Findex: Search Result Categories Help Users when Document Ranking Fails

Mika Käki
Department of Computer Sciences
FIN-33014 University of Tampere, Finland
mika.kaki@cs.uta.fi
+358-3-215 6181

ABSTRACT

Long web search result lists can be hard to browse. We demonstrated experimentally, in a previous study, the usefulness of a categorization algorithm and filtering interface. However, the nature of interaction in real settings is not known from an experiment in laboratory settings. To address this problem, we provided our categorizing web search user interface to 16 users for a two month period. The interactions with the system were logged and the users' opinions were elicited with two questionnaires. The results show that categories are successfully used as part of users' search habits. They are helpful when the result ranking of the search engine fails. In those cases, the users are able to access results that locate far in the rank order list with the categories. Users can also formulate simpler queries and find needed results with the help of the categories. In addition, the categories are beneficial when more than one result is needed like in an exploratory or undirected search task.

ACM Classification: H5.2 [Information Interfaces and Presentation]: User Interfaces; H3.3 [Information Storage and Retrieval]: Information Search and Retrieval — Selection process, Clustering

Keywords: Web search; search user interfaces; categorization; clustering; information access

INTRODUCTION

Web search engines are essential tools for finding information from the World Wide Web (web). The vast amount of information and the variety of the topics pose a major challenge for the web search engines. Displaying results in an ordered list according to a document ranking scheme is the most popular way of presenting the search results, but the solution is not without shortcomings. It is known that the typical query contains just one or two search terms [14] and people typically view only the first result page (10 results) [13]. Knowing this, it is understandable that document ranking can fail in addressing the user's information need. The problem is widely known and result categorization [5, 4, 8, 9, 24, 25] and query refining aids [1, 3], among other things, are proposed in the literature to overcome the difficulties.

We addressed this problem in an earlier study by developing an online search result categorization algorithm and a filtering user interface to improve the result evaluation phase of the search process. We tested the solution in an experiment where it was compared to the rank order list with ten results per page (*de facto* standard). The results showed [17] that the use of categories improves the search speed and increases the accuracy of the selected results. The first evaluation, however, was a strictly controlled experiment and was thus unable to provide insight about the use of the categories in real settings.

Thus the current research question is: *How is the category user interface used in real settings and is its use beneficial?* We addressed this question by providing access to a new web-based search user interface, Findex, for 16 users for two months. The use of the system was not controlled in any way and the participants were encouraged to use the system according to their own habits. The participants' interaction with the system was recorded in log files and opinions were collected using two online questionnaires, first in the early stages and then after the two month usage period.

The results of the study show that users do benefit from the categories, but the factors affecting the category use are more diverse than the earlier study suggested. The participants use the categories regularly, for every fourth search, and adopt them as a part of their search habits. When categories are used, the results located low in the listing are found and selected.

We will first discuss the related work and then move to a description of the tested system. Then the method of the evaluation is described followed by the presentation of the results. Finally the results are discussed.

RELATED WORK

There are three relevant areas of research for this study. First, the research on search user interfaces with categorization features is a long lived and recently quite popular area that sets the background for our user interface solution. Second, the development of query refinement user

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.

interfaces has recently gained more momentum as a few large web search engines have employed such features. This approach is closely related to our categorization technique. Third, research augmenting the general understanding of web searching is also relevant. Such studies provide us with valuable insights about why and how users are searching the web. The findings will further explain our results.

Result Categorization

Two main approaches are traditionally used in the categorization algorithms. *Clustering* algorithms form categories by grouping similar documents together on a given set. *Classification* refers to a process where documents are automatically assigned to a predefined set of categories.

Dumais et al. [10] were among the first in the HCI community to suggest categorization (more specifically clustering) techniques for improving the access to textual information. Later, Scatter/Gather [5] was one of the first systems where the clustering approach was tested. On one hand, the test revealed problems in the naming of the automatically computed clusters as Scatter/Gather seemed both slower and less precise in information search tasks [18]. On the other hand, Scatter/Gather demonstrated the great potential of the categorization approach. A later study showed that users were able to identify the most relevant categories in information gathering tasks [11].

Our previous study [17] used a clustering technique for enhancing the access to web search results. The results of the laboratory experiment indicated that accessing the results was both 40% faster and more accurate compared to the ranked list approach.

Zamir and Etzioni have also developed a clustering technique specifically for the web environment and a web search user interface called Grouper [25, 24]. Most notable differences to our work are the clustering algorithm and the user interface. In addition, we have conducted a laboratory experiment with our system which allows us to make comparisons to the current longitudinal study. The evaluation in the work of Zamir and Etzioni [24] was based on logs and their experiences and results are employed in this study.

Zeng et al. [26] followed largely the same approach in result clustering as Zamir and Etzioni but employed learning techniques in the process. Their user interface proposition is close to ours.

The SWISH prototype by Dumais and her colleagues [4, 8, 9] is also a categorizing web search user interface. Empirical evaluations of the system have confirmed that its use is efficient in certain tasks. In contrast to Grouper, SWISH uses a classification method for the categorization. DynaCat [19] is another prototype system that uses the classification approach. DynaCat applies classification to enhance access to medical information. A study showed improved performance over the conventional user interface.

In this context, the third technical approach in addition to clustering (as in Scatter/Gather) and classification (as in SWISH) is term extraction. One such a system, Kea [15], is part of the New Zealand Digital Library project efforts. Kea and the automatically extracted keywords were used to find related documents as well as to categorize search results on small devices [16]. In the web environment, a similar technique was utilized by Wu et al. [23]. They used the technique for displaying hierarchical overviews of web search results. Drori, on the other hand, used both keywords and categories in his experiments on web search engine user interfaces [7].

Additionally, categorization techniques have recently been employed in commercial search engines, such as Vivísimo [32], WiseNut [33], and iBoogie [30]. However, we are not aware of any studies concerning their actual usefulness.

Query Refinement

Query formulation is known to be difficult for typical web users and even experts have problems in it. To help users in query formulation, one can provide automatic suggestions for query refinements. A few major search engines (e.g. AltaVista [28] and Teoma [31]) have adopted this solution. This approach is closely related to the result categorization as the selection of a category can be seen as one kind of query refinement. The difference is that a query refinement generates a new result set whereas result categorization alters the presentation of the existing one.

Anick conducted a log study [1] to determine whether query refinement suggestions are actually used and how successfully. He concluded that query refinements are still mostly done manually, but when the suggested refinements are used, they are effective. Dennis, McArthur and Bruza have studied a prototype that provides similar functionality. Their studies have concentrated on comparing their approach to others [3] and measuring users' cognitive load while using the system [6]. They found support for their hypothesis that suggested refinements decrease users' cognitive load.

Search Behavior

Our general understanding of web searching has increased lately, too. For example, studies have shown that search topics have changed over the years, but the query formulation skills and habits have remained the same [21]. People still use few terms and few operators in their queries and therefore need help in result evaluation. In addition, our understanding of search types has sharpened as taxonomies of web searches are proposed [2] and further specified [20]. In particular, so called navigational and undirected informational search goals are interesting here. In navigational search the goal is to reach a particular site [2], such as a company web site. The link between the query and the page is typically direct and the rank ordering can place the correct page on the top of the list. On the

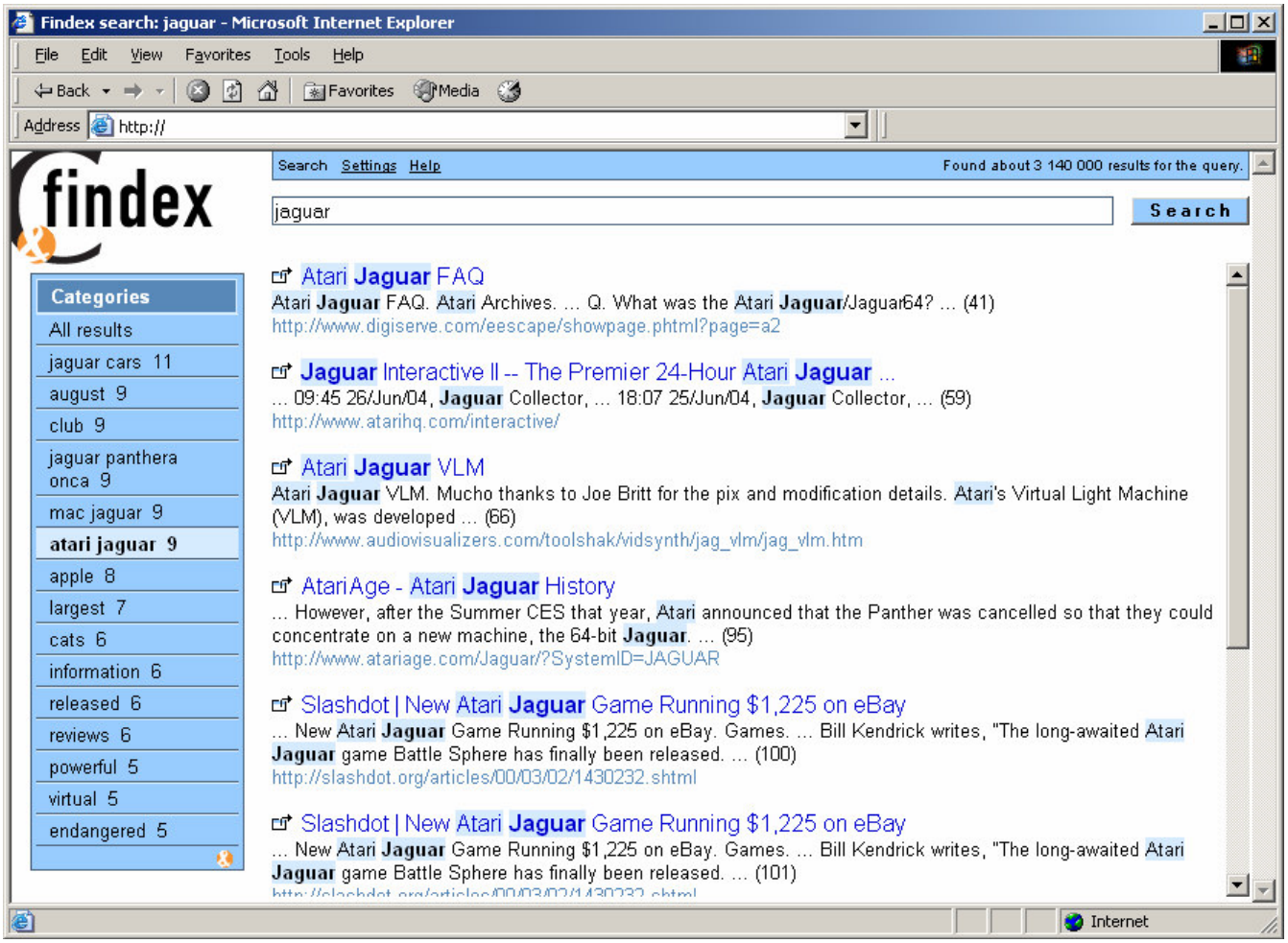

Figure 1. Filtering Findex search user interface. Categories on the left, filtered results on the right.

other hand, in undirected informational searches multiple results and broader understanding about the topic is desired. Categorizing approach can help in such a task.

Furthermore, an orientating search strategy [22] is seen to lessen the cognitive burden of searchers. Categories are a way to help users better orientate themselves to the results. Another study of web searches also revealed that novice users preferred making a new query over choosing a result document [12] in a hard situation. Categories could help those users in result evaluation as interesting documents can be found further down in the result listing.

SYSTEM DESCRIPTION

Our search interface, Findex (Figure 1), is based on filtering the result list using automatically computed categories. The categorization algorithm, in turn, is based on the word frequencies in result summaries. Categories are computed automatically for each result set and they are displayed to the users as a list next to the actual results. The solution aims to improve the search result categorization by avoiding complex functionality that prevents users from understanding the basic principles of the system. The actual searches are done through Google Web API [29].

In addition to more understandable categorization, the other major difference to previous work is the filtering user interface. It allows the users to take advantage of the ranked result list when it works and provides an extra means for coping when the ranking fails. In practice, the users first see the results as the underlying search engine returns them. The categories are available on the side and users are not required to use them. Selecting a category filters the result list to show those results belonging to the selected category.

Result Categorization

Users need to understand the functionality of the software in order to feel in control of it and simplicity is a key to this. To achieve this, our categorization is based on word frequencies among the result summaries. We select the *n* most common words and phrases and use them as the categories. Slightly modified, publicly available stop word lists are used to prevent articles, pronouns, and other non meaningful words from getting selected (modification means adding common words in the web like 'homepage'). The resulting category contains all the results where the word or phrase appears. This technique is expected to be

Phase 1. Original Result Listing

1	Atari Jaguar FAQ Atari Jaguar FAQ. Atari Archives. ... Q. What was the Atari Jaguar/Jaguar64? ... http://www.digiserve.com/eescape/showpage.phtml?page=a2
2	Atari Jaguar VLM Atari Jaguar VLM. Mucho thanks to Joe Britt for the pix and modification details. Atari's Virtual Light Machine (VLM), was developed ... http://www.audiovisualizers.com/toolshak/vidsynth/jag_vlm/jag_vlm.htm
3	Jaguar Cars Road test with BMW, Mercedes and Audi. Beating BMW, Volvo and Audi rivals. ... http://www.jaguar.com/uk/
4	JagWeb - Worlwide directory of Jaguar restoration, trimming ... JagWeb - The original and largest Jaguar directory since 1996! HOT Site: JagADS.com for Jaguar ads!! ... The official website of Jaguar Cars Ltd. ... http://www.jagweb.com/

Phase 2. Extracted Words and Phrases

Words	Result	Phrases	Result
atari	1, 2	atari jaguar	1, 2
faq	1 (unique)	atari jaguar faq	1 (unique)
archives	1 (unique)	atari archives	1 (unique)
cars	3, 4	jaguar cars	3, 4
jagweb	4 (unique)	worldwide directory	4 (unique)
. . .		. . .	

Phase 3. Category Selection from Non-unique Candidates

Candidate	Result	Frequency	Selected
atari jaguar	1, 2	2	yes
atari	1, 2	2	no (atari jaguar)
jaguar cars	3, 4	2	yes
cars	3, 4	2	no (jaguar cars)
. . .			

Figure 2. Partial example of the result categorization scheme for 'jaguar' query.

easily understandable and illustratable in the user interface (e.g. by highlighting the words that explain why the result is in the selected category). The technique is presented schematically in Figure 2.

A problem in string matching is that inflections of the words make them different (e.g. 'car' and 'cars' would be two different words). This is misleading as categories aim to cluster similar results together. We tried a proper stemmer (Snowball stemmer by Martin Porter [27]) for removing the word endings, but it introduced undesired complexity from the user's perspective. Thus we use a non-exact string matching algorithm. In the algorithm, the first 80% of characters (of the shorter word) of the strings are compared and matching strings are not allowed to differ more than 3 characters in length. This approach works well in this context and reduces the language dependency of the algorithm. It is tested to work acceptably in English, Finnish, French and German.

The use of this non-exact string matching algorithm does not remove the language dependency completely. Stop word lists are language sensitive because a stop word in one language may be a relevant word in another. The necessary language detection is based on stop word lists and works by counting the frequency of stop word occurrences in a text (the document summary).

The category building process starts by creating a list of all the words appearing in the results. Each word is associated with information about the result(s) it appears in. In order to get more meaningful categories, we search the most commonly occurring word strings (phrases) in addition to plain words (Phase 2 in Figure 2).

To find the most frequent phrases, we form all possible consecutive word strings within a sentence. When two phrases are tested for equality each word within the phrases is compared by using our non-exact match algorithm.

As we have a list of words and phrases, we remove unique instances, uninteresting, and overlapping categories. These include categories that consist only of words found in the query terms and sub phrases of larger super phrases (e.g. category 'atari' is removed because category 'atari jaguar' exists). Finally, the list of words and phrases is sorted according to the occurrence frequency and the *n* most frequent items are selected as the final categories (Phase 3 in Figure 2).

Final categories can be overlapping and results do not need to belong to any category. The non-exclusive nature of the categories is seen as strength. The meaning of information depends on the context and this feature allows us to present the same information in different contexts. This should make the identification of the interesting results easier. The problem of accessing results that do not belong to any category is not serious because the complete result list is also available in the user interface.

Optimizations and algorithm improvements have considerably improved the computational performance of our categorization system. In its current form, it takes, on average, about 200 milliseconds to compute categories for 150 results. 150 results, in our experience, is a good compromise between the speed and the scope. If more than 150 results are used in the computation, it takes more time, but the categories tend to remain roughly the same. In comparison, Zeng et al. [26] reported that it takes about 4 seconds to categorize the same amount of results with their algorithm.

User Interface

For this study we implemented a new web based user interface for the system (in the laboratory tests we used a standalone application). The implementation uses Java Servlets and it consists of two components. The first component performs the actual searches and computes the categories. The second component is responsible for the user interface. The application is used through any standard web browser.

The user interface follows the basic conventions used in most graphical email clients and Windows Explorer, where a set of collections is on the left and the right side shows the contents of the selected collection, like files in a folder. In Findex, this means that a list of categories is on the left and the right side displays the corresponding results (see Figure 1). The basic design is the same as in the software used in our previous experiment [17] as it was found to be easy to understand and adopt. The category list and the result display are tightly coupled so that changing the category selection changes the contents of the result view.

In contrast to Dumais et al. [9] and Zamir & Etzioni [24] proposals, we have separated the categories completely from the result listing and allow the users to use the plain result list if desired. Categories can be regarded as an added feature to the normal search user interface. We also made the category selection to be as cheap as possible so that the category meaning can be evaluated by looking at the contents of it.

The selected category is highlighted in light blue colour. The same colour is used in the result listing for highlighting the category word(s) (see Figure 1). The aim is to show the working principle to the user in an unobtrusive yet apparent way.

The first item in the category list is a special built-in category for viewing all the results in the result set ('All results') and it is automatically selected after each search. As a result, the system seems like any other search engine. The user gets a ranked list of documents and the categories are offered to the user on the side.

The speed of use is a major factor for web search engines. Category computation unavoidably poses a delay, especially if one server hosts a lot of users. We addressed this problem by a user interface functionality which provides the first ten results to the user as fast as the underlying search engine returns them. Subsequently, the remaining results are retrieved and the categories are computed. When the computation is complete, the rest of the results are appended to the result list and the categories are displayed. Such functionality increases the perceived performance considerably.

The order of the results in the list is determined by the ranking system of the underlying search engine. When 'All results' is selected, the result listing is the same as that returned by the search engine. When another category is selected, the relative order of the results is again determined by the original listing, although they do not typically form a continuous sequence. The result items have an ordinal that shows in parenthesis the position of the result in the rank order of the search engine (see Figure 1).

The status bar, on the top, provides access to user preferences (e.g. user interface language setting) and a short help text. It also displays the total number of hits found for a search. Below the bar is a field for entering the query and the 'Search' button. There is no paging functionality for the results. Instead the first 150 results are displayed on one page.

EVALUATION

We conducted an evaluation to collect usage information about the system in real settings. We wanted to have participants who normally conduct web searches regularly in their work or otherwise. In addition, the participants were not given any particular tasks or instructions on how to use the system. In short, the participants were encouraged to utilize their own search styles with the new system.

Participants

We recruited 16 (8 male, 8 female) participants from national universities. The participants had background in various disciplines such as medicine, geology, astrophysics, and history. None of the participants worked in computer science or software engineering. The ages of the participants ranged from 26 to 59 the mean being 38 years.

All participants had a relatively long history of computer use (*mean* = 18 years, *sd* = 4.2) as well as the web (*mean* = 8 years, *sd* = 1.4) and web search engine (*mean* = 7 years, *sd* = 2.3) use. They reported using computers and the web 'daily' and web search engines 'many times a week'. All participants mentioned Google as one of the search engines they use and 7 of them reported using some other search engines as well. None of the participants had previously used a categorizing search engine (such as Vivísimo, WiseNut, or iBoogie).

Invitation to the study was sent to about 80 people chosen randomly from the faculty listings of university departments. Of those invited, about 20% participated. The percentage is quite low and could affect the results. However, the low percentage is partly explained by a difficult timing of the study (it began in the early summer and some invited people had already started their one month vacation). The effect on the sample is random and therefore acceptable. Another reason was that we accepted only participants who use the web regularly.

Data Collection

The most important data source was application logs like in the study by Zamir and Etzioni [24]. The log gathering software was designed especially for this study and the log files contained information, for example, about the queries, the selected categories, and the selected results. We acknowledge that this data is not exhaustive, because we do not know, for example, the relevance of the selected results for the user. Collecting such information would have, however, affected the searches too much so we decided to exclude the acquisition of it.

The other major source of information was questionnaires. The questionnaires were designed to provide understanding about the motives and rationale of the users and their behavior. Furthermore, we wanted to collect their subjective attitudes about the system. We feel that despite the possible shortcomings the questionnaires are a good complement to the log data. They can give valuable insight to the perceived usefulness of the system without severely affecting search behavior of the users.

Procedure

The participants were invited to the study via email. Prospective participants were required to use web search engines at least a few times a week. Those who were

willing and suitable for the study were sent further instructions that encouraged them to first explore the new system and then use it as they like. The categorization scheme was not introduced at this point.

The first step of participation was to fill in a web questionnaire about background information. After its completion, the users were sent directly to the search engine. The search engine front page contained a short text outlining its functionality. No other instructions were given to the users about the tested system. No training was provided.

After this, the participants were allowed to use the search engine freely. The first subjective questionnaire was sent after a week or two to collect initial impressions and a second one was sent at the end of the observation period. We collected log data for approximately two months of use. Because the study was organized during the summer, the actual observation period was about a month longer (three in total) to compensate for vacations.

RESULTS

We recorded log events for 3099 queries. With these queries the users accessed 3232 result pages and made 1915 category selections. We will present the results by first looking at the queries the users made and then by describing what kind of results were selected. Finally, the category selections will be discussed followed by the presentation of the questionnaire results.

In the analysis we divided the data into queries. A query consists of sending a query, browsing the results and selecting them. The queries were classified according to the category use while browsing the results. If the user selected a category while evaluating the results of a query, we say that the search used categories. The other group includes all the searches where the categories were not used.

Queries

The query clauses in our study are consistent with previous studies. People used only a few search terms (*mean* = 2.07, *sd* = 0.8) as observed previously, for example, by Jansen et al. [14]. The logs also show that the use of query operators is rare except for one participant who used them regularly (typically using the + operator unnecessarily).

The analysis based on category use revealed that simple query properties do not explain the use of categories. The length of the query is the same in both cases as people used 2.04 and 2.10 words for searches with and without category use, respectively. The operator use in queries does not differ either. The participants used operators only rarely.

A more interesting observation concerns the goal of the search. We applied a simple and conservative method for determining the search goal. We assumed that queries that contained only a proper name for which there is a well-defined home page (such as personal or company home page) were navigational. We used a sample of 2704 queries and found 467 navigational queries using this method. The proportion of navigational queries was thus 17% which corresponds to the previous studies where the proportion varied between 11.7% and 24.5% [2, 20].

In the related work section we hypothesized that in navigational searches the conventional result ranking could work well. This seems valid because 78.6% of the navigational searches were completed without category selection and only in 21.4% of the cases the categories were used. The difference is statistically significant as $\chi^2_1 = 8.38$, $p < .005$.

Result Selections

For the characterization of the selected results, we continue to divide the results according to the use of categories during the search. The total number of selected results was

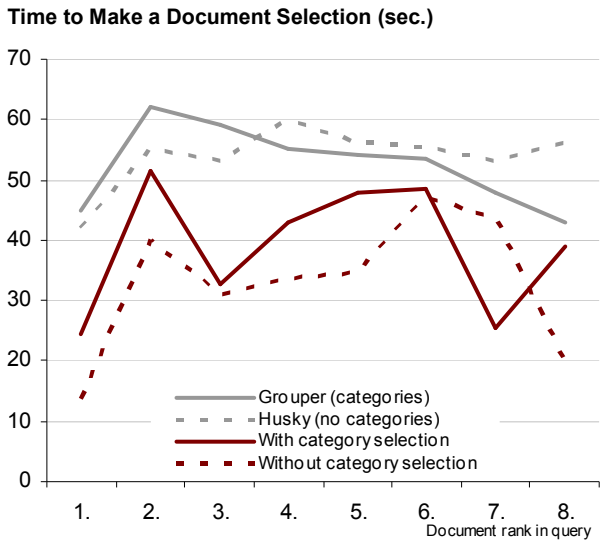

Figure 3. Time used to select documents using different search user interface tactics. Result for Grouper and Husky are from a study by Zamir and Etzioni [24].

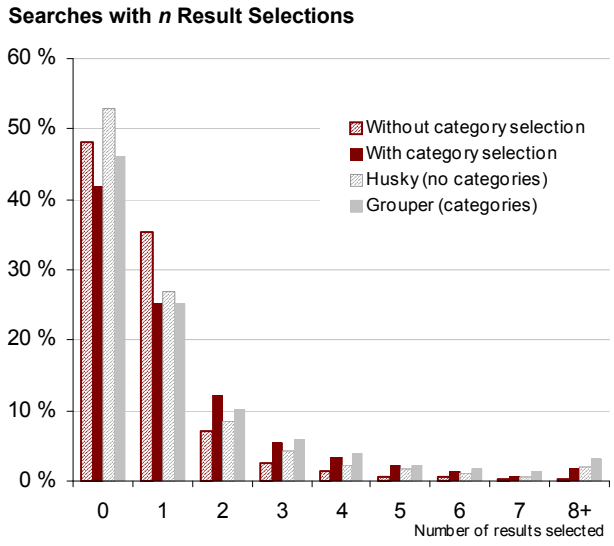

Figure 4. The proportion of the searches where users selected a certain number of documents in different conditions.

3232 and the majority of them, 2380 (74%), were accessed directly from the rank order list. On the other hand, 852 (26%) result selections were made using the categories.

The time to select a result is an important figure for characterizing the success and the manner in which the user interface is used. We adopted the same measuring scheme as Zamir and Etzioni [24] in their log study. They measured the time used to find a document from the results based on the time when the document was accessed. This time includes the time required to scan and evaluate the results, choose among them, as well as network delays and evaluation of the document itself. However, it is a useful measure of search efficiency when a more accurate measure is not available.

Time-based measures show that category use takes initially almost two times longer as the first document is selected in, on average, 24.4 seconds as opposed to 13.7 seconds without category use. The difference is statistically significant as independent samples *t*-test gives $t(30) = 2.5$, $p < .05$. Note also that the time measurement for the first document selection is easier to interpret than the rest because it does not include the time to evaluate the result document itself. In addition to marginal network delays, it simply contains the time to scan the results and to choose among them.

Figure 3 shows our results in a relation to the results by Zamir and Etzioni (gray lines; Grouper uses categories, Husky presents results in rank order list). Note that the measurements after the 5th selection are rather unreliable because the data sets cannot contain many such instances as the average number of selected results was close to one. In the comparison, our system seems a bit faster, but the difference may be due to possible differences in the timing implementation. Anyhow, in our case the category use is constantly slower than using simply the rank order list whereas Zamir and Etzioni observed a crossover point after which the category use was faster.

Interestingly we did not observe a difference in the number of opened results per search in contrast to Zamir's and Etzioni's result. In our study, users selected 1.05 results when they did not use the categories and 1.04 when they did. In contrast, in the Grouper study, users selected on average 1.0 results per search using non-categorizing Husky interface and 40% more with Grouper. Our results are different in this respect.

Although there is no difference in the average number of selected results, interesting differences are found in cases where more than one result is needed. Figure 4 suggests that the use of categories increases the frequency of selecting more than one result for a query. When categories are used, users select in 28.6% of the queries more than one result. The corresponding figure when categories are not used is 13.6%. The difference is statistically significant as $\chi^2_1 = 58.1$, $p < .005$. The same trend of selecting more results with categories is also visible in Grouper usage, but the difference is not as clear.

Another interesting detail can also be seen in Figure 4. A large amount (42% and 48% with and without category use, respectively) of searches yields no result selections. The phenomenon is hard to explain, but our results are congruent with the results of Zamir and Etzioni. The number of searches without any result selections is lower when the categories are used and the difference is statistically significant ($\chi^2_1 = 3.94$, $p < .05$).

An important observation about the selected results concerns the position of the selected result in the original result list. When categories are used, the median position of the selected result in the rank list is 22nd ($sd = 38$). The average position without category use is 2nd ($sd = 8.6$). Even though the variation is high, the difference is significant as independent samples *t*-test gives $t(30) = 3.9$, $p < .005$. The result is expected because the categories typically combine results from different locations.

Category Selections

There were a total of 1915 category selections made in 817 searches where the categories were used. The categories are used in 26.4% of the searches. During the last four weeks of the use, the proportion of searches with category use stayed above the average ranging from 27% to 39% (weekly averages).

The users selected on average 2.3 categories in searches where categories were used. The label of the selected category contained on average 1.9 words whereas the category names in general contained on average 1.4 words. The difference between the length of the selected and all category names is statistically significant. Independent samples *t*-test gives $t(30) = 3.2$, $p < .005$.

Findex had a default setting to compute 15 categories for a result set. Although it was possible to change this setting, it was hardly ever touched. On this list of 15 categories, the users selected the 5th (*median, sd* = 4.2) category. Users favored the beginning of the category list as the first quartile locates at 2nd, the median at 5th and third quartile at 9th category.

The number of results selected from each sequential category selection varied only moderately. From the first category, users selected on average 1.7 results. The second and third category resulted in 1.5 selections, on average. The differences are not statistically significant.

Questionnaires

There were two usage-related questionnaires. The first questionnaire was presented after a week or two of use to collect the first impressions. The second questionnaire was presented in the end of the study. The intention was to elicit users' opinions and experiences based on substantial experience.

The first questionnaire included three sections covering: 1) the usability, 2) comparison to the search engine used prior to the test, and 3) usefulness of the categories. We got

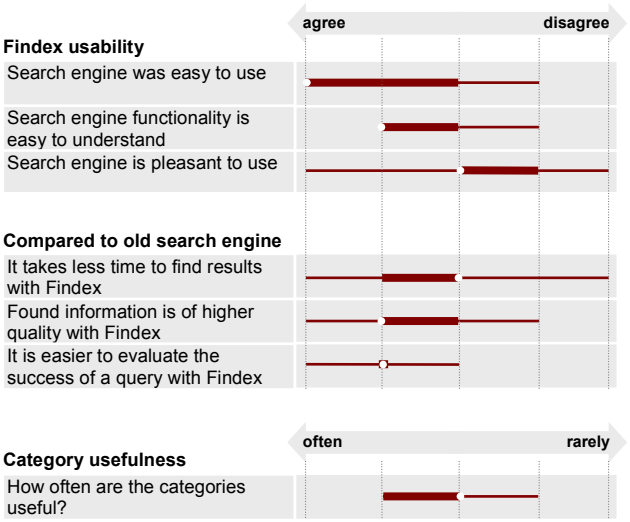

Figure 5. Answer distributions as box plots for the first questionnaire. Thin line represents range of observations, thick 50% of them and white dot the median.

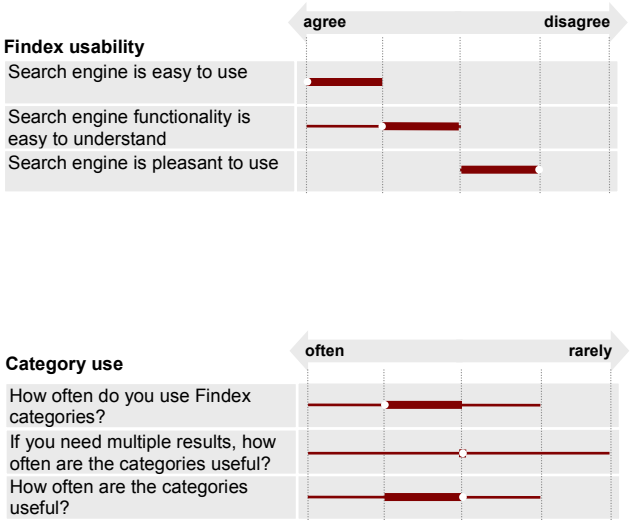

Figure 6. Answer distributions as box plots for the second questionnaire. Thin line represents range of observations, thick 50% of them and white dot the median.

answers from all of the 16 participants to this questionnaire.

The usability factors included the ease of use, understandability, and pleasantness of the user interface. Results are summarized in Figure 5. Ease of use and understandability were rated high although the variation is quite large in the responses about ease of use. Pleasantness, on the other hand, distributes the opinions even more widely. We found some explanations for this from answers to an open-ended question, such as: "the font is too small", "I would like to search only the sites in certain language", or "the layout is a bit confusing". It seems that issues not directly related to the categorization have affected the perceived pleasantness of the system.

In the second section the new search engine was compared to the one the participants were using prior to the study. Speed of use got neutral responses, but results are seen to be a little better compared to those given by the users' old search engine. In addition, the users almost unanimously agreed that the category interface helps in evaluating the success of the query. We hypothesized that one possible way of using the categories is to check if the query succeeded. The responses support this view.

The last question concerned the most important aspect of our user interface: the usefulness of the categories. The distribution of the responses to the question about how often categories are useful, is weighted towards the 'often' side and the median is 'sometimes'. This corresponds to the findings in our log data. Categories are used in 26% of the searches and in those cases about 2 category selections are made.

In addition to the closed questions summarized above, we collected the users' experiences in open-ended questions. To a question about when the categories are most useful, six respondents (out of the 11 who answered this question) replied by referring to situations when the original query was vague, broad, general, or contained words that have multiple meanings.

The second questionnaire was also divided into three sections concerning: 1) usability, 2) search habits after using Findex, and 3) use of the categories. For this questionnaire we got responses from 11 participants and the results are summarized in Figure 6.

Questions concerning the usability of Findex were exactly the same as in the first questionnaire to make comparison possible. From the results (Figure 6) we see that the variation in opinions is reduced. The main results are similar to the first questionnaire and the usability of the system is at an acceptable level. However, there is room for improvement in pleasantness.

The section on search habits shows that participants felt that their search habits changed due to the categorization interface. In particular, they stated that they use less precise search terms (45% of respondents) and that they consider the query formulation less than before (27% of respondents). On the other hand, they also felt that they use less time to evaluate the results (36%) and that they use categories for evaluating the search results (82%). These findings support a view that the categories make it possible to use less precise queries. The findings are further enforced by free comments on search habits where participants stated that: "[categories help] when the query fails (hundreds of results)" and "I didn't use phrase search ("") […] but I could write it [the query] without them. I didn't either use AND, OR, etc. operators." Participants did not, however, change their search habits completely. They reported to continue using Google on special occasions. These included searches for images or results in certain

language. Both are features that our user interface does not support.

Answers on category use are consistent with the log data findings. The answers confirm that the categories become a part of the participants' search habits. Respondents reported using categories 'quite often' (median) and benefiting from them 'sometimes' when looking for multiple result documents. In overall, categories were seen to be useful 'sometimes' (median). The distribution has moved a bit towards the 'often' side compared to the first questionnaire.

DISCUSSION AND CONCLUSIONS

In this study we wanted to find out how the category user interface is used in real life and whether its use is beneficial. We conducted a longitudinal study where 16 participants used the prototype system for two months. The users' behavior was not controlled. The research produced four main results that support the conclusion made in the earlier study: our form of result categorization is beneficial for the users.

Compared to our earlier experiment, these new results show that the question of usefulness is more complicated. In laboratory settings, the use of categories outperformed the ranked list user interface but here the list use was faster because the users could choose which method to use. When desired result is found in the top of the list, the list is faster, but when the order is not such, categories are valuable. This affects also the frequency of category use.

The results of the study confirmed that the **categories are used in real settings and that users find relevant results** with them. Participants used categories regularly (in 25% of searches) throughout the study and they found, on average, more than one result from each selected category. In addition, category use was moderate as just 2.3 categories were selected for a search. This means that users were able to discriminate the beneficial categories from the irrelevant ones. However, the usefulness was not as simple matter as the previous experiment suggested.

Both the logs and users' comments suggest that **categories are beneficial when result ranking fails**. Participants commented that they benefited from the categories when the query was vague, broad or general. In such queries the result ranking is likely to fail. This is further enforced by the observation that in navigational searches the categories were used less as the rank ordering is more likely to work.

The fact that it takes about twice the time to select the first document if categories are used explains the situation. It means that in addition to scanning the short category listing and the fairly short result listing inside the selected category, users have time to scan the first results in the original rank order (which is always initially visible). In other words, when the user does not find interesting results in the top of result listing, that is, when the rank order does not support the user's need, the categories are used. In these cases, the selected results locate further down the result list than without category use. Finding those hidden results is necessary when the ranking fails.

Observations indicate that **categories can forgive mistakes in searching**. When the categories were used, there were fewer cases where user did not find any results. This means that when the query formulation fails, the user may still be able to find some results using the categories. This view is consolidated by the users' comments stating that they use less precise query terms with categories, which is common when search topic is unfamiliar or task is exploratory in nature.

There is evidence that **categories make it easier to access multiple results**. Category use increases the number of cases were two or more results are selected. In addition, the fact that each category selection yields more than one document selection, and that typically two categories are selected in one search, point to the same conclusion. This is especially important in undirected informational searches where a broad understanding of the topic is needed.

FUTURE WORK

The basic functionality of the result categorizing is promising and we plan to continue the work. We will look into improving the solution further by experimenting with different categorization algorithms such as focusing solely on the query term contexts. The stronger emphasis on multi worded categories (where a category is defined by a phrase) is also interesting as the current results indicated that users slightly favour longer category names. In addition, we plan to integrate a software component that explains the query interpretation (including the default operator, operator misuse etc.) in natural language into the system.

ACKNOWLEDGEMENTS

This work was supported by the Graduate School in User-Centered Information Technology (UCIT). I would like to thank Anne Aula, Tomi Heimonen, Natalie Jhaveri, Kari-Jouko Räihä and Harri Siirtola for invaluable comments and discussions that contributed to this study.

REFERENCES

1. Anick, P.: Using Terminological Feedback for Web Search Refinement - A Log-based Study. *Proceedings of ACM SIGIR'03 (Toronto, Canada),* ACM Press 2003.
2. Broder, A.: A Taxonomy of Web Search. *SIGIR Forum,* Vol. 36, No. 2, ACM Press 2002, 3-10.
3. Bruza, P., McArthur, R., and Dennis, S.: Interactive Internet Search: Keyword, Directory and Query Reformulation Mechanisms Compared. *Proceedings of ACM SIGIR'00,* ACM Press 2000, 280-287.
4. Chen, H. and Dumais, S.: Bringing Order to the Web: Automatically Categorizing Search Results. *Proceedings of CHI'2000 (The Hague, Neatherlands)*, ACM Press 2000, 145-152.

5. Cutting, D., Karger, D., Pedersen, J., and Tukey, J.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of SIGIR 1992 (Copenhagen, Denmark)*, ACM Press 1992, 318-329.
6. Dennis, S., McArthur, R., and Bruza, P.: Searching the World Wide Web Made Easy? The Cognitive Load Imposed by Query Refinement Mechanisms. *Proceedings of the Third Australian Document Computing Conference, ADCS'98 (Sidney, Australia)*, University of Sidney, TR-518, 1998, 65-71.
7. Drori, O.: Display of Search Results in Google-based Yahoo! vs. LCC&K Interfaces: A Comparison Study. *Proceedings of Informing Science 2003 Conference (Pori, Finland)*, Informing Science 2003, 309-320.
8. Dumais, S. and Chen, H.: Hierarchical Classification of Web Content. *Proceedings of SIGIR 2000 (Athens, Greece)*, ACM Press 2001, 256-263.
9. Dumais, S., Cutrell, E., and Chen, H.: Optimizing Search by Showing Results in Context. *Proceedings of CHI'2001 (Seattle, USA)*, ACM Press 2001, 277-284.
10. Dumais, S., Furnas, G., Landauer, T., Deerwester, S., and Harshman, R.: Using Latent Semantic Analysis to Improve Access to Textual Information. *Proceedings of CHI'88 (Washington DC, USA),* ACM Press 1988, 281-285.
11. Hearst, M. and Pedersen, J.: Reexamining the Cluster Hypohesis: Scatter/Gather on Retrieval Results. *Proceedings of ACM SIGIR'96 (Zürich, Switzerland)*, ACM Press 1996.
12. Hölscher, C. and Strube, G.: Web Search Behavior of Internet Experts and Newbies. *Proceedings of the 9th International World Wide Web Conference (Amsterdam, Neatherlands),* 2000.
13. Jansen, B., Pooch, U. A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 3, 2001, 235-246.
14. Jansen, B., Spink, A., Bateman, J., and Saracevic, T.: Searchers, The Subjects They Search, and Sufficiency: A Study of a Large Sample of Excite Searchers. *1998 World Conference on the WWW and Internet (Orlando, USA)*, 1998.
15. Jones, S. and Paynter, G.: Automatic Extraction of Document Keyphrases for Use in Digital Libraries: Evaluation and Applications. *Journal of the American Society for Information Science and Technology (JASIST),* Vol. 53, Issue 8, Wiley Periodicals 2002, 653-677.
16. Jones, S., Jones, M., and Deo, S.: Using Keyphrases as Search Result Surrogates on Small Screen Devices. *Personal and Ubiquitous Computing.* Vol. 8, Issue 1, Springer-Verlag 2004, 55-68.
17. Käki, M. and Aula, A.: Findex: Improving Search Result Use through Automatic Filtering Categories. To appear in *Interacting with Computers*, Elsevier.
18. Pirolli, P., Schank, P., Hearst, M., and Diehl, C.: Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. *Proceedings of CHI'96 (Vancouver, Canada),* ACM Press 1996, 213-220.
19. Pratt, W. and Fagan, L.: The Usefulness of Dynamically Categorizing Search Results. *Journal of the American Medical Informatics Association.* Vol. 7, No. 6, 2000, 605-617.
20. Rose, D. and Levinson, D.: Understanding User Goals in Web Search. *Proceedings of the WWW 2004 Conference (New York, USA),* ACM Press 2004, 13-19.
21. Spink, A., Jansen, B., Wolfram, D., and Saracevic, T.: From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer,* Vol. 35, No. 3, 2002, 107-109.
22. Teevan, J., Alvarado, C., Ackerman, M., and Karger, D.: The Perfect Search Engine Is Not Enougth: A Study of Orienteering Behavior in Directed Search. *Proceedings of CHI 2004 (Vienna, Austria),* ACM Press 2004, 415-422.
23. Wu, Y., Shankar, L., and Chen, X.: Finding More Useful Information Faster from Web Search Results. *Proceedings of the ACM CIKM'03 (New Orleans, USA),* ACM Press 2003, 568-571.
24. Zamir, O. and Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results. *Proceedings of the 8th International World Wide Web Conference WWW8 (Toronto, Canada)*, Elsevier Science 1999.
25. Zamir, O. and Etzioni, O.: Web Document Clustering: A Feasibility Demonstration. *Proceedings of ACM SIGIR'98 (Melbourne, Australia),* ACM Press, 1998, 46-54.
26. Zeng, H-J., He, Q-C., Chen, Z., Ma, W-Y., and Ma, J.: Learning to Cluster Web Search Results. *Proceedings of ACM SIGIR'04 (Sheffield, UK),* ACM Press, 2004, 210-217.
27. http://snowball.tartarus.org
28. http://www.altavista.com
29. http://www.google.com/apis
30. http://www.iboogie.com
31. http://www.teoma.com
32. http://www.vivisimo.com
33. http://www.wisenut.com