.

[6] Kozma, L., Klami, A. and Kaski, S. 2009. GaZIR: gaze-based zooming interface for image retrieval. In *Proceedings of the 2009 international conference on Multimodal interfaces*.

[7] J. San Agustin, H. Skovsgaard, J.P. Hansen, and D.W. Hansen. Low-cost gaze interaction: ready to deliver the promises. *In Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 4453–4458. ACM, 2009.

[8] Walber, T. and Scherp, A. and Staab, S. 2012. Identifying Objects in Images from Analyzing the Users' Gaze Movements for Provided Tags. *Advances in Multimedia Modeling*, 138-148.

[9] Carson, C., Thomas, M., Belongie, S., Hellerstein, J. and Malik, J. 1999. *Blobworld: A system for region-based image indexing and retrieval.* Visual Information and Information Systems.

Using Wordclouds to Navigate and Summarize Twitter Search Results

Rianne Kaptein
Oxyme
Amsterdam, The Netherlands
rianne@oxyme.com

Abstract

This paper describes an application in which wordclouds are used to navigate and summarize Twitter search results. A search on Twitter can return thousands of relevant tweets. By just looking at the first few result pages you will not get an overview of what is discussed in all search results. Our application summarizes sets of tweets into wordclouds, which can be used to get a first idea of the contents of the tweets. Also the application provides the option to zoom in on a certain part of the search results to inspect them in more detail. The application has not been formally evaluated, but we do provide some insights and points for discussion.

1 Introduction

One of the most common problems in Information Retrieval is information overload: there is simply too much relevant information available for the users to process. Therefore applications are needed to help users deal with large amounts of data. In this paper we describe an application which was developed for this purpose. The use of wordclouds in the application serves two purposes:

1. To summarize
2. To aid navigation

This application was developed with the following two user scenarios in mind:

1. General Twitter search
   Nowadays many people express their opinions about products, services and companies on Twitter. When you want to get a broad overview of what people are tweeting in general about a company or event, it does not suffice to read the first few pages of search results. You want to get a feeling for the most frequently discussed topics overall, and dive into particular subtopics of special interest, such as product recommendations.
2. Searching fragments of categorized data
   Besides Twitter there are many more places on the Web where people express their opinions. These opinions can be collected and annotated with labels such as sentiment, source, market etcetera. When you have a large amount of annotated data available, it is interesting to see for example what are the different topics discussed in positive and in negative messages.

In this paper we will focus on the first user scenario: General Twitter search, since Twitter data is abundant and publicly available.

Humans have a great capacity to notice terms which are out of the ordinary. When looking at a wordcloud there will always be some unexpected terms which catch your attention and are good pointers for further investigation. In tweets about public transport you can expect for example tweets about delays, but you might not expect certain tweets about recent events such as a new colour of the trains. What we try to do in the wordclouds is to emphasize the words that are noteworthy from a statistical point of view, and leave it up to the user to decide which messages to explore further.

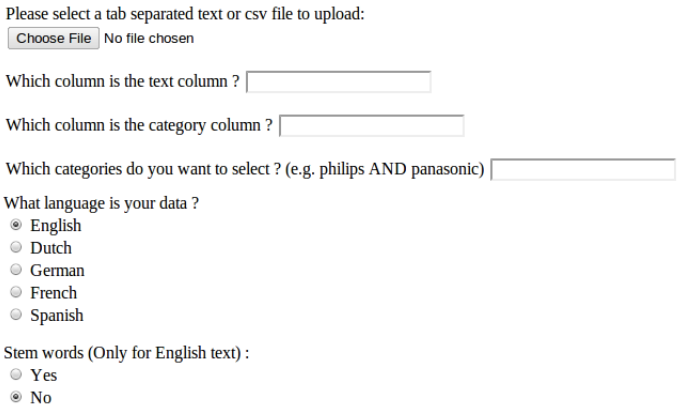
Although the usefulness of tagclouds for navigation is still a topic of debate [2], exploratory applications which make use of wordclouds for summarization and navigation of search results have been moderately successful on specific domains such as web documents [1] and PubMed publications in biomedical literature [5].

The search results that we are investigating in this paper have three characteristics:

- A search result is a short textual message. By design a Twitter message cannot contain more than 140 characters.
- The number of search results is large. If this would not be the case, since the results are short texts, you could simply read through all of them.
- There are many, equally relevant search results. In web search there are usually not more than a handful highly relevant search results. Many of the search results contain copied or redundant information, or only mention the search words occasionally. Although Twitter search results also contain redundant information, i.e. repeated tweets and retweets, the set of relevant tweets can still consist of thousands of equally highly relevant tweets.

In the next sections of this paper we will present our approach (Section 2), a case study (Section 3), and finally our conclusions (Section 4).

Figure 1. First part of the inputscreen



2 Approach

The application consists of two screens. The first screen handles the input, the second screen displays the results based on your input.

On the first screen the system offers a number of selections that can be made to make sure you generate the wordclouds that are best representing your data and your analysis purpose. The input is collected using textfields, radiobuttons and checkboxes. The first part of the inputscreen is shown in Figure 1.

The following selections can be made:

- File selection, a tab separated text file is required as input.
- Text selection, which column in the dataset to use as textual input for the wordcloud generation.
- Category selection, based on a value in any column of your dataset your data can be categorized. It is also possible to create categories based on the presence of words in the contents of your data, e.g. to create a category for all tweets containing the term 'happy'.
- Language, used for the removal of standard stopwords.
- Optionally, additional stopwords can be specified. These words will not occur in any of the wordclouds.
- Stemming, currently available only for English. The Krovetz stemmer is used, because this stemmer always stems words into other valid English words.
- Exclude numbers, when your data includes many numbers such as product prices it can be desirable to exclude these numbers from the wordcloud.
- Exclude retweets / repeated posts, when your data contains a tweet that is retweeted very frequently, this one tweet will dominate the wordcloud which can be undesirable.
- Include only usernames, for Twitter data only, keep only the usernames, i.e. all the words starting with @.
- Include only hashtags, for Twitter data only, i.e. all the words starting with #.

The second screen shows the output, which consists of wordclouds for the categories you have specified, as well as a wordcloud for all the search results.

Wordclouds for categories are generated using a parsimonious language model. This model compares the frequency of words in a set of documents to the average term probability in a background collection containing similar documents to extract the most noteworthy terms. In this case the background collection are all the retrieved search results. Terms that are only mentioned occasionally in the set of documents and terms which have a similar or higher probability of occurrence in the background collection will not be included in the parsimonious language model [4].

The parsimonious language model [3] is an extension to the standard language model based on maximum likelihood estimation, and is created using an Expectation-Maximization algorithm. Maximum likelihood estimation is used to make an initial estimate of the probabilities of words occurring in the set of documents.

$$P_{mle}(t_i|S) = \frac{tf(t_i,S)}{\sum_t tf(t,S)} \tag{1}$$

where S is the set of documents, and $tf(t,S)$ is the text frequency, i.e. the number of occurrences of term t in set of documents S. Subsequently, parsimonious probabilities are estimated using *Expectation-Maximisation*:

$$\begin{aligned} \text{E-step:} \quad & e_t = tf(t,S) \cdot \frac{(1-\lambda)P(t|S)}{(1-\lambda)P(t|S)+\lambda P(t|C)} \\ \text{M-step:} \quad & P_{pars}(t|S) = \frac{e_t}{\sum_t e_t}, \text{ i.e. normalize} \end{aligned} \tag{2}$$

where C is the background collection model. In the initial E-step, maximum likelihood estimates are used for $P(t|S)$. We set the smoothing parameter λ to 0.9. In the M-step the words that receive a probability below a threshold of 0.001 are removed from the model. The iteration process stops after a fixed number of iterations.

In the next section we present a case in which the generated output of the application is presented.

3 Case

Using an example search we will demonstrate how we use wordclouds in our application to navigate and summarize the search results. We executed a search on Twitter using the Twitter search API[1] for the query '#london2012' over the last 5 days, saved all the 30,504 search results in a .csv file and load this file into our application. Looking at the wordcloud over all the results that is shown in Figure 2, we see the term 'torch' is frequently used, and we zoom in on this aspect of the '#london2012' search. By clicking on the word 'torch' a list of messages is shown that all contain the term 'torch', so these messages can be inspected in more detail. This list of messages is still quite long however, consisting of 1,046 tweets. We can zoom in further on these tweets by going back to the input screen and specifying 'torch' as a category. Now, a parsimonious wordcloud is created from the 1,046 tweets that contain the term 'torch'. The resulting wordcloud is shown in Figure 3. The figure is a screenshot of the screen that is displayed when the word 'Sheffield' is clicked, showing the tweets containing the word 'Sheffield'.

Words which occur frequently in all of the '#london2012' messages, such as '#london2012', '2012', and 'olympics', receive a lower score from the parsimonious model, and almost none of these

[1] https://dev.twitter.com/docs/api/1/get/search

Figure 2. Wordcloud of all #london2012 Twitter search results, showing the tweets containing the term 'torch'

Figure 3. Wordcloud of #london2012 Twitter search results containing the term torch, showing the tweets containing the term 'sheffield'

words occur in the 'torch' wordcloud. Also general words that occur frequently in all of the messages, such as 'get', and 'will' are filtered out. Instead the cloud contains words that occur more frequently in the subset of messages that contain the word 'torch', for example some of the cities that the torch passes through such as Sheffield, Leeds and Manchester. Every result in this cloud by definition contains the word 'torch', therefore it takes a prominent place in the wordcloud. You can choose to not show the word 'torch' in the wordcloud by specifying it as a stopword on the input screen.

Clicking on a term in the wordcloud has the same effect as query expansion, i.e. adding that term to your query and retrieve another set of results. When you use the Twitter API to search Twitter without query operators, only results will be returned that contain all of the search terms in the Tweet, username or hyperlink. This means adding a term to your query will not lead to more search results. Only if you remove the original query terms, other results will be returned.

Observations

We have not had the chance to evaluate our application through means of a user study. However, we do want to point out the following observations. Given the nature of our data, i.e. a collection of tweets, there might be some improvements possible that exploit this particular type of data. Tweets can contain special elements in the text, namely usernames, hashtags, links, and emoticons. We make the following observations:

- Usernames and hashtags are currently considered in the sense that we remove all punctuation except the characters '@' and '#' which are the indicators of usernames and hashtags respectively. There is an option to generate wordclouds containing only usernames, or only hashtags. In the default settings usernames and hashtags are included as is in the wordcloud. For future work we want to discuss and investigate two open issues:
    1. Can a word with a hashtag be considered as the same word without the hashtag? While a hashtag term does not always have to be a real word, e.g. #london2012, in many cases it is, e.g. #london. For the wordcloud

should the terms ‘london’ and ‘#london’ be merged? Sometimes usernames are used in a similar way as hashtags to address companies, e.g. in this tweet: ‘Ambush marketing at the Olympics! Well played, @Nike. bit.ly/N4zAUc #London2012’.

2. A related issue is the importance or term weights of usernames and hashtags. Is a hashtag a stronger signal, and should it therefore be featured more prominently in the wordcloud? Similarly for usernames, but usernames could also be considered a weaker signal, so should they be featured less prominently?

Both of these questions can also be considered when you want to optimize a retrieval algorithm.

- Besides the ‘@’, and ‘#’ all other punctuation is removed during text preprocessing. This means all emoticons like ‘:)’ are removed. Sometimes these emoticons are used as indicators of sentiment, i.e. tweets containing ‘:)’ are classified as positive messages, and tweets containing ‘:(’ as negative messages. In this sense the emoticons do indeed represent valuable information that could be included in the wordcloud. When an emoticon appears in the wordcloud, clicking on it can give you all the messages associated with for example a positive emoticon.

Feedback from users is required to determine the most useful improvements for the application.

4 Conclusions

In this paper we have shown how wordclouds can be used to summarize and navigate search results, and in particular Twitter search results. Wordclouds are a quick way to summarize and get a first overview of large amounts of data. Using human observation skills it is easy to zoom in on a group of messages in which you are interested, i.e. all messages that contain a specific term from the wordcloud. In future work we would like to evaluate the usefulness of wordclouds for navigation and summarization of search results in a user study.

5 References

[1] T. Gottron. Document Word Clouds: Visualising Web Documents as Tag Clouds to Aid Users in Relevance Decisions. In M. Agosti, J. L. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *ECDL*, volume 5714 of *Lecture Notes in Computer Science*, pages 94–105. Springer, 2009.

[2] D. Helic, C. Trattner, M. Strohmaier, and K. Andrews. Are tag clouds useful for navigation? A network-theoretic analysis. *IJSCCPS*, 1(1):33–55, 2011.

[3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious Language Models for Information Retrieval. In *Proceedings SIGIR'04*, pages 178–185. ACM Press, New York NY, 2004.

[4] R. Kaptein, D. Hiemstra, and J. Kamps. How Different are Language Models and Word Clouds? In *Advances in Information Retrieval: 32nd European Conference on IR Research (ECIR 2010)*, volume 5993 of *LNCS*, pages 556–568. Springer, 2010.

[5] B. Y.-L. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson. Tag clouds for summarizing web search results. *Proceedings of the 16th international conference on World Wide Web WWW 07*, 196:1203, 2007.

Do users benefit from controlled vocabularies in search interfaces?

Ying-Hsang Liu
School of Information Studies
Charles Sturt University
Wagga Wagga, NSW 2678, Australia
yingliu@csu.edu.au

Paul Thomas
CSIRO
GPO Box 664
Canberra, ACT 2601, Australia
paul.thomas@csiro.au

Jan-Felix Schmakeit
Research School of Computer Science
Australian National University
Canberra, ACT 2601, Australia
jan-felix.schmakeit@anu.edu.au

Tom Gedeon
Research School of Computer Science
Australian National University
Canberra, ACT 2601, Australia
tom@cs.anu.edu.au

ABSTRACT

Search providers in domains from medicine to news have long labelled documents with controlled vocabularies, to help users explore their collections. These vocabularies are expensive to build and use, however, and seem to be useful mostly for domain experts.

This paper describes an on-going gaze-tracking study which asks whether users notice controlled vocabularies when they are exposed in a search interface; whether they make use of them; and whether this improves search. We also hope to learn what effect several standard search interfaces have on the use of controlled vocabularies.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.5.2 [**User Interfaces**]: User-centered design—*performance measures*

General Terms

Experimentation, Human Factors

Keywords

Search results presentation, individual differences, gaze behaviour, MeSH terms

1. INTRODUCTION

It has been recognised that people engage with different kinds of searching behaviours, but current information retrieval (IR) systems are primarily designed for specified search [1]. The simple search box is still the dominant interaction mode in modern search engines. However, a user-centred approach to interface design that takes into account individual differences, search goals and tasks, has the potential to support users interacting with IR systems more efficiently and effectively.

To this end researchers have advocated "natural" search user interfaces, arguing they are easier to use and require less user training [e.g. 9, 20]. It is however challenging to design natural interfaces because of the complexity of information problems and associated searching behaviours. For instance, user studies have demonstrated that user queries are typically very short representations of complex information needs [3, 11], and users have difficulty formulating queries to represent information problems. User interaction with IR systems is inherently interactive and exploratory [e.g. 2, 17], so usable interfaces for query formulation are important in support of natural search interactions. (See Wilson [24] for a recent comprehensive review of search interfaces, and Wacholder [22] for a review of interactive query formulation.)

One way to support query formulation is with a controlled indexing language, where each document is assigned terms from an predefined list or hierarchy of indexing terms. Examples include Medical Subject Headings (MeSH) terms and Library of Congress Subject Headings (LCSH). The usefulness of MeSH terms in biomedical searching is especially important because of the extreme popularity of the PubMed database[1], the publicly accessible version of MEDLINE on the web.

Controlled vocabularies are expensive to build, use, and maintain, and they may contribute to clutter in a search interface. There is some evidence that domain experts benefit from controlled vocabularies, but results have been mixed for ordinary users (e.g., [10, 15, 19]). Given these costs, and the unclear benefits for most searchers, we are interested in whether and how users make use of controlled vocabularies when they are available.

[1] `http://www.ncbi.nlm.nih.gov/pubmed`

This paper describes an on-going eye-tracking study of user gaze and search behaviours searching clinical search topics, with particular reference to the user's attention to and use of the document surrogates (i.e., MeSH terms, title and abstract). The specific research questions are:

1. What components of document surrogates do searchers look at when reformulating their queries? Do searchers even notice MeSH terms in standard search interfaces?

2. If they do notice them, how do searchers use the displayed MeSH terms in their search processes?

3. If they are used at all, do MeSH terms lead to better search performance and efficiency?

2. RELATED WORK

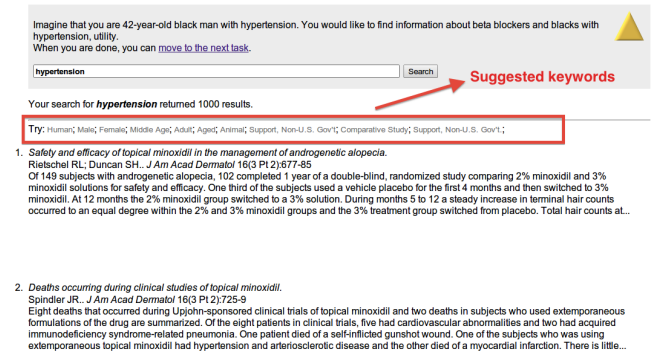
Past work has considered system designs to support query reformulation. From a system perspective, researchers have proposed visualizing document inter-relationships [21], explicit term distribution information [8] and search interfaces in support of search results navigation [18] to help users refine their queries. From a user perspective, research has revealed that searchers prefer to use such search interfaces for reformulating their queries and to have some degree of control over the search process [e.g. 12, 8, 13, 23]. In a recent study of search interfaces in support of interactive query expansion [7], it was found that displaying expanded terms and corresponding changes in summaries of search results was useful for the decision-making process in query reformulation; particularly for difficult search topics. However, it is still unclear whether users pay attention to these system features, and whether the use of these features contributes to better search performance and efficiency.

Recent HCI and IR research has focused on users' cognitive aspects in search interactions by measuring the gaze patterns, an indicator of searcher attention (see e.g. Dumais et al. [5] or Logio et al. [16]). The use of eye-tracking equipment for capturing searchers' fixation patterns provides a rich set of data to understand whether searchers read document surrogates (e.g. summary and metadata) and more importantly, how searchers attend to different components of search results or search interfaces [4, 14]. We are adopting similar techniques in our study.

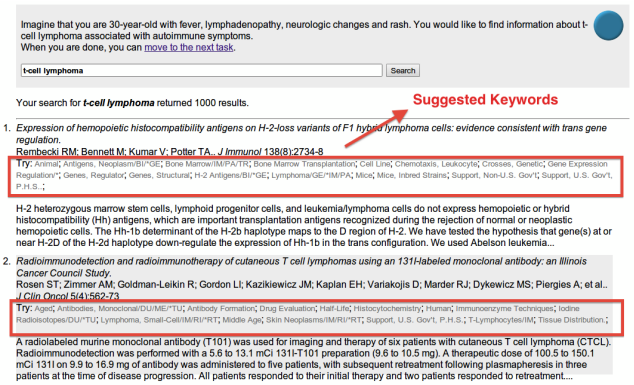
3. METHODS

We are conducting a user experiment to assess the effect of displayed MeSH terms on search behaviors and performance. The search task is to perform searches on clinical information for other patients, and find the best query to obtain as many relevant documents as possible. Our recruits are undergraduate and postgraduate students with search engine experience but without advanced academic background in the biomedical domain. Each user searches 8 topics in total, with a 7-minute limit for each topic, and the experiment takes about 90 minutes in total.

Participants are given brief instructions about the search task and system features, followed by a practice topic and then the searches proper. User interaction data is recorded: we are noting all queries, mouse clicks, retrieved documents, time spent, and eye movements. Electroencephalogram (EEG) readings are also captured.



(a) Screenshot of Interface "B", suggestions per-query and displayed at top.



(b) Screenshot of Interface "D", suggestions per-document and displayed with the document.

Figure 1: Two of the four search interfaces in our study.

3.1 Search interfaces

Participants search on four different search interfaces using a single search system. The four search interfaces are distinguished by whether the MeSH terms are presented and how the displayed MeSH terms are generated:

Interface "A" mimics web search and other search systems with no controlled vocabulary. This interface has a brief task description at top; a conventional search box and button; and each result is represented with its title, authors, publication details, and abstract where available.

Full text is not available, so the results are not clickable. Users must judge their success on the titles and abstracts alone.

Interface "B" (Figure 1(a)) adds MeSH terms to the interface. After the user's query is run, MeSH terms from all results are collated; the most frequent ten are displayed at the top of the screen. This mimics the per-query suggestions produced by systems like ProQuest[2].

MeSH terms are introduced with "Try:" and are clickable: if a user clicks a term, their query is refined to

[2]For example, see `http://www.proquest.co.uk/en-UK/products/brands/pl_pq.shtml`

Imagine that you are 63-year-old male with acute renal failure probably 2nd to aminoglycosides/contrast dye. You would like to find information about acute tubular necrosis due to aminoglycosides, contrast dye, outcome and treatment.

Figure 2: An example OHSUMED search topic, reworded for our participants.

include the MeSH term and then re-run. We hope that the label, and the fact they work as links, will encourage users to interact with them.

Interface "C" uses the same MeSH terms as "B" but displays them alongside each document, where they may be more (or less) visible. It is a hybrid of interfaces "B" and "D".

Interface "D" mimics EBSCOhost[3] and similar systems that provide indexing terms alongside each document. As well as the standard elements from interface "A", interface "D" displays the MeSH terms associated with each document, as part of that document's surrogate (Figure 1(b)).

Again, terms are introduced with "Try:" and are clickable.

Each interface is labelled with a simple figure—a square, circle, diamond, or triangle—which we refer to in our exit questionnaire.

3.2 Design

This experiment is a 4×4 factorial design with four search interfaces and four topic pairs. We are using a 4×4 Graeco-Latin square design [6] to arrange the experimental conditions. We expect to enroll 32 participants from the campus of a large university, which will give good statistical power (when $N = 32$, ANOVA $\beta < 0.01$ for "medium" effect of $\Delta = 0.75$).

Entry and exit questionnaires are collecting demographic information and information on participants' cognitive styles and their perception of the search process. We also ask participants' opinions of the tasks and the interfaces.

3.3 Topics

Search topics used here are a subset of the clinical topics from OHSUMED [10], originally created for batch-mode IR system evaluation. We have re-written the topics slightly so they read as instructions to our participants (see Figure 2 for an example).

We selected topics to cover a range of difficulties: we sorted the topics according to the number of judged relevant documents and selected two topics, at random, from each quartile. These eight topics were then randomly paired off to produce four pairs of topics. A final topic, the same for all participants, is used for training.

3.4 Software and hardware

The search system is built on Solr[4], with the search results ranked by default relevance score. The MeSH terms are not specifically weighted.

Gaze tracking uses FaceLab[5] software and hardware. We use Eyeworks software[6] for recording and basic analysis. EEG data is recorded with an Emotiv headset[7].

3.5 Analysis

With the design above, we expect to answer the three questions from Section 1.

Where do people look?. Recordings will be analysed to see how often there are fixations in different parts of document surrogates, and therefore how often people have looked at each part. In particular, for interfaces B, C and D we will consider how often participants look at the controlled vocabularies ("Try:..."). Any effect on gaze patterns due to interface would tell us which interfaces make the extra information easiest to discover.

Our exit questionnaire also asks whether users noticed the controlled vocabularies: we would not be surprised if there were differences between the self-reported data and the gaze data, for example if participants were trying to please us.

Do they use the controlled vocabulary?. Our software records all clicks on terms from the controlled vocabulary, so it will be easy to note how often it is used and whether there is any correlation with interface, task, sequence, or user. Again, an effect due to interface would suggest which style of interface makes features like the controlled terms most attractive.

Participants who merely read and re-type the controlled vocabulary may be picked up in query logs.

Again, we intend comparing these recordings with self-reports.

If so, does it help?. Assuming some participants do make use of the MeSH terms, we anticipate four ways to address this question. First, as before, we will consider self-reports of task difficulty to see whether these correlate with the use of controlled vocabulary features. Second, since participants' final queries on each topic should be the ones the like best, we can check how many of these use MeSH terms. Third, the judgements associated with OSHUMED topics will allow us to measure the actual effectiveness of queries with and without controlled terms. Finally, if participants do not use all their allocated time for each task, variations in completion time may be interesting.

4. FIRST RESULTS AND NEXT STEPS

We have conducted a small-scale pilot to test our design and instruments.

Our participants did glance at MeSH terms: 8% of fixations were on MeSH terms in interfaces B to D, which compares to 6% on document titles and 12% on abstracts. However, they were very seldom used – only one query, of 44 queries issued on these interfaces, used any MeSH terms at all. There are also some indications of a per-interface effect, with the MeSH terms at the top of interface D receiving little attention We will shortly be recruiting for the full-scale experiment. We hope this will offer some insight into the relationship

[3] http://www.ebscohost.com/
[4] http://lucene.apache.org/solr/
[5] http://www.seeingmachines.com/product/facelab/
[6] http://www.eyetracking.com/Software/EyeWorks
[7] http://www.emotiv.com/

between interface, reading patterns, search behaviour, and search effectiveness.

5. ACKNOWLEDGMENTS

Ying-Hsang Liu has been supported by the School of Information Studies Research Fellowship from Charles Sturt University and working as Visiting Fellow at Research School of Computer Science, The Australian National University.

6. REFERENCES

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

[2] N. J. Belkin, P. G. Marchetti, and C. Cool. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.

[3] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: I. Background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[4] E. Cutrell and Z. Guan. What are you looking for?: An eye-tracking study of information usage in web search. *Proceedings of the SIGCHI Conference*, pages 407–416, 2007.

[5] S. T. Dumais, G. Buscher, and E. Cutrell. Individual differences in gaze patterns for web search. *Proceeding of the Symposium on Information Interaction in Context (IIiX '10)*, 3:185–194, 2010.

[6] R. A. Fisher. *The design of experiments*. Hafner Press, 9th edition, 1971.

[7] N. Gooda Sahib, A. Tombros, and I. Ruthven. Enabling interactive query expansion through eliciting the potential effect of expansion terms. *Lecture Notes in Computer Science*, 5993:532–543, 2010.

[8] M. A. Hearst. Tilebars: Visualization of term distribution information in full text information access. *Proceedings of the SIGCHI Conference*, pages 59–66, 1995.

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

[10] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the ACM SIGIR Conference*, 17:192–201, 1994.

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

[12] D. Kelly and X. Fu. Elicitation of term relevance feedback: An investigation of term source and context. In *Proceedings of the ACM SIGIR Conference*, pages 453–460, New York, 2006. ACM.

[13] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. *Proceedings of the SIGCHI Conference*, pages 205–212, 1996.

[14] B. Kules and R. Capra. Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. *Journal of the American Society for Information Science and Technology*, 63(1):114–138, 2012.

[15] Y.-H. Liu and N. Wacholder. Do human-developed index terms help users? an experimental study of MeSH terms in biomedical searching. *Proceedings of the American Society for Information Science and Technology Annual Meeting*, 45(1):1–16, 2008.

[16] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

[17] G. Marchionini and R. White. Find what you need, understand what you find. *International Journal of Human-Computer Interaction*, 23(3):205–237, 2007.

[18] X. Mu, H. Ryu, and K. Lu. Supporting effective health and biomedical information retrieval and navigation: A novel facet view interface evaluation. *Journal of Biomedical Informatics*, 44(4):576–586, 2011.

[19] M. L. Nielsen. Task-based evaluation of associative thesaurus in real-life environment. *Proceedings of the American Society for Information Science and Technology Annual Meeting*, 41:437–447, 2004.

[20] K. A. Olsen and A. Malizia. Interfaces for the ordinary user: can we hide too much? *Commun. ACM*, 55(1):38–40, 2012.

[21] R. C. Swan and J. Allan. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. *Proceedings of the ACM SIGIR Conference*, 21:173–181, 1998.

[22] N. Wacholder. Interactive query formulation. *Annual Review of Information Science and Technology*, 45:157–196, 2011.

[23] R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Information Processing and Management*, 43(3):685–704, 2007.

[24] M. L. Wilson. Search user interface design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(3):1–143, 2011.

User-Centred Design to Support Exploration and Path Creation in Cultural Heritage Collections

[1]Paula Goodale, [1]Paul Clough, [1]Nigel Ford, [1]Mark Hall, [1]Mark Stevenson, [1]Samuel Fernando, [1]Nikolaos Aletras, [2]Kate Fernie, [3]Phil Archer, [4]Andrea de Polo

[1]University of Sheffield, Sheffield, United Kingdom
[1]p.goodale | p.d.clough | n.ford | m.mhall | m.stevenson | s.fernando | naletras1@sheffield.ac.uk
[2]MDR Partners, United Kingdom; [3]iSieve Technologies, Greece; [4]Alinari 24 ORE Florence, Italy
[2]kate.fernie@mdrpartners.com, [3]phil@philarcher.org, [4]andrea@alinari.it

ABSTRACT
In this paper, we present the results of the user requirements and interface design phase for a prototype system, designed to enhance interaction with cultural heritage collections online through means of a pathway metaphor. We present a single user interaction model that supports various work and information seeking tasks undertaken by both expert and non-expert users within the context of collection exploration and path creation. The user interaction model is shown to enable seamless movement between interaction modes, with the potential over time to encourage deeper engagement and learning.

Categories and Subject Descriptors
H.5.m [Miscellaneous]: Interaction framework

General Terms
Design, Human Factors, Theory.

Keywords
Cultural Heritage, Paths, Information Access, User Requirements, Interaction Model, Exploration.

1. INTRODUCTION
Large-scale projects for the digitisation of cultural heritage (CH) have become commonplace in recent years, and yet complex issues arise with regard to information access. Specialist metadata and the often variable quantity and quality of object descriptions make it difficult for users to navigate vast, structured and often very scholarly collections. It is therefore difficult to locate resources of interest, especially for those without advanced levels of subject and domain knowledge [6]. User experience online is thus far removed from that of visiting a museum or gallery in person, where guidance through a much smaller selection of carefully curated objects is the norm, for example, via the medium of visitor-friendly object labels, guide books, audio tours and activity trails. Exhibit information is designed for general rather than academic audiences, with additional materials tailored for family groups and learners, amongst others.

Guided tours and activity trails are commonplace offerings to aid visitor orientation at physical cultural heritage sites, and offer a range of opportunities for immersive and more highly engaged visitor experiences [4], often utilising technological solutions, and even extending to the latest mobile devices [15]. They are though much less in evidence online, despite the fact that the idea of documents or other items linked together in the form of hypertext trails is considerably older than the web itself [3].

Online paths and trails are seen as a means of aiding navigation, exploration and learning [10] in general and educational online environments, and there are many examples of research [11] [12] [13] and commercial activity in offering tools to develop paths from web pages (e.g. www.trailmeme.com) and social media content (e.g. www.storify.com). However, very few examples are domain-specific and/or pertain to digital library collections, and in consequence, it is rare for all of the associated exploration, authoring and use activities to be integrated within the same space. Through our current research we therefore aim to exploit opportunities to utilize paths to support diverse groups of users in the complete cycle of information seeking, exploration, path creation and interaction within CH digital collections, opening up their use to more widespread educational and leisure audiences.

2. RELATED WORK
Research on information user behavior in CH digital collections is scarce, especially when considering the needs on non-expert users, i.e. those without detailed subject and domain knowledge. Expert users regularly engage in both simple fact-finding and more complex information gathering tasks, amongst others, with the latter having multiple variations and components such as topic searches, exploration, collecting/combining [1], all of which are relevant to our current study. Similarly, non-expert users [14] also engage in known-item searching and exploration. Visual representations of artefacts are highly important in this context, and the process of meaning-making through contextual information and the derivation of personal inferences and connections is also strongly evidenced [14].

For known-item or fact-finding searches, some knowledge of the metadata and collection structure is imperative, but such knowledge is much less likely to be used effectively, if at all by non-expert users than expert users [7]. In addition, information retrieval tools in CH collections, and the web more generally, are much less likely to effectively support the needs of users more in more open-ended exploratory tasks.

Exploratory search extends the idea of basic lookup into the areas of learning and investigation, which in turn incorporate extended information processing, evaluation and annotation [9]. Aligned with these variations of exploratory search are the concepts of

serendipity [5], where the user encounters information that they were not actively looking for, and berry-picking [2], which is an extended, iterative and adaptive search process that also incorporates the idea of collecting information objects as the search progresses over time. Solutions for these more complex user needs are yet to fully exploited, with greatest potential in adaptive systems that take account of patterns of user behaviour [8] [9], and the use of paths or trails as a means of capturing items of interest [11].

3. METHODS

In the absence of an existing system, extensive requirements gathering [6] was conducted with *potential* users, as the first stage in a user-centered design process. The goals of this research were to:

- Develop a detailed understanding of the characteristics and needs of potential users across four primary domains: heritage, education, professional/commercial, and general/leisure.
- Explore the meanings and potential applications of the path metaphor in the context of digital CH.
- Gain an understanding of the path-creation process and the types of paths that might be created.
- Determine the current availability and functionality of path-creation tools in CH collections.

In order to achieve these goals, mixed methods were employed, gathering a variety of complementary qualitative and quantitative data. First, an online user survey was used to collect data from 79 expert and non-expert users, comprising questions about their personal and cultural participation characteristics, and information behavior and use in the CH context. This was complemented by in-depth semi-structured interviews conducted with 22 expert users, which focused on exploring the meaning of the path metaphor in CH environments, and understanding the process of development and use of paths in this context.

Secondary data was used to scrutinise the features of published paths from various sources, to ascertain their core elements. Similarly, a comparative analysis of general and cultural-heritage specific systems offering path-creation functionality was conducted, to identify common features and standard approaches to the proposed core functionality. These findings were validated via user participation in path-creation tasks, utilsing low- and medium-fidelity techniques.

Analysis of these various complementary data enabled the development of detailed domain and role-specific information user profiles; a user interaction model supporting four key modes of interaction; and, use cases illustrating some of the primary user interaction scenarios. From these we extrapolated detailed user requirements, and in turn, interface designs and functionality for the first PATHS prototype. The resulting system is intended to support all elements of the interaction model, allowing users to move seamlessly between modes of use.

4. RESULTS

Given the breadth and depth of data, this paper focuses on the findings relating to paths and their uses in CH, and in turn, how related user tasks are incorporated within a single user interaction model, to be implemented in the prototype system.

4.1 Existing Path Forms

Analysis of existing paths and trails found that online and offline paths both have similar characteristics. *Nodes* are the essential building blocks of all paths, representing collection objects. Each node has associated metadata and primary content (e.g. descriptions, images) relating to the object. *Connections* between nodes enable navigation through the path and often represent meaningful relationships between objects. In the online environment, additional features of paths included *navigation tools* (e.g. path overviews and back/forward arrows), *annotations* added by the path creator to give context and guidance for use, and occasionally *links* to other related content, both within the same collection, and/or in external web sites. These findings largely support the initial vision for PATHS and can all be seen in the first prototype design.

In addition, it was found that most existing online paths are *static* and pre-published by an author, *linear* in form, rather than a more complex map or network structure, and *standalone*, without inter-connections with other paths. These findings fall somewhat short of the PATHS vision, limiting the possibilities for exploration and discovery, although for pragmatic reasons, they form the core functionality of the first prototype, with more advanced variations of paths coming later.

4.2 The Path Metaphor in Cultural Heritage

Interviews with potential expert users in the heritage, education and professional domains found a strong affinity with the path metaphor, revealing a range of different interpretations of what it means in the CH context, and similarly about what form paths might take, and how they could be employed in an online environment to engage with key audiences. Eight interpretations of the path metaphor emerged:

1. Path as search history
2. Path as information seeking journey
3. Path as linked metadata
4. Path as a starting point or way in
5. Path as a route through
6. Path as augmented reality
7. Path as information literacy journey / learning process
8. Path as transaction process

The first three of these are closest to the idea of hypertext trails [3], with trails defined by user interaction in 1 and 2, and trails defined automatically, by the system in 3. Variations 4-6 are more creative interpretations, all suggesting opportunities for guiding the user into and through collections, encouraging exploration and/or offering an immersive experience. In addition to expert-defined routes, 5 also incorporates the idea of users being able to see and follow "well-trodden paths" defined by the cumulative interactions of other users, thus extending the opportunities for utilizing search histories. Lastly, 7 and 8 are both process oriented, although 7 is experiential, user-defined, learning-oriented, typified by trial and error and unique to the individual, whilst 8 is a rigid process designed to escort all users consistently through a standard process of pre-defined steps.

4.3 Desired Characteristics of Paths

Expected characteristics of paths were explored, and views contrasted markedly with the existing path formats enabled by

path-creation tools currently available. Linearity is rarely seen as the best option for maximizing the potential of paths as exploration devices. Allied to this is the belief that starting and end points for paths should be mutable rather than fixed, allowing different users to explore a path in different ways according to their preferences and needs.

In the absence of linearity, some form of organization is still required to aid the accessibility and navigation of the path. The most popular option is for path content to be aligned to themes, with other alternatives including date, location, narrative and author, where the latter might present multi-layered paths offering the differing perspectives of several path-creators on the same topic. An over-arching conceptual framework for the path is also desirable, in order to tie together the themes and other ideas.

As a way-finding or navigational aid, paths are seen to support both guided and exploratory behavior, with the latter seen as the more desirable goal for user interaction. Features that are needed to enable way-finding include path overviews, navigational context in the form of next/last and nearby nodes, branching opportunities where paths converge and diverge, visualization, e.g. in the form of timelines or maps, and some degree of object level information at the node and overview display.

Path content must be carefully selected or 'curated' by the path-creator, with the addition of context and interpretation so that the objects within the path convey a narrative or meaning. Content may be derived from one collection, but there are significant benefits from including objects from diverse collections, along with other materials from external web sites. It may also be beneficial for interpretation of the path content to be extended by user-generated content and/or annotations of various kinds.

Many of these characteristics are seen in existing path systems, but limitations arise from the linearity that is commonplace. Exploration and deeper levels of engagement within collections requires more complex path structures, carefully curated content, interpretation and narrative, and interconnectedness of paths and other content within and outside of the system. The fact that most of these more advanced characteristics are rare, and that linearity prevails also suggests that these are complex issues yet to be adequately resolved.

4.4 Potential Applications of Paths

Many opportunities for the use of paths in CH were suggested. Two major themes emerging from these are the use of paths to achieve learning, and to support exploration and browsing. For learning to occur there needs to be strong contextual information, along with questions and other exercises to structure the learning process. Exploration and browsing activities implicitly enable meaning making and learning to take place, as users become more familiar with a topic and select or interpret the objects they encounter.

Specific instances of learning activities that may be delivered via paths are collection or subject familiarization, story-telling, individual or collaborative inquiry-based learning utilising path creation, modeling the research process, and comparative analysis of differing view-points on a topic of interest.

In addition to learning,, paths may also serve to deliver entertainment and an enjoyable interaction experience for more general audiences. In practical terms, paths may simply be used as a means of introducing people to a collection and its stories, and in due course, encouraging them to venture further in a more independent fashion. Paths facilitate topic-based information retrieval typified by the berry-picking mode of interaction [2], rather than known item searching. Furthermore, paths may be a useful tool for personal information management in both formal and informal research scenarios, enabling the user to record, reuse and share their research activity, or helping them to organize their ideas. Creativity is also encouraged, as user-generated paths provide the means to repurpose CH objects into users' own narratives for private or public consumption.

5. USER INTERACTIONS WITH PATHS

By consolidating findings across the various data collection methods, we were able to discern five core elements of interaction with CH collections relating to activities that encompass creating, using and sharing paths as a means of exploration and engagement.

Findings from the qualitative data collected via interviews and path-creation tasks revealed a set of five core activities relating to the creation, use and sharing of paths; developing a concept for a path; collecting items in include in a path; creating a path from items collected; communicating about paths found and about paths created; and, consuming (following or exploring) paths created by others. All elements of the model may be undertaken by expert and non-expert users, in any sequence, and with varying degrees of iteration, according to the user's preferences and behavioural traits.

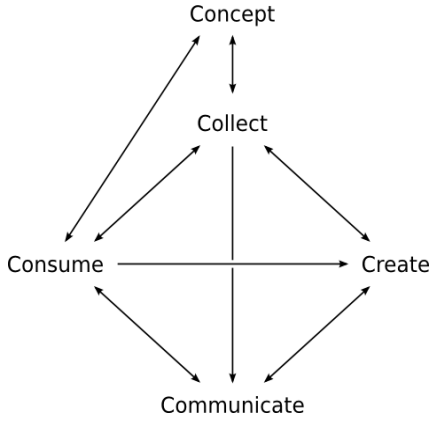


Figure 1. PATHS user interaction model.

Initially, we expect users to begin by *Consuming* paths created by others, using them as a means of exploration and familiarization with the collection and the system. *Collecting* items of interest when exploring a collection is a natural behaviour in berry-picking mode, and is implicit in the process of creating a path, or as a by-product of a user's information seeking history. When the path creation activity is purposeful, it is likely that an over-arching *Concept* is devised, which may come from activities undertaken outside of the system, but also may be developed via a process of exploration within the collection and any pre-existing paths. The concept may also evolve alongside the collection and path creation activities, through a process of iteration and meaning-making. A path is *Created* once a number of appropriate items have been collected, and this activity may include ordering the items into a narrative, and adding contextual information and/or metadata. In a web 2.0 environment, it is also important to allow for *Communication* activities in support of the interaction experience. These may include sharing paths that have been

created or discovered, both within and outside of the system (e.g. via social media), commenting on and rating content, and adding narrative to personal paths as a means of making meaning.

It is imperative in an adaptive web environment that systems do not prescribe modes of interaction or enforce sequences of activities. During the design of PATHS we have uncovered four primary interaction modes, all of which are supported by the user interaction model, but each with a somewhat different typical interaction flow.

Path consumers are the most passive users, and likely to be in the majority. By using paths as a guided tour or means of simple exploration of the collection and its content, we expect users to become more interested in communicating their discoveries with others and exploring further within the main collection. Over time we would expect some of them to move onto collecting and creating paths of their own, as they develop into more independent and active users of the system.

Path creators will likely be in a minority in the early stages, and primarily expert users such as curators and educators, and perhaps a few more independent non-expert users. In *expert* path-creation mode we believe interaction will be purposeful and systematic, with a goal of creating a path about a defined topic. Topics and styles of paths may vary by domain, and we expect that educators are more likely to adapt ideas from existing paths, whilst CH experts will try to develop something novel, showcasing elements of a collection or subject expertise. In contrast, *non-expert* path creators are more likely to develop their concept as they explore the collection, and their paths may be more idiosyncratic, evolving over time, or in the education domain, may even be directed in the task by an expert in a path facilitator role.

Path facilitators are most likely to be found within educational settings, where inquiry-based learning is prevalent. These users may not create paths themselves, but may curate a broad collection of objects from which a group of non-expert users are encouraged to create their own paths. (for instance, as a homework project). Facilitators are more interested in enabling deeper engagement with CH materials, and in fostering communication and reflection on the activity and the content of the paths created in this way.

6. CONCLUSIONS

We have presented the findings of our user requirements study on the creation and use of paths as a means of aiding information access and exploration in CH digital collections. It has been shown that paths support many of the needs for exploratory information behavior, and have applications for diverse users across multiple domains. Users interactions with paths comprise five core elements, integrated into a single user interaction model and can be used in varying sequences, illustrated by four primary modes of interaction.. An initial prototype has been developed from the user interaction model, which is currently being evaluated within a task-based user-centred evaluation setting.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°. 270082. We acknowledge the contribution of all project partners involved in PATHS (see: http://www.paths-project.eu).

8. REFERENCES

[1] Amin, A., et al. 2008. Understanding cultural heritage experts' information seeking needs. In *Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (Pittsburgh, PA, June 16-20, 2008) JCDL'08, ACM New York, NY. DOI= http://dx.doi.org/10.1145/1378889.1378897

[2] Bates, M. 1989. The design of browsing and berry picking techniques for the online search interface. *Online Review*, 13:5, 407-431.

[3] Bush, V. 1945. As We Think. *Atlantic Monthly*, Boston.

[4] Camhi, J. 2008. Pathways for communicating about objects on guided tours. *Curator: The Museum Journal* 51, 3, 275-294.

[5] Erdelez, S. 1997. Information encountering: a conceptual framework for accidental information discovery. In: Vakkari, Savolainen & Dervin, eds..*Proc. Information seeking in context: 14-16 August, 1996, Tampere, Finland*, 412-421.

[6] Goodale, P., Hall, M., Fernie, K., and Archer, P. 2011. Paths Project D1.1 User Requirements Analysis. http://www.paths-project.eu/eng/Resources

[7] Koolen, M., Kamps, J. and de Keijzer, V. 2009. Information retrieval in cultural heritage. *Interdisciplinary Science Reviews, 34*, 2-3, 268-284.

[8] Kruschwitz, U., et al. 2011.. Moving towards adaptive search in digital libraries. In: *Advanced Language Technologies for Digital Libraries*. Springer.

[9] Marchionini, G. 2006. Exploratory search: from finding to understanding. *Communications of the ACM* 49, 4, 41-46.

[10] Peterson, D. and Levene, M. 2003, Trail records and navigational learning, *London Review of Education* 1, 3.

[11] Schraefel, M.C., Zhu, Y., Modjeska, D., Wigdor, D, and Zhao, S. 2002. Hunter gatherer: interaction support for the creation and management of within-web-page collections. In *Proc. 11th International Conference on World Wide Web* (Hawaii, May 7-11, 2002). WWW'02 ACM, New York, NY. DOI= http://dx.doi.org/10.1145/511446.511469

[12] Shipman III, F.M., Furuta, R., Brenner, D., Chung, C., and Hsieh, H. 1998. Using paths in the classroom: experiences and adaptations. In *Proc. 9th ACM Conference on Hypertext and Hypermedia* (Pittsburgh, PA, USA, June 20-24 1998), HT'98. ACM, New York, 267-270. DOI= http://dx.doi.org/10.1145/276627.276656

[13] Shipman, F., Furuta, R., Brenner, D., Chung, C., and Hsieh, H. 2000. Guided Paths through Web-Based Collections: Design, Experiences, and Adaptations. *Journal of the American Society of Information Science* 51, 3, 260-272.

[14] Skov, M. & Ingwersen, P. 2008. Exploring information seeking behaviour in a digital museum context. In *Proc. 2nd Int. Symposium on Information Interaction in Context* (London, October 14-17, 2008). IIiX 08, ACM, New York, NY. DOI= http://dx.doi.org/10.1145/1414694.1414719

[15] Walker, K. 2006. Story structures: building narrative trails in museums. *Technology-mediated Narrative Environments for Learning*, 103-111.

Supporting Serendipitous and Focused Search

Junte Zhang
Meertens Institute, Royal Netherlands Academy of Arts and Sciences
Amsterdam, the Netherlands

ABSTRACT

People with complex information needs are for example Humanities researchers, who need advanced search engines to investigate their research questions. Much can be gained by combining research datasets, reusing tools and serendipitously discovering new insights for further research. Humanities researchers have different (large-scale) research datasets and tools, which are described differently with metadata.

We present a highly interactive advanced search engine for Humanities researchers that semantically converges differently structured metadata records from different collections and institutions. It has features that support serendipitous and focused search in context based on the structure of the metadata used. This single system serves Humanities researchers by allowing them to search interactively across yet unexplored (research) data, discover patterns, locate relevant data for new insights, and find existing tools that could provide novel use cases.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital Libraries**]: Systems issues, user issues; H.5.2 [**Information interfaces and presentation**]: Graphical user interfaces (GUI)

General Terms

Design, Human Factors

Keywords

information retrieval, metadata, user interfaces, ehumanities

1. INTRODUCTION

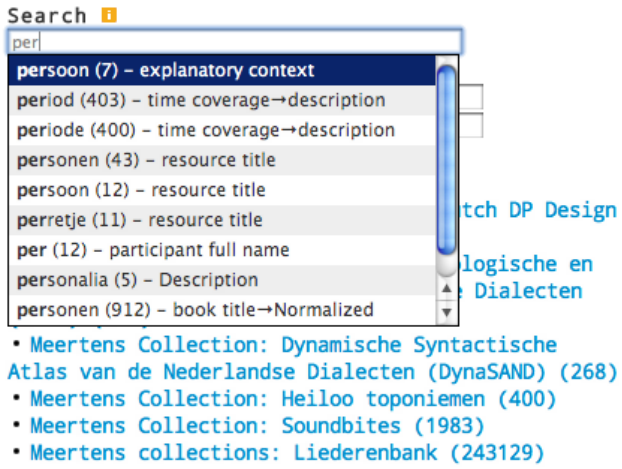
The Common Language Resources and Technology Infrastructure (CLARIN) initiative seeks to establish an integrated and interoperable research infrastructure of language resources and its technology.[1] Descriptive metadata is used to characterize large number of (legacy) research data resources (collections) and tools (e.g. Web services) to facilitate their management and discovery. The Search & Develop (S&D) project within CLARIN in the Netherlands uses the Component MetaData Infrastructure (CMDI; [4]) with ISOcat [6, 12] to open up the sharing of resources and Web services for people and machines first within the collections of a single institution, then across institutions in the Netherlands and eventually across Europe as whole. This infrastructure enables new research methods in language research and stimulates the Digital Humanities, where new insights can be gained by combining and reusing resources from different institutions and domains, and existing tools can be more effectively found and reused based on new insights.

How to use the CMDI framework with ISOcat to search for data and services, which can be understood by both people from varying disciplines and machines? The challenge is that the data is heterogenous both in content and structure, and can be massive in amount. In [11], we show how to deal with such heterogeneously structured data in the CMDI MI Search Engine. Users of the CMDI framework are mostly Humanities researchers. What type of system is needed driven by CMDI that matches with the search behavior of these users? This paper presents a proposition that has been implemented on a live system.

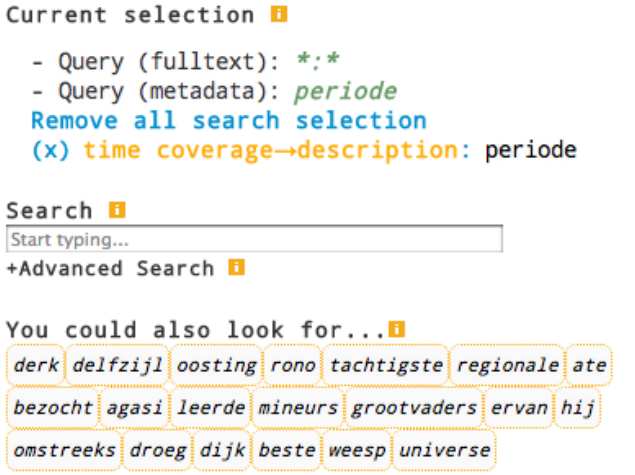
2. USING CMDI FOR FOCUSED AND SEMANTIC ACCESS

CMDI has grown out of the need to facilitate access, reuse, and interoperability using metadata [4]. A CMDI file in XML consists of a `<Header>`, `<Resources>`, and `<Components>`. The former two are fixed in structure, while the content and structure within `<Components>` is flexible and can encapsulate any data in any structured form. An XML schema can be used to make CMDI files coherent in structure for a (sub)collection and it contains references to ISOcat data categories (DC) stored in the Registry (DCR; [7, 6]). The DCR was established by the *ISO Technical Committee 37, Terminology and other language and content resources* based on the ISO 12620:2009 standard. Because multiple elements may refer to the same DC, semantic interoperability can be achieved across different datasets. A specification using the DCR and projected for example in an XML schema is called a metadata *profile* and can be (re)used for describ-

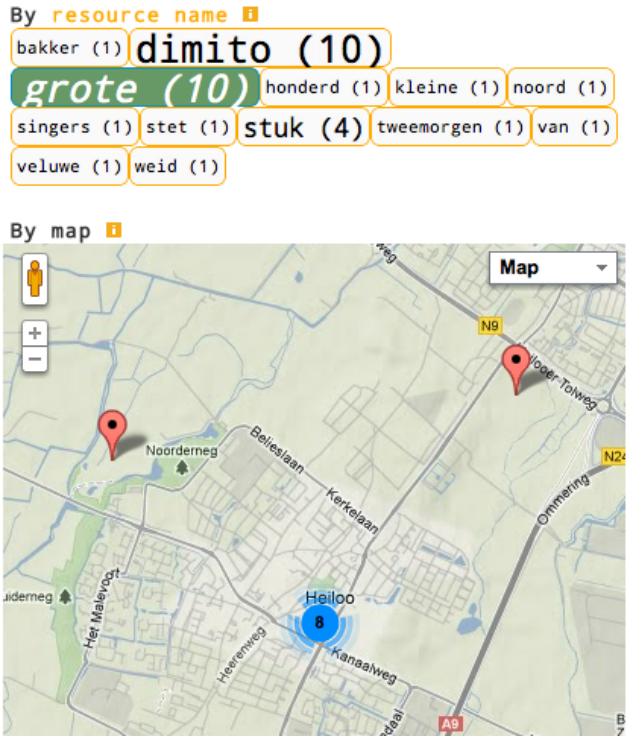
[1]See http://www.clarin.eu/external/index.php?page=about-clarin



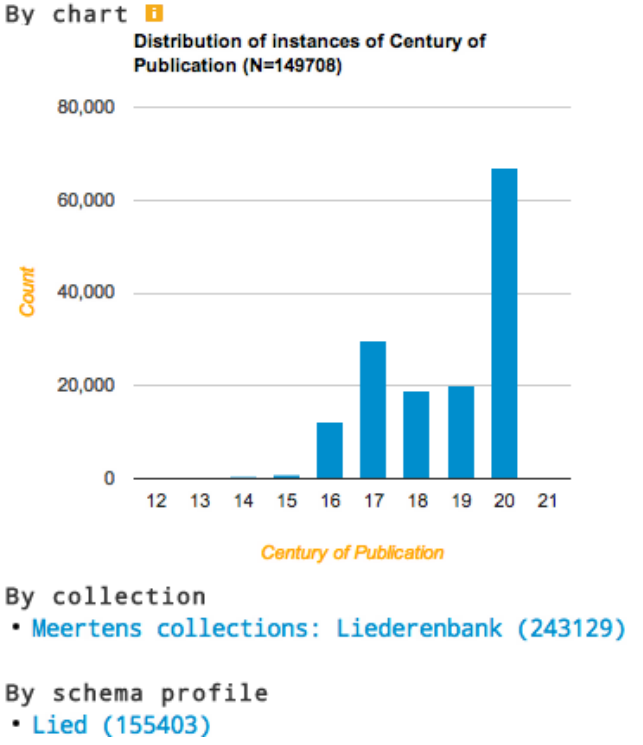
(a) Query autocompletion based on the count that a query occurs in a tag within the result set. By default the query box is content-centric, but searching directly in a tag is possible with Advanced Search (can be collapsed with a click). Users can express queries using the metadata or only the fulltext of the document by discarding autocompletion.



(b) The selection widget that allows users to keep overview of the search trail and change it, while updating the result list. Here, the query stored is "periode" (*period*) within the tag *time coverage→description*. Interesting terms are suggested by presenting the top TF∗IDF terms, which people can use to start a parallel search episode.

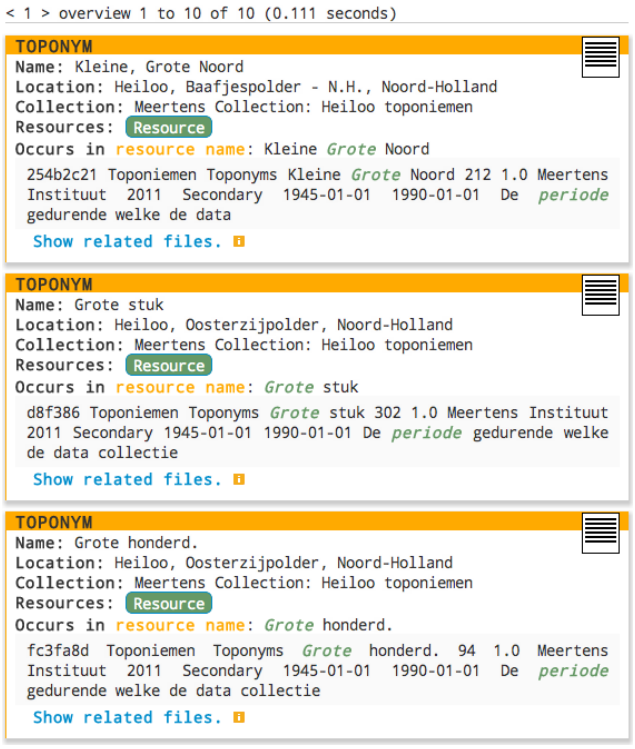


(c) To further support query expansion and serendipitous information seeking, a dynamic tag cloud is generated based on the last retrieved result list and used metadata label with keyword highlighting. Moreover, retrieved geo-referenced documents are projected on a map and clustered by markers.

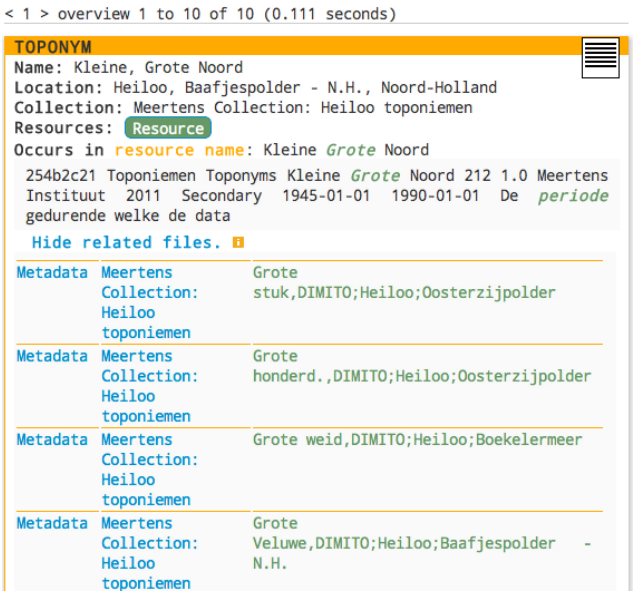


(d) The distribution of retrieved time-referenced documents (given the tags *Century of Publication* and *Year of Publication*) are visualized in bar or line charts. Users can click in the charts to narrow down the result set. The distribution of results in tags *collection* and *schema profile* always appear.

Figure 1: The CMDI MI Search Engine (1).



(a) Retrieved list of results with the display of the list of results with 'fixed' contextual information, snippets and keywords in context within the last searched metadata label and the presentation of all used keywords in context given the fulltext. There is links to the fulltext of the metadata record and the actual resource in the digital archive.



(b) For each retrieved result in the list, there is a recommendation (when available) of related results based on the content similarity of the last used metadata label. A recommendation consist of a link to the record, the collection it belongs to, and a snippet (can be collapsed with a click).

Figure 2: The CMDI MI Search Engine (2).

ing datasets and for eventual access. Moreover, RELcat [10] goes a step further by allowing for the storage of arbitrary relationships between data categories to assist crosswalks and to specify ontological relationships for further semantic search, which in the future can be used in the CMDI MI Search Engine using field collapsing.

We have indexed 246,728 CMDI files from 18 different profiles consisting of 143 different types of elements in a single stream, which shows our indexing method for CMDI files is robust enough to deal with complex data [11]. By indexing metadata in CMDI on the XML element level, the search engine can provide focused access [8]. We use straight-forward information retrieval techniques only. The 'Liederenbank' (*Dutch Song Database*) alone has 9 different profiles (XML schemas), which is equivalent to a sub-collection, ranging from very differently structured descriptions about songs to singers. How to provide interactive access to such heterogeneously structured data for Humanities researchers?

3. SERENDIPITY IN CONTEXT

When a user with no a priori intentions interacts with a node of information and acquires useful information, then serendipitous information retrieval occurs [9]. The success of serendipitous discovery is not just the find itself, but being able or willing to do something with it, so that users get more insight and can enhance the domain expertise [1]. Humanities researchers are the type of users who can be greatly supported in their research tasks with serendipitous IR, because their information-seeking behavior can be described as an idiosyncratic process of constant reading, "digging," searching, and following leads [2]. This confirms with the Berrypicking model of [3], such as that queries are not static, but rather evolve, and users "gather information in bits and pieces instead of in one grand best retrieved set."

Since the CMDI MI Search Engine should serve Humanities researchers, we design it to support serendipitous search and be highly interactive. The system has been designed to maximize the user's ability to explore. This is our focus. The user interface of the system is depicted in Fig 1. It uses the JavaScript library AJAX Solr[2], which has been heavily modified and extended by us with JQuery. It allows for faceted search [5] as we treat the indexed elements of the CMDI files as one large category hierarchy.

A user can improving the search episode (session) by effectively reducing the information space step by step. These steps are stored as part of the search trail, so the overview is kept. There are different search strategies possible. Users can search by fulltext by entering a query. This makes sure users can always search in everything. The query get highlighted in context given the fulltext, but the dynamic tag cloud widget that supports query expansion is not activated, see Fig.1(a). Users can also do a focused search request by using structure, i.e. within the content of a specified tag, and get the content of these tags returned. This can be content-centered, as users enter a keyword and the autocompletion widget returns a list consisting of keyword plus field name and hit count. It can also be structure-centered (using the Advanced Search option) by looking up a tag and then entering a keyword also with the autocompletion feature. When the last two options are used, then the keyword highlighting also occurs within the context of the retrieved

[2]See https://github.com/evolvingweb/ajax-solr

snippets of the searched tag, see Fig.2(a).

A challenge is how we can support serendipitous search given the diversely structured metadata in CMDI. Hence, we introduce and propose the concept of serendipitous search in context. We can use the heterogeneous structure of different collections to provide context to the user in a single search engine. We propose the following contextual system features that aim to support serendipitous and focused search.

- Help users by automatically completing the query that the user is entering while simultaneously and directly giving the hit count for the suggested queries in conjunction with a tag, see Fig.1(a).
- Provide inline suggestions (*Did you mean...*) based on a spell checker whenever applicable.
- Suggest a new parallel search episode (*You could also look for...*) by presenting interesting terms based on the content of the first few retrieved results after each used query, see Fig.1(b). This increments and becomes more focused as a search episode gets more queries.
- Offer different overviews of the retrieved results and allow for query expansion by directly presenting a dynamic tag cloud of the aggregated content within the metadata label used and highlighting the query entered in this context, see Fig.1(c).
- Preserve the overview of a search episode by storing the search selection (see Fig.1(b)), and the overview on collection level by the result type, e.g. the metadata profile 'lied' (*song*) in the Dutch Song Database, and the collection a document belongs to (see Fig.1(d)).
- Aggregate and visualize collection-specific search features in extra widgets, such as projecting and clustering the list of retrieved geo-referenced resources on a map (see Fig. 1(c)), and displaying the date ranges of the documents in charts that can be clicked to narrow down a result set (see Fig. 1(d)).
- Entice users to explore further by recommending related resources using the content similarity by presenting a link to the metadata record and a snippet of a recommendation, see Fig.2(b).

So the context consists of different modalities and features existing in the structure of the metadata of a collection, and used in the retrieval and visualization of information. This can be displayed on a aggregated level based on the set of retrieved results. And it can be displayed with different displays of the result types given the metadata profile. Eventually, the user finds the links to the resources in the digital archive using the metadata, and can use the found resources for further research or development. However, there is no real definite end of the search episode as people still can continue searching using the above proposed system features.

4. CONCLUSIONS

We have presented a working proposition for serendipitous and focused search by describing the CMDI MI search engine. The novelty is that it provides semantic access to diversely structured language and digital heritage resources with different metadata schemas for users such as researchers with very specific and complex information (research) needs. The search engine provides faceted search and has serendipitous features that maximize the user's ability to explore any metadata in CMDI in context, such as query autocompletion, tag clouds, and recommendation of related resources, while keeping track of the search trail. It is a tool that provides interactive and focused access to heterogeneous metadata, gives new perspectives on legacy (research) data and tools, and provides new insights for research and development. It has been released as live, and can be used at www.meertens.knaw.nl/cmdi/search.

5. ACKNOWLEDGMENTS

This work is part of the Search & Develop project at the Meertens Institute, and funded by CLARIN-NL.

6. REFERENCES

[1] P. André, m. schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: designing for (un)serendipity. In *Proceedings of the seventh ACM conference on Creativity and cognition*, C&C '09, pages 305–314, New York, NY, USA, 2009. ACM.

[2] A. Barrett. The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*, 31(4):324 – 331, 2005.

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

[4] D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In *LREC*, 2010.

[5] M. A. Hearst and C. Karadi. Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *SIGIR*, pages 246–255, New York, NY, USA, 1997. ACM.

[6] M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. E. Wright. ISOcat: remodelling metadata for language resources. *IJMSO*, 4(4):261–276, 2009.

[7] M. Kemps-Snijders, C. Zinn, J. Ringersma, and M. Windhouwer. Ensuring semantic interoperability on lexical resources. In *LREC*, 2008.

[8] M. Lalmas. *XML Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[9] E. G. Toms. Serendipitous information retrieval. In *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[10] M. Windhouwer. RELcat: a relation registry for isocat data categories. In *LREC*, 2012.

[11] J. Zhang, M. Kemps-Snijders, and H. Bennis. The CMDI MI Search Engine: Access to language resources and tools using heterogeneous metadata schemas. In *TPDL*, volume 7489 of *Lecture Notes in Computer Science*. Springer, 2012.

[12] C. Zinn, C. Hoppermann, and T. Trippel. The isocat registry reloaded. In *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 285–299. Springer Berlin / Heidelberg, 2012.

Vague Query Formulation by Design

Marcus Nitsche
Faculty of Computer Science,
Otto-von-Guericke-University,
Germany
marcus.nitsche@ovgu.de

Andreas Nürnberger
Faculty of Computer Science,
Otto-von-Guericke-University,
Germany
andreas.nuernberger@ovgu.de

ABSTRACT

When users search for information in domains they are not familiar with, they usually struggle to formulate an adequate (textual) query. Often users end up with repeating re-formulations and query refinements without necessarily achieving their actual goals. In this paper we propose a user interface that is capable to offer users flexible and ergonomic interaction elements to formulate even complex queries simple and direct. We call this principle *vague query formulation by design*. By this formulation we like to point out its design-driven origin. The proposed radial user interface supports phrasing and interactive visual refinement of vague queries to search and explore large document sets. The main idea is to provide an integrated view of queries and related results, where both queries and results can be interactively manipulated and influence each other. Changes will be immediately visualized. The concept was implemented on a tablet computer and the usability was stepwise evaluated during a formative and a summative user study. The results reveal high usability ratings, even if the concept was completely unknown to our test users.

Keywords

Search User Interface, Query Reformulation, Query Refinement, Information Retrieval.

Categories and Subject Descriptors

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

General Terms

Design, Human Factors, Management.

1. INTRODUCTION

When users try to handle complex information needs they often end up in conducting exploratory searches [11]. One of the main characteristics of exploratory searches is that users often do not know how to formulate their information need. Often this problem coexists with an unfamiliarity with the domain they search in [17]. In this work we like to tackle this problem of querying appropriate queries by offering users dynamic user interface (UI) elements that they can manipulate directly by touch gestures to give them *a feeling* for a certain query configuration that matches a certain result set. Thereby learning and exploring aspects will covered as well [17, 11]. This concept of interactive visual filtering of relevant information in a more natural way enables data processing in cases, where standard algorithms can not be applied since these algorithms might filter out relevant data. We introduced the concept of this paper back in 2011 [15], where we described the basic idea and did some pre-studies with a digital mockup prototype. In this paper, we first introduce a running implementation and a more detailed user study towards this concept. Therefore we present some related work aspects in Section 2, followed by a presentation of the UI concept in 3 and the description of the implementation, evaluation concept and results of the final user study in Section 4. Finally, we conclude and discuss possible future work in 5.

2. STATE-OF-THE-ART

User-specific context aware data filtering is not a new challenge. In the following we show two tools, that can also be used for these application domains. The VIBE-system [10, 16] supports users in finding relevant information using magnets to attract relevant documents to specific screen points (Fig. 1).

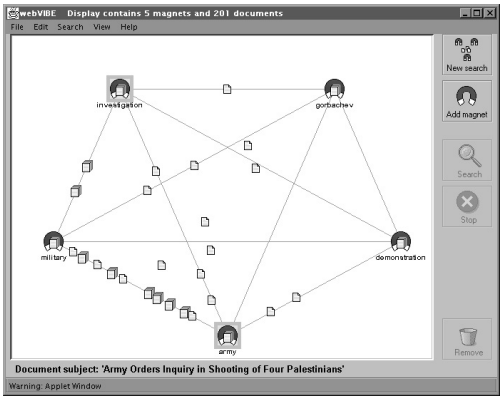


Figure 1: webVIBE, a variant of [10, 16].

This system follows the principle of *dust-and-magnet* [18]. Our proposed concept uses this principle also as one aspect of interaction. In contrast to VIBE we offer users of our system an interactive visualization without any classical WIMP-interface elements

(Windows, Icons, Menus, Pointer). By this, no virtual mapping of functions is necessary and users might be able to use the interface in a more firm and reliable way. Cousins et al. [5] developed a system that follows a direct manipulation approach like done here. But in contrast to our proposed solution it is divided into different UI elements and different views. It is less integrated in a single view. Therefore user's work load might be higher since he needs to face various mode switches. Commercial systems, like the Vis4you concept[1], are more focused on visualization than on interaction via direct manipulation. Furthermore, this system is designed to be used on desktop computers with a mouse (*single point and click*-principle), no multi-touch-support. In the next section we like to present our concept in more detail.

3. CONCEPT & DESIGN

Due to the increasing amount of data and complexity, it is necessary to apply and improve the concepts of visual information filtering and retrieval. This goes along with the underlying methods and tools. Considering clustering algorithms (e.g., k-means [3]), we thought about the concept of *vague query formulation*: Since users sometimes do not know what they are searching for, we like to support them by the opportunity to formulate vague queries. Here the user is asked to narrow the search results by dragging user interface (UI) elements, so called widgets, with query terms, see also Fig. 2.

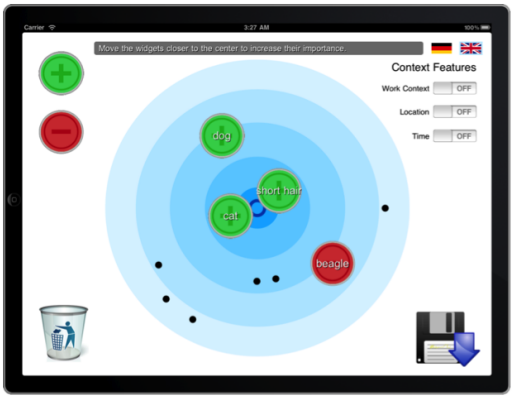


Figure 2: Radial design of the implemented UI.

The concept follows the idea that more relevant data are centred. Note, this is equivalent to filtering an overcrowded desktop, cf. Fig. 3 (left picture)[2], where the more centralized documents are possibly more important (highlighted in the right picture).

The system was designed to be a multi-user system. Therefore a number of multiple users need to be supported at the same time, also considering security aspects [14]. To offer each user the same possibility to interact with the system we use for the interface a radial form. Furthermore, an underlying multi-touch device is a hardware requirement, that enhances the combination of tool and application domain significantly. Another appealing advantage is, that multi-touch also supports users in a more natural way of interaction [9]. Other radial user interfaces for selecting or filtering often offers fixed places for items. In contrast to this our system is supposed to be more flexible since users are allowed to position UI elements where they like.

We offer users a dimension merging according specified weights, similar to the result listing of search engines, where also different

[1]http://www.vis4you.com/vis4you/ (accessed on 04.07.2012)
[2]http://lawprofessors.typepad.com/ (accessed on 04.07.2012)

Figure 3: Desktop: more relevant documents are centred.

weights can be linked to specific query terms (Fig. 2). Data points represent the data space. Query objects, so called widgets, can be entered via a virtual keyboard and can also be dragged by the user to formulate more complex or vague queries. Selecting a specific data point supports the user with additional information on this data point and highlights all related data points.

The distance of a certain term is directly connected to its importance for the user. In other words, if a user thinks a specific term is more relevant to its actual filter-/search-task, she or he positions the corresponding UI-element more to the center, which influences the weight of this term when computing its Term Frequency / Inverted Document Frequency (TF/IDF)-value [2], which in fact is a calculated weight to influence the ranking of the data space and this in return the visualization (Fig. 6). Thereby, users do not need to specify a concrete position of UI elements on the screen, we support this by a non-determined precision. The widget-induced relevance of a query term is calculated according to the formula in Fig. 4.

$$R_{Widget} = \frac{d_{Center\ Point}}{r_{Search\ Area}}$$

Figure 4: Widget-induced relevance of a query term.

Result elements are placed near to corresponding query elements. The formula for calculating the relevance of a SearchResult object (result dot) is shown in Fig. 5.

$$R_{Search\ Result} = \frac{\sum_{i=0}^{\#\ relevant\ widgets} R_{Relevant\ Widget_i}}{\#\ relevant\ widgets} - \left(\frac{\sum_{i=0}^{\#\ non\ relevant\ widgets} R_{Non\ Relevant\ Widget_i} + \sum_{i=0}^{\#\ relevant\ exclusion\ widgets} R_{Relevant\ Exclusion\ Widget_i}}{\#\ non\ relevant\ widgets + \#\ relevant\ exclusion\ widgets} \right)$$

Figure 5: Relevance of a SearchResult object.

The calculated relevance determines the distance to the center, considering further result objects.

To address various types of end devices such as multi-touch desktops or mobile interfaces with large displays, we use direct manipulation as a central interaction paradigm. Only the relative distance of an UI element to the center is relevant for the system. Thus, we provide users with a direct linking to the data they like to filter. By this interaction concept, we propose to achieve more precise results. Additionally, we support users with the concept of *What-if*-queries, which supports a fault-tolerant interaction system, using a ghosting technique: Dragging an element and holding it on a specific position triggers the system to show the user how many items are in the center point of interest (POI) after releasing the element. Thereby, users are able to explore the impact of possible next steps.

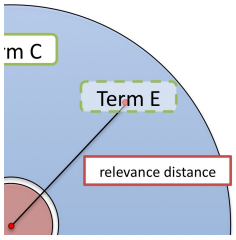


Figure 6: Concept of relevance mapping.

Changes of the query configuration also effect the data points to provide the user with a direct link to the data (interactive visualization). By the underlying metaphor of magnets, we offer an integrated feedback, comparable to *Dust-and-Magnet* [18]: When users drag a specific UI element to a certain point, relevant data points follow this UI element. Data points that have the same TF-IDF value (equal relevance to a query configuration) are drafted with a minimal distance to each other to prevent occlusions.

3.1 Features

The UI supports direct feedback since the relevance value is simultaneously shown while users interact with the widget (Fig. 7).

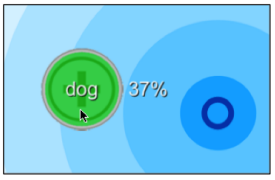


Figure 7: Direct feedback: relevance value next to the widget.

Results, corresponding to a specific query object are visually highlighted and grouped to each other (Fig. 8).

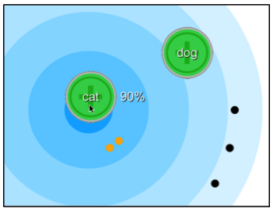


Figure 8: Corresponding results are visually highlighted to group them (e.g. highlighted results for the search term 'cat').

Detailed information on particular result objects, like a website preview, is provided after clicking on the result dot (Fig. 9).

4. IMPLEMENTATION, EVALUATION & RESULTS

Since this contribution is basically driven by fields of human factors and user interface design, we are using common methods from these research areas. Such as user centred design (UCD) processes [7], formative evaluation methods [12], questionnaires [6], think-aloud-protocols [8], and cognitive walkthroughs [4].

To proof the concept of the proposed user interface, a prototype was implemented. This was done by using an Apple iPad. Therefore the application was written in ObjectiveC using the xCode[3] environment. The backend architecture is the CARSA system [1], an information retrieval framework for research purposes. For a detailed overview about the system's architecture see Fig. 10.

Figure 9: Prototypical search result popover as a website preview feature, here a result for 'Labrador Retriever'.

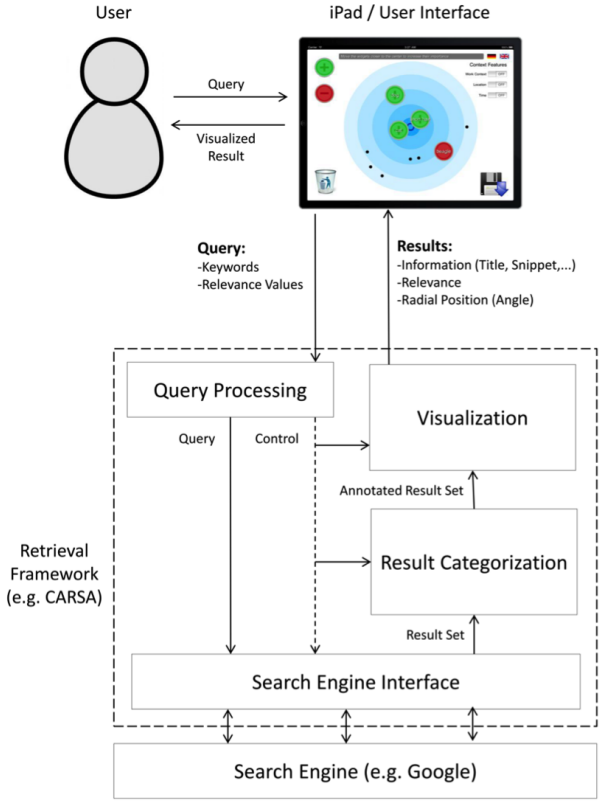


Figure 10: System architecture & UI interaction, cf. [1].

The evaluation concept followed a formative evaluation principle where several usability testings were conducted. Also in parallel to the development process: To identify at least 85% of all usability issues this mock-up was evaluated according to Nielsen and Landauer [13] with only a small number of test users since most usability issues will be mentioned repeatedly by users. The sixth tested user would report new usability issues in only 15%

[3]developer.apple.com/xcode/ (accessed on 04.07.2012)

of all cases. Therefore we decided to ask only eight users. The results of this first user test seem to be promising that this concept works as desired. Users were introduced in the main features and were asked afterwards to formulate a filter query consisting of three terms to find all relevant documents while visualizing most important relations to other potential interesting data. After going through a cognitive walk-through of a movie filtering task our eight test users (six male, two female, average age: 23.4) answered seven usability questions by filling out a 7-step Likert scale from 1 (very bad) to 7 (very good). Next to cognitive walk-throughs, we used think-aloud-protocols and questionnaires. The usefulness of the prototype was rated high, the functionality was praised by test users, performing tasks were rated as *very easy* and test users were satisfied with this prototype. Terminology, attractiveness, and consistency were rated lower. Our final evaluation revealed the results you can see in Fig. 11. Even if there is room for improvement the results reveal overall a good usability, several test users mentioned that it was fun to use it, which might is reflected by a high rating of *joy of use* measurings.

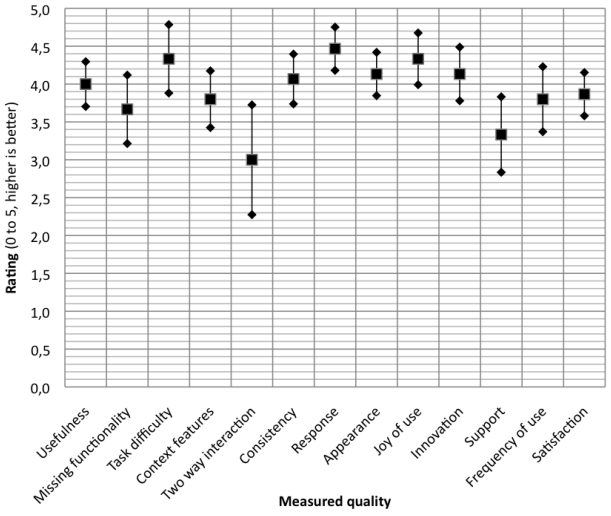


Figure 11: Results of final usability testing.

5. DISCUSSION & OUTLOOK

We described a newly designed user interface for filtering, exploring and managing data via direct manipulation supporting multiple reference systems to support context sensitive interaction techniques. We proposed a UI concept for visual filtering, that is

- flexible: parameters can be adapted or enhanced by users
- context-sensitive: initial parameters are extracted from the current use case
- easy to learn: through work environment metaphor and direct manipulation

In near the future a more detailed and larger user study will be conducted to identify further improvements of our tool and the overall concept. Also a plan to re-design slightly is already in place.

6. ACKNOWLEDGEMENT

Part of the work is funded by the German Ministry of Education and Science (BMBF) within the ViERforES II project (no. 01IM10002B). We also thank Martin Schemmer for the implementation of the presented concept during his diploma thesis.

7. REFERENCES

[1] Bade, K., De Luca, E. W., Nürnberger, A., Stober, S.: CARSA - an architecture for the development of context adaptive retrieval systems. In: Proceedings of Adaptive Multimedia Retrieval: User, Context, and Feedback, Volume 3877/2006, Lecture notes in computer science, pp. 91-101, Springer Berlin / Heidelberg (2006).

[2] Baeza-Yates, R. and Ribeiro-Neto, B.: Modern Information Retrieval, pp. 29-30. Addison Wesley / ACM Press, NY (1999).

[3] Bradski, G. and Kaehler, A.: Learning OpenCV Computer Vision with the OpenCV Library. O'Reilly, p. 479 (2001).

[4] Busemeyer, J. R.: Choice behavior in a sequential decision-making task. In: Organizational Behavior and Human Performance 29 (2), pp. 175-207(1982).

[5] Cousins, S. B., Paepcke, A., Winograd, T., Bier, E. A., Pier, K.: The digital library integrated task environment (DLITE). In: Proceedings of the second ACM international conference on Digital libraries (DL '97). ACM, New York, NY, USA, pp. 142-151 (1997).

[6] Czaja, R. and Blair, J.: Designing Surveys. Pine Forge Press. A useful resource for factual-style surveys, including material on interviews as well as mail surveys (1996).

[7] Eason, K. D.: User centred design for information technology systems. In: Physics in Technology 14 (5), p. 219 (1983).

[8] Ericsson, K. A. and Simon, H. A.: Verbal reports as data. In: Psychological Review 87 (3), pp. 215-241 (1980).

[9] Han, J. Y.: Multi-touch interaction wall. In: Proceedings of ACM SIGGRAPH 2006 Emerging technologies (2006).

[10] Koshman, S. L.: VIBE User Study Technical Report LS062/IS97001, University of Pittsburgh (1997).

[11] Marchionini, G.: Exploratory search: from finding to understanding. In: Communications of the ACM 49 (4), pp. 41-46 (2006).

[12] Moxley Jr., R. A.: Formative and non-formative evaluation. In Instructional Science 3(3), pp. 243-283 (1974).

[13] Nielsen, J. and Landauer, T. K.: A mathematical model of the finding of usability problems. In: Proceedings of ACM INTERCHI'93 Conference, pp. 206-213, Amsterdam, The Netherlands (1993).

[14] Nitsche, M., Dittmann, J., Nürnberger, A., Vielhauer, C., Buchholz, R.: Security-relevant Challenges of selected Systems for Multi-User Interaction. In Proceedings of the 7th International Workshop on Adaptive Multimedia Retrieval (2011).

[15] Nitsche, Marcus and Nürnberger, Andreas: Supporting vague query formulation by using visual filtering. In Proceedings of Lernen, Wissen, Adaption (2011).

[16] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G.: Visualization of a Document Collection: the VIBE System. In: Information Processing & Management, 29(1), pp. 69-81 (1993).

[17] White, R. W. and Roth, R. A.: Exploratory search: Beyond the Query-Response paradigm. In: Synthesis Lectures on Information Concepts, Retrieval, and Services, Ed.: G. Marchionini, Morgan & Claypool Publishers (2009).

[18] Yi, L. S., Melton, R., Stasko, J. and Jacko, L.: Dust & Magnet: multivariate information visualization using a magnet metaphor. In: Information Visualization, pp. 239–256 (2005).