Presenting Query Aspects to Support Exploratory Search

Mingfang Wu, Andrew Turpin, Simon J. Puglisi, Falk Scholer and James A. Thom

School of Computer Science and Information Technology
RMIT University,
GPO Box 2476 Melbourne, Australia, 3001
Email: {mingfang.wu, andrew.turpin, simon.puglisi, falk.scholer, james.thom}@rmit.edu.au

Abstract

Successful information search requires a joint effort from both syntactic matching provided by current search engines and semantic matching performed by human users. Word-based syntactic matching schemes work well for tasks such as homepage finding or fact finding, but they are less effective in supporting exploratory search tasks such as learning and investigation. One way to overcome this limitation of syntactic matching is to capture the search journeys of other users with semantically related queries, and use them as a roadmap to guide exploratory search.

This paper presents our investigation on the utilization of query semantics derived from query logs, to 1) increase the diversity of a search result; and 2) devise new interfaces that display a search result to support exploratory search. We conducted a user study to evaluate our initial interface prototypes. The evaluation shows that, with the interface that explicitly supports their task, subjects acquire more knowledge and are more confident about their task completeness. The differences between subjects' preferences suggest that we may need to provide a range of interfaces that can not only support users' search tasks, but also suit their personal styles.

Keywords: Information Retrieval, Exploratory Search, Search Interface, Web Search, Task based Design.

1 Introduction

A successful information search requires a joint effort from both the syntactic matching provided by current search engines and the semantic matching performed by human users. Word-based syntactic matching schemes work well for tasks such as homepage finding or simple fact finding, but are limited for more complicated needs such as learning and investigating tasks (labeled as exploratory search tasks by Marchionini (2006)) or research searches (Guha et al. 2003). For example, if a high school student is learning about the concept of "acid rain" and is asked to write an informative and comprehensive essay on the topic, a query containing "acid" and "rain" may yield many relevant pages, however if all these pages are about the definition of "acid rain", that would not satisfy the student. With current information retrieval systems (typically represented by Web search engines), the student has to explore and gather information about all possible aspects of "acid rain" by iterating through a process of query reformulation and search result browsing. In the end, the student may be still unable to obtain a broad overview of the issue. To provide better support for such learning or exploratory search tasks, an information retrieval system should go beyond word-based match and the expectation that information needs can be answered with a single query. It should proactively provide users with related information, and an intuitive interface that guides users' search and browsing activities.

In the Web search context, aggregated search logs from many users provide a rich information source for mining aspectual semantic relations about a topic. They provide a link between different aspects of a topic that are semantically related to the original query, but where the documents describing the aspects do not necessarily contain the original query words. For example, we followed and extracted sequences of related queries that were sent before and after the query "acid rain" by nearly 30 users from a query log. As we can see from Figure 1, each query chain indicates a user's search and learning journey about "acid rain" or a particular aspect of "acid rain". While some of the aspects contain the keywords "acid rain", many do not. Representative queries from each chain can be used to populate the search results for the query "acid rain", allowing a new user with the same query to read-and-choose, rather than reformulate the same query or choose from suggested queries, and so make more complex decisions about their next search.

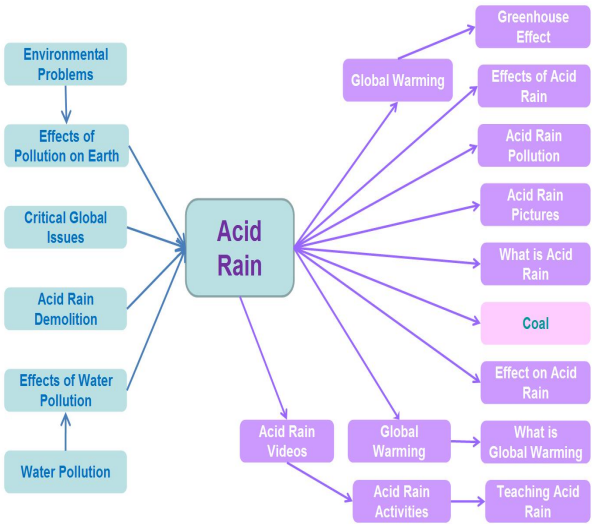

Figure 1: Example of query chains of "acid rain"

Two recent papers study extraction of semantically related queries from a query chain (Radlinski and Dumais 2006, Parikh and Sundaresan 2008). The suggested queries, as provided by major Web search engines (such as Google, Yahoo! and LiveSearch/Bing), include most query aspects of a topic that one would reasonably expect. In our study presented here, we focus on the use of aspectual semantic relations to support a user's interaction in exploratory search tasks. In particular, we investigated how to use aspectual semantic relation to: 1) increase the diversity of a search result; and 2) devise new interfaces that display a search result to support exploratory search.

As we focus on the search result organization and presentation in this study, we follow the typical cycle of user interface design and evaluation. Firstly we analyzed the exploratory search task, and users' search behavior when under taking such a task, through reviewing related work. Next, we designed interfaces based on this understanding, and then conducted a pilot study to evaluate and compare the candidate interfaces with a small number of users and topics. This paper will present this design process and what we have found and learnt through the pilot study.

2 Related Work

2.1 Exploratory Search

An understanding of a user's search context and search intent behind a query is essential to the design and development of effective search algorithms and search interfaces. Marchionini (2006) classified users' search intents into three categories: *lookup*, *learn* and *investigation*. When users perform a lookup search activity, they know exactly the resource for which they are searching, and can express their needs through keyword queries. Examples of lookup activities include fact finding, known item search and question answering. Such types of lookup searches are supported reasonably well by current Web search engines.

The other two search activities, learn and investigation, are labeled by Marcionini as exploratory search (alternately, Guha et al. (2003) labels them research searches). An exploratory search task involves learning and gathering enough information about a topic so that users can accumulate new knowledge about the topic through the information seeking process. There are two major difficulties associated with exploratory search tasks: 1) the topic domain is usually unknown to users, thus it is difficult for users to describe what they are looking for through a query; and 2) the answer space is usually multi-dimensional and users' information needs cannot be satisfied with a single query. Thus, when users start such a search task, their initial query is usually vague and broad. They then refine their queries according to what they learn from search results of previous queries, and collect relevant information on the way. Compared to the lookup search task, exploratory search requires users to perform higher level cognitive processing of each query search iteration: users must spend a significant amount of time browsing search results from each query, reading, interpreting, comparing, evaluating and synthesizing unfamiliar information, revisit what they found, and aggregate what they learn. This culminates in the formulation of a new query to retrieve information that either confirms what they have learnt, or leads to new dimensions of the answer space. This is a berrypicking information seeking process as described by Bates (1989).

2.2 Search Interface

A search interface is a communication channel between a user and an underlying search engine, hence an important part of an information retrieval system. Current search interfaces usually provide a query box and a ranked list of retrieved documents, regardless of a user's search task or search context. Significant effort has been invested in tailoring search and ranking algorithms to particular search tasks in the last decade, as evidenced by various TREC (Text REtrieval Conference) tracks such as home page finding, new topic detection and question answering (Voorhees and Harman 2005). However, search interfaces have remained largely unchanged over the same period. A query box and a ranked list of retrieved documents may be good enough for lookup search tasks, where users know what they are looking for and can recognisee relevant documents from a few top ranked pages, but they are inadequate to support complex information seeking tasks in which users are required not only to recognisee relevant documents, but also to make sense and use of these documents.

An ideal interface should direct and focus users' attention to their information search tasks and allow them to interact with information effortlessly. Designing a search interface to support a user's information task is highly overlapped with design goals of human computer interaction (HCI). In fact, we can regard the design of a search interface as an example of HCI or interface design. Norman (1988) emphasized that when we design HCI, we ought to be asking what tasks people need to accomplish and what tools and technologies are most appropriate for those tasks. Task-based user interface design has been a key methodology of HCI design, where task analysis serves as the primary source for determining which tasks and sub-tasks can be well supported by a system, and how an interface should be structured (Wilson and Johnson 1996).

There have been successful examples of building task-based search interfaces that are specific to certain domains and users. Wu et al. (2004) demonstrated that when users were provided with an interface that explicitly supported their topic distillation task, their search performance significantly improved over the use of a generic interface that showed search results in a ranked list. Based on the understanding of users' needs and tasks for a personal information management task, the Phlat system (Cutrell et al. 2006) provides an intuitive interface that integrates search and browsing through a variety of associative and contextual cues. Apart from various querying features, Phlat allows users to manipulate search results according to features or tags that can be associated with documents. The Flamenco search interface (Yee et al. 2003) is an example of emerging faceted search systems, which uses hierarchical faceted metadata in a manner that allows users to both refine and expand the current query while maintaining a consistent representation of search results in the document collection's structure. The interface also provides guided or structured browsing. While such an interface is effective for searching and browsing a well-structured collection with rich metadata, it would be very challenging to apply such an interface to the web search context, where web pages are unstructured and heterogeneous (Teevan et al. 2008).

2.3 Evaluation

Evaluation of retrieval systems has largely followed the Cranfield methodology (Cleverdon 1967, Sparck-Jones 1981), where a test collection includes a collection of documents; a set of queries or topics; and rel-

evance judgements (usually binary) about each document with respect to each query. A system is then fed the set of queries and scored according to its ability to retrieve relevant documents. This methodology is still widely used today in the TREC (Text REtrieval Conference) evaluation framework (Voorhees and Harman 2005) with increased size of test collections and more realistic search tasks. The advantages of this methodology include the ability to isolate and compare individual components or algorithms embedded within a retrieval system by manipulating algorithmic parameters. Moreover, the experiments are (usually) repeatable by others, and are much less resource intensive than live user studies of information retrieval systems. This system driven evaluation methodology have played an important role in the advancement of search engines.

However, recent papers question whether many of the reported advancements in search engine technology measured with such a methodology are real, or an artefact of the evaluation method (Armstrong et al. 2009*a*,*b*). Furthermore, the existence of an information retrieval system is to assist people to find relevant information that allows informed decision making, or to bridge their knowledge gap (Belkin 1980). Ingwersen and Jarvelin (2004) argue that evaluation of an information retrieval system should be enlarged to include users and their problem context. Thus an information retrieval system should also be evaluated with its potential users and with those users interactively seeking during the retrieval processes.

In 1995 (TREC 4), TREC initiated an interactive track. The goal of this track was two fold: to investigate searching as an *interactive* task by examining the *process* as well as the outcome; and to develop better methodologies for the evaluation of interactive information retrieval systems (IIR) (Over 2001). The evaluation method as formalized by this track has become a defacto standard followed by many researchers to evaluate IIR systems in a laboratory controlled environment (Dumais and Belkin 2005).

In the TREC IIR evaluation framework, subjects (recruited users) are given a number of topics and asked to perform searches on given topics. During their search process, subjects would make their own judgments on what information is relevant and what is not. Subjects' interaction with systems under evaluation are logged, and their experiences of the systems are collected through questionnaires or think-aloud methods. The evaluation criteria are a mix of IR criteria (such as user performance) and HCI criteria (such as a system's accessibility and usability, and user satisfaction).

The TREC IIR evaluation framework has its limitations. Influenced by the system-driven evaluation approach, TREC IIR evaluation focused on the comparison of users' performance with different systems. The systems usually had part of their components specifically designed to support the users' search task. As a result, if a user's performance is poor, it is hard to identify the cause: is it because of a poor design of the search algorithms or a poor design of the search interfaces? In the TREC IIR evaluation, search topics were described only by pre-constructed information requests. Borlund (2000) argued that an essential component of an IIR evaluation method should be a simulated work task situation. As such, we ground our topics in tasks, as explained in Section 4.1. Some researchers also suggested longitudinal studies, especially for those IIR systems that are designed to support more complicated search tasks than simple fact finding (Kelly et al. 2009).

3 Utilizing Query Semantics to Support Exploratory Search

As discussed in Section 2.1, when a user conducts an exploratory search, the problem domain investigated is usually new to the user. Users devote much of their search time examining and comparing search results; every piece of new information may lead them to pursue new directions and formulate new queries (Bates 1989). Thus to support exploratory task effectively, we need to develop a more intuitive interface that could: reveal dynamic, multiple relationships among retrieved documents; provide a better explanation of the relationship between retrieved documents and the query than query word highlighting; proactively collect document attributes (such as document type and creation date) and show them in an appropriate context; as well as to provide better feedback or query recommendations to users.

In this section, we describe two attempts to improve the exploratory search process. Both approaches use an interpretation of the semantics of an initial query as a resource, but one enriches search result presentation, while the other diversifies the coverage of the search result list. We first discuss how the semantics of a query are derived, and then describe our two approaches.

3.1 Semantic Relations

Semantics is the study of meaning, and the meaning is usually represented through a relationship between two or more word entries or concepts. There exist many types of semantic relations. For example, WordNet (Miller 1990) includes relations such as synonymy, antonymy, hypernymy and hyponymy between words. The KL-One knowledge representation System captures the subsumption or super-concept relation between concepts (Brachman and Schmolze 1985).

Semantic relations can be formally defined and captured in manually (or semi-manually) constructed systems such as WordNet and KL-One. It can also be extracted automatically from document corpora (Anyanwu et al. 2005). Manually constructed relations are generally more accurate than those that are automatically extracted, however they are more costly to build, and difficult to extend when new concepts arise.

In this study, we try to support the learning and knowledge acquisition task by utilizing concepts that are semantically related to a query topic. For example, the concept "acid rain" in a high school science subject, would not only involve a definition, but also the following related concepts: "causes of acid rain"; "the relationship between acid rain and global warming"; and "the relationship between acid rain and pollution". We refer to these related concepts that are essential for understanding a target concept as *aspects* of the target concept.

Now the question is, given a topic as expressed by a query, how is the set of aspects derived? Query logs from Web search engines provide a rich information source to (partially) solve this problem in the context of Web search. Studies on query logs have shown that if a query was issued by many users, it is likely that the information need of those users differ (Teevan et al. 2005, Wu et al. 2008), and that successive reformulated queries by these users usually represent different aspects of the initial query (Wu et al. 2008). The larger the number of users who issue a particular query, the richer and wider the coverage of aspects derived from query reformulations.

There have been a number of studies attempting to extract semantic relations from query logs (Parikh

and Sundaresan 2008, Radlinski and Dumais 2006). In fact, major search engines (such as Google, Yahoo! or LiveSearch/Bing) typically extract semantically related queries from query chains and show them to users as "related search". The main purpose of the "related search" is to assist users in clarifying their information needs should their search results be unsatisfactory. In this paper, we investigate how a set of aspects can be effectively used to improve search result organization and presentation.

3.2 Tabular Aspect Interface

A problem with the related search functionality provided by major search engines is that it only provides an *overview* of possible semantic dimensions of a search topic, but lacks a *preview* of each aspectual dimension. Such an overview typically does not provide the user with enough information to assist them to decide whether to explore a particular aspect further. As defined by Greene et al. (2000), an overview is constructed from, and presents, a collection of objects of interest, while a preview is extracted from, and acts as a surrogate for, a single object of interest. An appropriate representation and display of overview and preview can facilitate rapid elimination of aspects that are not of interest to a user. For example, the query biased summaries typically presented as part of a a ranked list are previews of a webpage and influences a user's decision whether to view the webpage page or not (Tombros and Sanderson 1998, Wu et al. 2001, Turpin et al. 2009).

The new search interfaces we explore present a ranked list of previews for each aspect of the initial query in a two dimensional table. Figure 2 shows an example for the "acid rain" query. Each row of the table is an aspect, and previews of ranked documents for each aspect appear across the columns of the row. The preview in each cell is simply the query biased summary of the document generated by Microsoft LiveSearch using each aspect as a query. We expected that, by using this tabular aspect interface, a user can ascertain possible coverage of the topic vertically, while exploring individual aspects horizontally. Users can focus their attention on examining, understanding and relating the available information, instead of expending cognitive effort formulating search strategies.

3.3 Diversified Search Results

Instead of explicitly presenting semantically related aspects of a query topic in the tabular aspect interface, an alternate approach is to present these aspects implicitly through a diversified search result list. As most users are comfortable interacting with a ranked list interface (such as those used by the major Web search engine companies), we assume many users may prefer the simplicity of a list based approach.

Diversifying a search result list has been advocated by IR researchers for over a decade. A diversified list may satisfy users with ambiguous queries, or broad queries with multiple aspects (Carbonell and Goldstein 1998, Paramita et al. 2009). To support an exploratory search task, a diversified list should cover as many different aspects of the query topic as possible. Here we take a simple strategy which involves two steps. Firstly, given a query, we collect its aspects through the "related search" suggestions of a major search engine and retrieve the top 20 webpages from the original query and each aspect query. To ensure accuracy and prevent topic drift, each query aspect must contain the original query words. Second, we rank the aggregated search results according to the number of aspects contained in a document, taking into account the number aspects that have not been covered by higher ranked documents. In other words, the top document has the most aspects, the second has the most aspects after the top document's aspects have been removed from consideration, and so on.

Figure 3 shows the presentation of a diversified list that is ranked based on a document's coverage and novelty. The top ranked pages usually cover more than one query aspect, thus the page has multiple previews: one for each aspect. There is (typically) not enough room on a single page to show all of the previews for all of the documents, and so each page is still represented by its title, a preview (randomly chosen amongst all aspects) and its URL. If the page has more than one aspect, a label "more summaries" will be placed next to the title, and if the user hovers her mouse over the label, a text box pops up and displays other previews for the same page.

It has been demonstrated that when users visit a webpage from a search results list, they may browse within the website, particularly if they feel the site can answer their information need (Pirolli and Card 1995). Hence we could consider presenting a list of websites, rather than documents, ranked by the site's coverage and novelty. There are two possible advantages of treating a website as a unit for ranking. Firstly, we could find an information-rich site that is devoted to the query topic; a list of ranked websites would function like a resource page or topic distillation page (Craswell et al. 2003). Secondly, users could focus their attention on possibly relevant pages from a site instead of spending time browsing the site to locate these pages. They may not even be able to find them if a website is not structured appropriately.

Accordingly, we built an interface that ranks sites, rather than documents, with previews of multiple documents from the same site being presented to users. As the number of previews to list for each site can be quite large, we opted for an in-line presentation of the lists, rather than the hover-based interface used for aspects of a single document. Figure 4 shows an example of the list of ranked sites, and then if a user selects "Click here for more matches from this site", the summaries of pages are shown inline. Figure 5 shows the result if the first site is selected from Figure 4. The expanded area will disappear if the user clicks the label again.

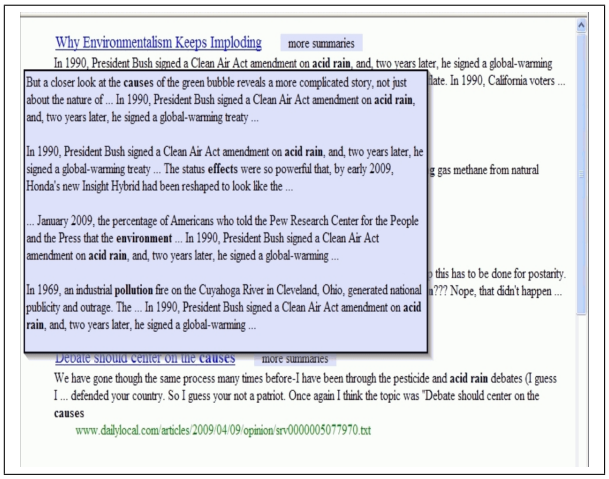

Figure 3: Diversified List with in-line previews interface. Summaries showing the aspects of one page are displayed as the result of a mouse-hover action.

Figure 2: The Tabular Aspect interface.

Figure 4: Diversified List with Site-based ranking interface.

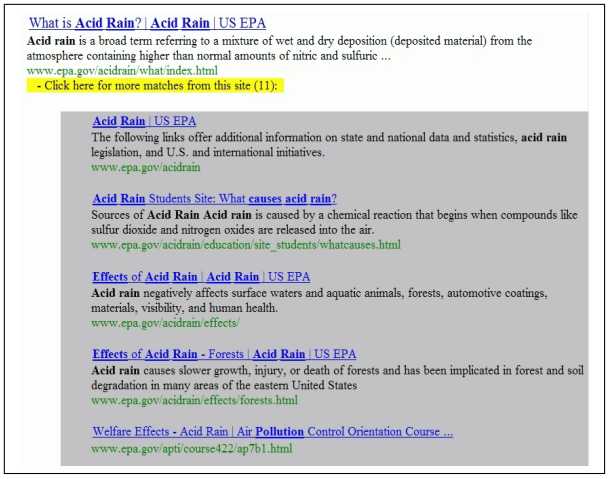

Figure 5: Diversified List with more matched pages from the first displayed site.

4 Evaluation

In the process of researching and designing an interface to support the exploratory search task, we followed design → evaluation → redesign cycle. This section introduces a discounted usability evaluation method (Nielsen 1989) that we adopted to evaluate our initial interface prototypes.

We recruited a small number of subjects and relied our findings on observation rather than on statistical analysis. The feedbacks we gathered from this evaluation will guide our further improvement to the interfaces.

4.1 Experimental Design

We developed two prototype interfaces: the Tabular Aspect interface (Figure 2) and the diversified list

with site-based ranking and in-line previews (Figure 4 and 5). We decided not to include the diversified list with page-based ranking and hovering preview lists, as initial user feedback indicated that while users preferred multiple snippets extracted from a page, they did not like the presentation style. We are exploring other possible presentation styles for lists of this nature.

For the two interfaces, we chose a similar interface from mainstream search engines as a baseline for comparison. For the tabular aspect interface, we chose the "related search" interface (Figure 6) from LiveSearch;[1] for the site-based diversified list, we chose a conventional ranked list as shown in Figure 7.

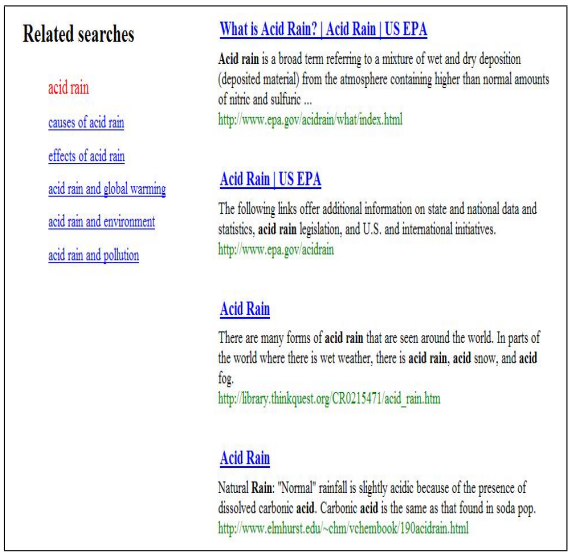

Figure 6: The Related Search interface.

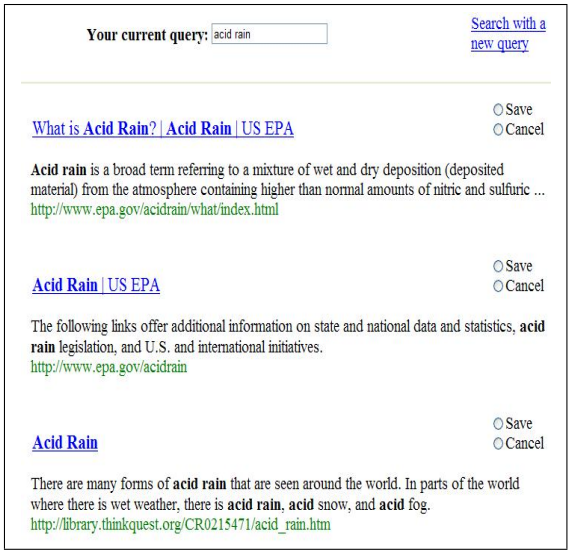

Figure 7: The Conventional List interface.

We hand picked four queries from a Web search log that might suit our subjects. For each topic, we developed a task scenario that simulates a search task in the real world. The four topics (and a training topic) are shown in Table 1.

[1]At the time of this research was conducted, Bing was not launched.

Table 1: The search topics.

Topic 1: Acid Rain
Imagine you are doing a science project on acid rain; you need to find as many facts about acid rain as possible and understand how it impacts our environment, so you can write a thorough and well-researched report.
Topic 2: Scientology
Imagine that you hear this word from all sorts of media now and then and you were also asked by your friends what Scientology is about: is it a science, a religion or a cult? You set out to search for the truth about Scientology so you could understand what it is and therefore help you to form your own opinion.
Topic 3: Nanotechnology
Imagine that you are a science reporter of a leading newspaper, and you are asked by your boss to write a news story on nanotechnology so the public can be informed about this new technology. Think what the public would like to know about a new technology and what information you need to provide in such an essay. And now please go and find the information that you think it would be required.
Topic 4: Impressionism
Imagine that your friend is inviting you to visit an impressionism exhibition this weekend. You would like to know more about impressionism before the visit so you can get the most out of the visit and also impress your friends.
Training topic: Diabetes
One of your close fiends told you that he was recently diagnosed with Diabetes. You might hear about this disease before but you have only limited knowledge. Now you would like to learn more and get more information about Diabetes so that you can communicate better with your friends in the future.

Table 2 shows a block of the experimental design. Each subject searched one topic per interface; the order of topics were fixed, while the four interfaces are rotated in each position. A complete design requires a group of four subjects.

We recruited eight subjects and divided them into two groups so that the block design was repeated twice. The four subjects from the first group were postgraduate students from our School of Computer Science, while the other four subjects were from various backgrounds: two of them were teenagers (both 16 years old) and another two were middle aged professionals.

Figures 2 and 4-7 show the four testing interfaces respectively. As shown in Figure 7, two buttons, "save" and "cancel", are placed next to each retrieved document. Subjects can save a document if they think that this entry is useful, or cancel their previous saving if they change their mind. When subjects started a new topic, they were initially presented with the same set of pre-saved documents. If these initial candidate documents could not satisfy a subject's need, they could formulate their own queries. Their queries were directed to LiveSearch through its API, and search results were displayed as a ranked list, as the related search topics were not included in the API.

During sessions, all significant events such as documents viewed and saved, and queries sent, were automatically logged and time-stamped.

4.2 Experimental Procedure

Subjects undertaking the experiment were instructed as follows. Firstly they read a plain English statement about the experiment and signed a consent form in compliance with our university's Ethics Board. The

Table 2: The experimental structure.

Topic 1	Topic 2	Topic 3	Topic 4
Conventional List	Related Search	Diversified List	Tabular Aspect
Related Search	Diversified List	Tabular Aspect	Conventional List
Diversified List	Tabular Aspect	Conventional List	Related Search
Tabular Aspect	Conventional List	Related Search	Diversified List

search task and the four testing interfaces were then explained and demonstrated to subjects. Subjects had a practice with each interface using the example topic, and were encouraged to ask any related questions. After the subjects acknowledged their understanding of the search task and search interfaces, they started their tasks, following the sequence of assigned topics and interfaces as per the block design.

For each topic, subjects filled in a pre-search questionnaire that gathered their familiarity with the search topic, they then commenced searching to gather information for the assigned task. When they finished interacting with the initial testing interface, but before continuing with their own search, they were asked to fill in an intermediate questionnaire that asked them how much they had learned about the topic and to describe their search experience. The same questions were asked again in a post-search questionnaire to those subjects who continued searching using their own search queries. After the subjects finished all four search topics, they filled in an exit questionnaire that asked them to compare the four interfaces.

Subjects had a maximum of 20 minutes to search on a topic. They could exit earlier if they felt that they found enough information for that topic.

4.3 Experimental Measurement

This experiment had an independent variable (the interface), which has four possible values: Conventional list, Related Search, Diversified List, and Tabular Aspect.

We devised two types of dependent variables to measure against the independent variable: *subjective measure* and *objective measure*.

For the objective measures, we count how many queries a subject sent, how many documents a subject read and saved, and, most importantly, the quality of their saved document set. In order to measure the quality of their saved set, the eight possible sets of saved documents for each topic (one per subject) were assessed for aspect coverage by the final four authors of this paper (who are experienced judges in information retrieval tasks). Note that none of these authors were involved in developing the topics, tasks, nor selecting the documents and lists for the experiments. During the assessment process, each set was rated on a 1-5 Likert scale: with 1 meaning "very narrow", 2 "narrow", 3 "somehow covered", 4 "better covered" and 5 "most covered".

This evaluation of the saved set is relative to the other subjects in the task. That is, if a particular topic is difficult, or the lists provided in the interfaces poor in some wider sense, then most subjects will have poor coverage of aspects in their sets, but some may still get a high score of 5 relative to the other subjects.

Our subjective measures include: subjects' preference of each interface; their confidence of task completion; and the amount of knowledge acquired after they searched. The values of these variables were collected from questionnaires and are now described in detail.

Acquired knowledge: To measure how much knowledge a subject acquired after an interface was used, we designed the following questions for the pre-search, intermediate search and post-search questionnaires. We first asked subjects the following question in the pre-search questionnaire; the subjects were required to choose an indicative scale as answer.

How much do you think you know about the topic?

1. Not at all
2. A little bit
3. Somehow
4. A lot
5. Extremely knowledgeable

After the subjects searched with each interface, a similar question was then asked again in the intermediate questionnaire.

How much do you know about this topic now?

We could use the difference of two scores from above questions to measure how much a subject learnt. However, if a subject does not know a topic very well, they might not be able to truly answer the first question well. As one subject said: "I thought I knew about acid rain a lot, but after I searched and read, I found I only knew a little bit." Hence we added the following question in the intermediate questionnaire

Now you have learnt about this topic, please re-evaluate how much did you know about this topic before your search?

The difference between the value chosen in the intermediate questionnaire and the one from this re-evaluated question was used to measure the amount of acquired knowledge.

Confidence of task completion: This variable was measured by subjects' response to the following question in the post-search questionnaire.

How certain are you that you have got enough information to fulfil your assigned task?

1. Not at all
2. A little bit
3. Somewhat
4. A lot
5. Extremely

Subjective preference: We asked subjects about their preference for the four interfaces in the exit questionnaire. Subjects were asked to say which they preferred out of the Conventional List interface and the Diversified List, and then which they preferred out of the Related Search interface and the Tabular Aspect interface. Subjects were also asked two open questions: 1) What did you like about each interface? and 2) What did you dislike about each interface?

5 Experimental Result

5.1 Objective Measures

Twenty minutes is not a lot time to carry out an exploratory or research type of search task. Most of our

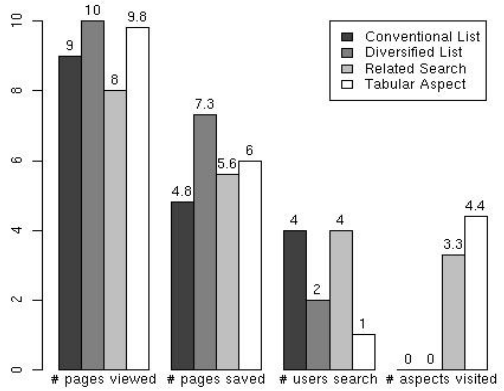

Figure 8: Objective measures (mean of 8 subjects).

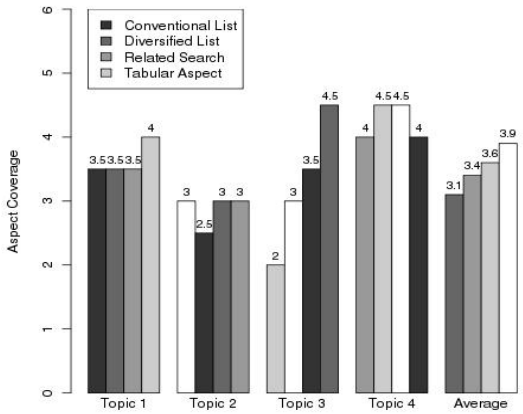

Figure 9: Aspect coverage of saved set. Each topic is the mean over 2 subjects. Average is the mean over all 8 subjects.

subjects spent twenty minutes in interacting with the initial answer lists or interfaces without doing much further search. Figure 8 indicates where the subjects, on average, spent their time. As we expected, the subjects read more pages with the Diversified List and the Tabular Aspect interfaces than the other two interfaces. The subjects who interacted with the Conventional List and the Related Search interfaces tend to issue their own queries (4 users issued queries for both of the former, compared with 2 and 1 for the latter pair, respectively). Between the two interfaces that explicitly showed query aspects, the subjects visited more aspect entries from the Tabular Aspect interface (4.4) than the Related Search interface (3.3).

Figure 9 shows the aspect coverage. The aspect coverage averaged over all four topics (final bars) is in an increasing order from Conventional List, Diversified List, Related Search, to Tabular Aspect interfaces. Topic by topic, the subjects with the aspectual interfaces (either explicit or implicit) performed better than the two list interfaces (except for Tabular on Topic 4).

Generally, the more pages a list contains, the more aspects it covers; except for those lists saved from the Diversified List interface. Although the subjects saved the most pages from the Diversified List interface, these pages did not have the most coverage. A possible reason could be that as most saved pages are from the same site, they tend to cover similar sets of aspects.

5.2 Subjective Measure

Acquired Knowledge: Figure 10 shows the different scores of the four interfaces topic by topic: subjects gave the highest score to the Conventional List for Topic 1, the Diversified List for Topic 3, and the Tabular Aspect for Topic 4. The Tabular Aspect and the Related Search interfaces are equivalent for Topic 2. Overall, as indicated in the group of the rightmost bars in Figure 10, subjects felt that they acquired more knowledge from the Tabular Aspect interface, followed by the Diversified, Related Search and Conventional List interfaces, although the difference was not statistically significant.

Confidence of task completeness: Subjects' confidence of task completeness also varies from topic to topic, as shown in Figure 11. On average, subjects were the most confident of getting enough information using the Tabular Aspect interface, followed by

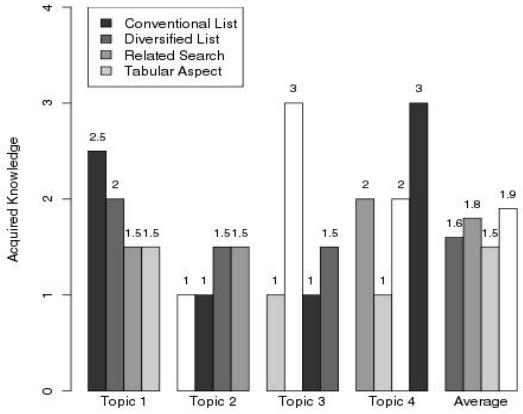

Figure 10: Acquired Knowledge. Each topic is the mean over 2 subjects. Average is the mean over all 8 subjects.

the Related Search and the Diversified List interface. The Conventional List interface inspired the least confidence in completeness of the coverage of information gained.

It is interesting to note that although the Spearman Rank Correlation is low (and not significant) between the acquired knowledge as self-assessed by the subjects and the aspect coverage as assessed by some authors ($\rho = 0.29$), the self-confidence of task completeness is significantly correlated with the aspect coverage ($\rho = 0.60, p < 0.01$) and the acquired knowledge ($\rho = 0.54, p < 0.03$).

Subjective Preference: Subjects were asked to choose a preferred interface between the two list interfaces and the two aspect interfaces. Among eight subjects, five of them preferred the Diversified List interface over the Conventional List interface, and the same number of subjects preferred the Tabular Aspect interface over the Related Search interface. Four of the subjects are postgraduate students with a Computer Science and Information Technology background. It is interesting to note that two of them pre-

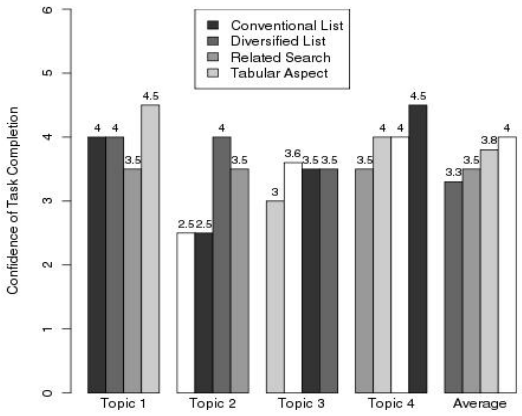

Figure 11: Confidence of Task Completion. Each topic is the mean over 2 subjects. Average is the mean over all 8 subjects.

ferred the Conventional List over the Diversified list and three of them preferred the Related Search interface over the Tabular Aspect interface. Seven subjects replied to the question that asked them to name a preferred interface: five subjects named the Tabular Aspect interface (only one of them is the postgraduate), two named the Related Search interface, and one the Conventional List interface.

The subjects' choice of preference indicates a trend that people from different backgrounds may prefer different interfaces: our postgraduate students preferred the Conventional List and the Related Search interfaces, they commented that there was too much text in the Tabular Aspect interface; while subjects from non-computing backgrounds showed the opposite preference. We conjecture that the postgraduates might have more knowledge on how a search engine works, and their search and browse behaviors are more analytical than the other group of subjects. The other four subjects who preferred the Tabular Aspect interface commented that all information is available from the interface, the two teenagers explicitly said that they could do less (search and click) with the Tabular Aspect interface. Their interaction behavior with the list interfaces provide some explanation: when they interacted with the list interfaces, they first clicked and opened quite a few pages in separate windows or tabs, which they then read one by one; unlike the postgraduates who tend to read the lists sequentially and followed this cycle: read page summaries from the list interfaces, click on a page and then read the page.

Our future work will include redesigning and refining the prototype interfaces and testing them with more users and users with wider backgrounds; and ultimately, when we get a reasonably accepted interface through laboratory controlled evaluation, we would like to conduct a longitudinal user study to test users' true acceptance of the interface.

6 Discussion and Conclusion

We explored different interfaces that support learning and investigative search tasks. We conducted a user study to verify whether these interfaces could achieve the desired goals of improving the amount of information that users could discover about a topic in 20 minutes. The feedback from this pilot evaluation will guide our further improvement of the interfaces.

Our results show that, by either self-assessment or objective assessment, subjects acquired more knowledge and are more confident with the interface that explicitly supports their learning and information gathering tasks. There is a significant correlation between subjects' confidence of task completeness and the acquired knowledge. However, subjects with different backgrounds preferred different answer presentation styles and showed different interaction behaviors. These findings indicate that we should combine the advantages of each interface and/or provide a range of presentation choices so that users could have some control on what should be included in the answer list and how the answer list should be presented.

While the number of users and topics is small, we feel that the observations made in the study are informative. In particular, the preference of postgraduate computing students for the "traditional" Conventional List interface is notable. Many user studies in the field of Information Retrieval make use of computing graduate students as subjects, and often the conjecture is made that the results would not carry over to other user populations. This study in part confirms that conjecture, emphasizing that extreme care must be taken when interpreting results of users studies on specific cohorts of participants.

Acknowledgements

We thank Microsoft Research for providing a search API to their LiveSearch, and for research funding for this study.

References

Anyanwu, K., Maduko, A. & Sheth, A. (2005), Semrank: ranking comples relationship search results on the semantic web, *in* 'Proceeding of the 14th International Conference on on World Wide Web', Chiba, Japan, pp. 117–127.

Armstrong, T. G., Moffat, A., Webber, W. & Zobel, J. (2009*a*), Has adhoc retrieval improved since 1994?, *in* J. Allan, J. Aslam, M. Sanderson, C. Zhai & J. Zobel, eds, 'Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Boston, USA, pp. 25–26.

Armstrong, T. G., Moffat, A., Webber, W. & Zobel, J. (2009*b*), Improvements that don't add up: ad-hoc retrieval results since 1998, *in* 'Proc. 18th ACM International Conference on Information and Knowledge Management', Hong Kong, China. To appear.

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

Belkin, N. J. (1980), 'Anomalous state of knowledge as a basis for information retrieval', *Canadian Journal of Information Science* **5**, 133–143.

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

Brachman, R. & Schmolze, J. (1985), 'An overview of the KL-ONE knowledge representation system', *Cognitive Science* **9**(2), 171–216.

Carbonell, J. & Goldstein, J. (1998), The use of MMR diversity-based reranking for reordering documents and producing summaries, *in* 'Proceedings of the 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval', pp. 335–336.

Cleverdon, C. (1967), 'The Cranfield tests on index language devices', *Aslib Proceedings* **19**, 173–192. Reprinted in K. Sparck Jones and P. Willett, editors. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., 1997.

Craswell, N., Hawking, D., Wilkinson, R. & Wu, M. (2003), Overview of the TREC-2003 web track, *in* 'Proceedings of TREC-2003', Available online at trec.nist.gov/pubs/.

Cutrell, E., Robbins, D. C., Dumais, S. T. & Sarin, R. (2006), Fast, flexible filtering with phlat - personal search and organization made easy, *in* 'Proceedings of ACM SIGCHI Conference', pp. 261–270.

Dumais, S. J. & Belkin, N. J. (2005), The TREC interactive tracks: Putting the user into search, *in* E. Voorhees & D. Harman, eds, 'TREC: experiment and evaluation in information retrieval', MIT Press, pp. 123–153.

Greene, S., Marchionini, G., Plaisant, C. & Shneiderman, B. (2000), 'Previews and overviews in digital libraries: designing surrogates to support visual information seeking', *Journal of the American Society for Information Science* **51**(4), 380–393.

Guha, R., McCool, R. & Miller, E. (2003), Semantic search, *in* 'Proceeding of the 12th International Conference on on World Wide Web', Budapest, Hungary, pp. 700–709.

Ingwersen, P. & Jarvelin, K. (2004), Information retrieval in contexts, *in* 'Information Retrieval in Context (IRiX) ACM-SIGIR Workshop', pp. 6–9.

Kelly, D., Dumais, S. & Pedersen, J. O. (2009), 'Evaluation challenges and directions for information-seeking support systems', *IEEE Computer* pp. 60–66.

Marchionini, G. (2006), 'Toward human-computer inforamtion retrieval', *Bulletin of the American Society for Information Science and Technology* .

Miller, G. A. (1990), 'Wordnet: An online lexical database', *International Journal of Lexicography* **3**(4), 235–312.

Nielsen, J. (1989), Usability engineering at a discount, *in* 'Proceeding of the 3rd International Conference on Human-Computer Interaction', NY, USA, pp. 394–401.

Norman, D. A. (1988), *The Psychology of Everyday Things*, Basic Books.

Over, P. (2001), 'The TREC interactive track: an annotated bibliography', *Information Processing and Management* **37**(3), 369–381.

Paramita, M., Sanderson, M. & Clough, P. (2009), Developing a test collection to support diversity analysis, *in* 'Proceedings of Redundancy, Diversity, and Interdependent Document Relevance workshop held at ACM SIGIR'.

Parikh, N. & Sundaresan, N. (2008), Inferring semantic query relations from collective user behavior, *in* 'Proceedings of the 17st ACM Conference on Information and Knowledge Mining', pp. 349–358.

Pirolli, P. & Card, S. K. (1995), 'Information foraging', *Psychological Review* **106**, 643–675.

Radlinski, F. & Dumais, S. (2006), Improving personalized web search using result diversification, *in* 'Proceedings of the 29st International ACM-SIGIR Conference on Research and Development in Information Retrieval', pp. 691–692.

Sparck-Jones, K. (1981), The Cranfield tests, *in* K. Sparck-Jones, ed., 'Information Retrieval Experiment', Butterworths, pp. 256–284.

Teevan, J., Dumais, S. & Gutt, Z. (2008), Challenges for supporting faceted search in large, heterogeneous corpora like the Web, *in* 'Proceedings of HCIR'.

Teevan, J., Dumais, S. T. & Horvitz, E. (2005), Beyond the commons: Investigating the value of personalizing web search, *in* 'Proceedings PLA 2005: Workshop on New Technologies for Personalized Information Access', pp. 84–92.

Tombros, A. & Sanderson, M. (1998), Advantages of query biased summaries in information retrieval, *in* 'Proceedings of the 21th International ACM-SIGIR Conference on Research and Development in Information Retrieval', Melbourne, Australia, pp. 2–10.

Turpin, A., Scholer, F., Jarvelin, K., Wu, M. & Culpepper, S. (2009), Including summaries in system evaluation, *in* 'Proceedings of the 32th International ACM-SIGIR Conference on Research and Development in Information Retrieval', Boston, USA, pp. 508–515.

Voorhees, E. M. & Harman, D. K. (2005), *TREC: experiment and evaluation in information retrieval*, MIT Press.

Wilson, S. & Johnson, P. (1996), Bridging the generation gap: From work tasks to user interface design, *in* J. Vanderdonckt, ed., 'Computer-Aided Design of User Interfaces', Presses Universitaires de Manur, pp. 77–94.

Wu, M., Fuller, M. & Wilkinson, R. (2001), Searcher performance in question answering, *in* 'Proceedings of the 24th International ACM-SIGIR Conference on Research and Development in Information Retrieval', New Orleans, LA, pp. 375–381.

Wu, M., Muresan, G. & et al. (2004), Human versus machine in the topic distillation task, *in* 'Proceedings of the 27th International ACM-SIGIR Conference on Research and Development in Information Retrieval', pp. 385–392.

Wu, M., Turpin, A. & Zobel, J. (2008), An investigation on a community's web search variability, *in* G. Dobbie & B. Mans, eds, 'Proceedings of the Thirty-First Australisian Computer Science Conference', Wollongong, Australia, pp. 117–125.

Yee, P., Swearingen, K., Li, K. & Hearst, M. (2003), Faceted metadata for image search and browsing, *in* 'Proceedings of ACM SIGCHI Conference', pp. 401–408.