ADAPTING TO EVOLVING NEEDS: EVALUATING A BEHAVIOUR-BASED SEARCH INTERFACE

Ryen W. White
Department of Computing Science
University of Glasgow
Glasgow, Scotland. G12 8QQ
ryen@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow, Scotland. G12 8QQ
jj@dcs.gla.ac.uk

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland. G12 8QQ
ir@cis.strath.ac.uk

ABSTRACT

In this paper we present an evaluation of a behaviour-based adaptive search interface that predicts the current state of a user's information need based on their interaction. We evaluate the hypotheses that our adaptive system selects additional query words that closely describe user needs and is able to accurately depict the degree of change in these needs. Our evaluation, with real users and different types of information seeking scenario, shows that these hypotheses hold.

Keywords

Evaluation, Adaptive Search Systems

1. ADAPTIVE SEARCH SYSTEMS

Adaptive search systems observe the user (via their interaction), model the user's information needs (based on this interaction), and anticipate these needs (based on the model they develop). Such systems can be classified as *behaviour-based* interface agents [4,7], that develop and enhance their knowledge of user needs incrementally from inferences made about their interaction. They have been developed to address the problems in query formulation that many novice users currently experience.

To predict what may be useful, adaptive search systems learn from a user's history of activity to improve the relevance and timeliness of their suggestions. These concepts are embodied by systems with a *just-in-time* (JIT) infrastructure, where information is brought to users as they need it, without requiring explicit requests [3].

Many adaptive systems are currently available. Letizia [5] learns current interests and predicts closely-related documents that may be useful. PowerScout [6] uses a model of user interests to construct a new complex query and search for documents semantically similar to the last relevant document. WebWatcher [1], in a similar way, accompanies users as they browse, but also acts as a *learning apprentice* [8]. Over time the system acquires greater expertise for the parts of the Web that it has visited in the past, and for the topics in which previous visitors have had an interest. Watson [3], uses text in the active document to proactively search distributed information repositories by devising a new query.

Despite the appeal of such systems, there are a number of shortcomings. With the exception of Watson and Letizia, these systems do not adequately consider a user's *current* goals. Instead, they develop a profile of user characteristics and previous search experiences founded on the belief that long-term information needs drive user interaction. Such systems allow users to communicate general interests, but take no account of active goals and that users may search for different things or wear various 'hats' during a day [10].

Adaptive search systems typically base positive relevance assessments on stereotypical search behaviours such as document retention (e.g. saving, printing, bookmarking). Users may retain a document for a number of reasons, only one of which is the relevance of its content. Also, these systems treat documents as the finest level of granularity and extract words from those active or assumed relevant. Documents can contain irrelevant parts, leading to the possible selection of erroneous or inappropriate words.

In this paper we present the evaluation of an adaptive search system that addresses these problems and is designed to predict current information needs and detect the degree of change (or evolution) in these needs. The system is compared against a strict baseline where the user is responsible for selecting query words and for indicating the extent to which their information need has evolved. The study tests how well the adaptive system detects the current state of a user's information need. Systems with a better understanding of their users' needs can produce higher quality queries and hence return more precise search results.

2. INTERFACES

We now describe the adaptive search interface and the experimental baseline against which it is tested.

2.1 Adaptive Interface

Our search interface utilises interaction with a number of representations *of the same document* to select words that represent the user's *current* information need better than their original query. These representations are highly relevant to this initial query, reducing the likelihood that erroneous words will be selected. We assume that a user's need drives their interaction, and potential query words are selected from the representations they interact with. During

the search session the system detects the degree to which the need has evolved and dynamically updates the display to depict this change. Large changes in perceived information needs result in new searches but smaller changes result in less radical actions on the user's behalf.

The systems consists of an interface (Figure 1), with underlying functionality, which connects to an existing web search engine. Once the underlying search engine has performed a retrieval, the system downloads and summarises the top thirty ranked documents. We chose this number to ensure the system responded in a timely manner. Summarisation is carried out using the sentence extraction method described in [9].

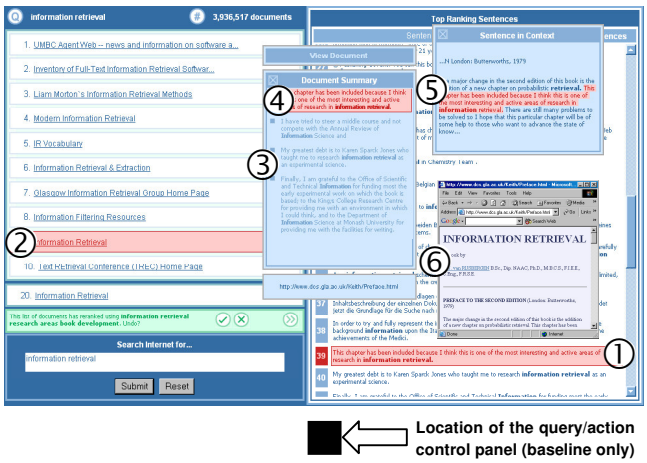

Figure 1: Adaptive search interface

Six different representations of the current document combine to form an ordered path at the results interface Firstly we have a list of sentences from all documents retrieved (thirty at a time) scored in relation to the query, we call these *top-ranking sentences* (TRS). We then have the title of the document, a summary of the document, a sentence in the summary of the document, that sentence in the context it occurs in the document and finally the document itself. There is such a path for each document, however only the title and top-ranking sentences appear on the interface at all times, the others appear on user demand.

TRS → Title → Summary → Summary Sentence → Sentence in Context → Document

① ② ③ ④ ⑤ ⑥

Figure 2: Document representations in a path (numbers correspond to Figure 1)

Passing the mouse over the title will show the summary, passing the mouse over a sentence in the summary will show the sentence in context, etc. Through their interaction, the user has control over which representations they view. A user does not need to complete a path for a document, they can stop when the information contained in the current representation is not of interest. The distance travelled along the path reflects a user's commitment to the information contained in the path and within the document.

The system selects and scores words *only* from representations viewed by the user and constructs a ranked list of words based on these scores. This ranking represents the utility of the words in describing what the user has looked at. By comparing successive lists of such words the system detects the degree to which the information need has evolved. The system then acts on the user's behalf. The action taken depends on the degree of evolution; minor changes result in re-ordering operations, major changes imply re-searching. The four possible actions, in increasing order of severity are: *no action* (for minor changes in the need), *re-order top-ranking sentences*, *re-order document list* and *re-search the web*. The top six words suggested by the system are used for this action. The system notifies the user with a message at the periphery of the interface when the action has occurred and highlights the part of the interface affected by the action. The message includes the revised query and gives the user the option to reverse the action's affect. There is no need for the user to respond to this message, and it will disappear after a short time.

2.2 Baseline Interface

To measure the adaptive system's ability to detect the current state of the user's information need, the baseline system requires the user to be responsible for selecting additional query words and choosing the action undertaken. This is the only difference between the two systems. Through comparing the performance of the adaptive system against the baseline we can measure how well the system perceives the user's information needs. Figure 3 shows the control panel allowing users to expand their initial query and choose their action.

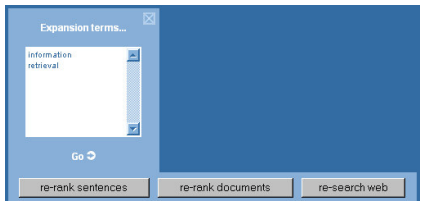

Figure 3: Query/action control panel

The words shown in the text box include the original query words. Therefore users can expand, and if required, replace their original query. There is no restriction on the number of words the user can enter. Figure 1 shows the location of control panel on the interface.

3. Evaluation

We now describe how the adaptive search interface was evaluated, including details on the experimental methodology employed, the subjects who participated and the tasks devised. We sought to test two hypotheses:

H_1: Words selected by the adaptive system closely relate to the current information need of the user.

H_2: The adaptive system's detection of information need change is an accurate depiction of a user's impression of this change.

3.1 Methodology

In our evaluation 24 subjects completed 4 search tasks, two tasks on each of the systems. The presentation of tasks to subjects was held constant: each subject performed the search tasks in the same order, however the order of presentation of systems was rotated across subjects according to a Greco-Latin square design. Subjects were given a maximum of 10 minutes to complete each task.

The subjects were given a short tutorial on the features that were incorporated into the two systems being tested and a training task to allow them to become accustomed to both systems. We also collected background data on aspects such as the subjects' experience and training in online searching. After this, subjects were introduced to tasks and systems according to the experimental design.

Once they had completed a search, the subject was asked to complete questionnaires regarding various aspects of the search. We used semantic differentials, Likert scales and open-ended questions to collect this data. In addition, we conducted semi-structured interviews after each search and after the experiment as a whole. Background logging was used to record user interaction.

3.2 Experimental subjects

We recruited 24 subjects for our experiments. Our recruitment was specifically aimed at targeting two groups of users: experienced and inexperienced users. The experienced users were those who used computers and searched the web on a regular basis. Inexperienced users were those who searched the web, used computers and the Internet infrequently. On average per week, inexperienced users spent 3.1 hours online, and experienced users spent 34.9 hours online. Overall our subjects had an average age of 26 with a range of 38 years (youngest 16 years, oldest 54 years). 14 males and 10 females participated in the experiments.

3.3 Experimental tasks

Each subject was asked to complete four search tasks from eight in total. There were four categories of search task: *fact* search, *decision* search, *background* search and *search for a number of items* [9] and subjects attempted one from each task category. The fact search asked subjects to find a single item of information (e.g. a named person's *current* email address), the background search asked subjects to find as much information as possible on a given topic, the decision search forced subjects to make a qualitative decision on the information they retrieved and the search for a number of items, asked subjects to find a number of items that meet certain criteria.

Each search task was placed within a simulated work task situation, [2]. This technique asserts that subjects should be given search scenarios that reflect real-life search situations and should allow the user to make personal assessments on what constitutes relevant material.

There were two tasks per category, each of a similar level of difficulty (verified by questions in the post-task questionnaire, and *a priori* with pilot testing) and subjects were asked to choose the task they would like to do. Users chose 51% of tasks because they were *interesting*, 21.8% of tasks because they felt they were *easy*, 19.8% because they were *familiar* with the topic area and 7.4% for *no reason*. Offering subjects a choice of tasks allowed them to select tasks that interested them and were familiar.

4. RESULTS & ANALYSIS

In this section we present the results of our system evaluation. In particular we concentrate on results pertinent to our two research questions: the adaptive system's ability to select appropriate words and to detect the extent to which the information need has evolved. Tests for statistical significance are given where appropriate with $p \leq .05$ unless otherwise stated. S_A and S_B are used to denote the adaptive system and the baseline respectively. M is used to denote the mean, and 5-point differentials are used throughout.

4.1 Information Need Detection

One of the main aims of the adaptive system is to select potentially useful words that accurately depict the user's current information need. Through monitoring user interaction with multiple representations of the same document and extracting popular words from this interaction the system is able to recommend potentially useful words and add these to the initial query. We measure the effectiveness of the system in this regard using a measure known as *term overlap*, user opinion on the usefulness of the words chosen and their informal comments.

4.1.1 Term overlap

This measure considers the degree of overlap between the words chosen by the user (in the baseline) and words chosen using the adaptive approach. A large degree of overlap would suggest that the system could accurately predict the user's *current* information need.

In the baseline system the adaptive term selection operates in the background, completely hidden from the user. The user is entirely responsible for selecting their own query words and never sees the output of the adaptive system's term selection. On average, some or all words selected by the user were also in the top 6 words proposed by the system at the same time on 72.1% of occasions. This suggests that the system selects words that are closely related to the user's own impression of their current information need.

4.1.2 Subject opinion

Participants were asked to rate on a semantic differential whether the words added to the initial query in the adaptive system were useful *always* (rating 1), *occasionally* (rating 3) or *never* (rating 5). There were no significant differences in the comparison between systems (M_B = 1.84

vs. M_A = 2.11, Wilcoxon Signed Rank test) or between groups ($M_{exp.}$ = 1.90 vs. $M_{inexp.}$ = 2.01, Mann-Whitney test). Also, it is worth noting that even though the words selected by users were useful, those selected by the adaptive system were also useful and less than the median rating (i.e. less than 3). The differences with the median were significant.

4.1.3 Discussion

The term selection in the adaptive interface was generally well received. Subjects suggested that it may be of most use when their information need is ill-defined (i.e. in the decision search), and they need support from the system. When the need is well-defined (i.e. in the fact search) they had an exact idea of what they are searching for and the adaptive system may not be of as much use. The words selected with the adaptive approach were a close match to those selected by the user.

4.2 Information Need Evolution

Another aim of the adaptive search system was to detect varying degrees of change in the user's information need. This allows the system to tailor the degree of interface support offered its user. To analyse this aspect, we elicited the opinions of those who participated and present the preliminary results in what follows.

4.2.1 Subject opinion

Participants were asked to indicate on semantic differentials their perceptions of the action, Table 1.

	Inexperienced		Experienced	
	S_B	S_A	S_B	S_A
Occurred at appropriate time	1.31	1.54	1.48	1.50
Accurately reflected change in need	1.52	1.72	2.01	2.20
Helpful for task completion	2.35	2.47	1.87	2.28

Table 1: Subject perceptions of action (lower = better)

In the baseline system, the user has control of the action. Therefore, we would expect their responses to be more positive than the implicit system. This is the case, however, what is interesting is that the differences between the systems within each group of subjects is not statistically significant using a Wilcoxon Signed Rank test for each of the differentials and a multivariate analysis of variance (MANOVA) across all differentials. The difference between groups is not significant using a Mann-Whitney test. For each of the three differentials, the results for the adaptive system are significantly less than the median value, a positive response to the action adopted.

Subjects were also asked to rate (on semantic differentials) the extent to which they felt in control of their search (M_B = 1.24 vs. M_A = 1.43) and that the system intruded on their search (M_B = 1.65 vs. M_A = 1.79). All inter-system and inter-group differences were not significant with a Wilcoxon Signed Rank test and Mann-Whitney test respectively.

4.2.2 Discussion

Subjects responded well to the action employed by the adaptive system on their behalf. All differentials were less than the median value of 3 and there was no significant difference with the baseline system, where the subject has complete control of the action. It seems the actions of the system adequately reflected the degree of change in a user's information need. The adaptive system could therefore reduce the cognitive burden imposed on the user by the need to reword and resubmit their query, selecting an action that is most appropriate for their current search.

5. CONCLUSIONS

In this paper we have presented the evaluation of an adaptive system that detects the current information need of a user and carries out actions on the user's behalf to help them in their seeking. We conducted a user evaluation comparing our system with a baseline where the user was responsible for term selection and indicating the degree of change in their information need. The results show that the adaptive system selects words that are useful and the detection of information need change is an accurate depiction of a user's own impression of this change. Devising systems that accurately perceive and adapt to current information needs is an important step in developing systems to help those struggling to find what they want on the Web.

6. ACKNOWLEDGEMENTS

The work reported is funded by the UK Engineering and Physical Sciences Research Council grant number GR/R74642/01.

7. REFERENCES

[1] Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T. (1995). 'WebWatcher: A learning apprentice for the world wide web'. *Proc. of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*.

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

[3] Budzik, J. and Hammond, K.J. (2000). User Interactions with Everyday Applications as Context for Just-in-time Information Access. *Proc. of the Annual ACM Conference on Intelligent User Interfaces*, 44-51.

[4] Lashkari, T., Metral, M. and Maes, P. (1994). Collaborative Interface Agents. *Proc. of the American Association for Artificial Intelligence*.

[5] Lieberman, H. (1995). 'Letizia: An Agent That Assists Web Browsing'. *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 924-929.

[6] Lieberman, H., Fry, C. and Weitzman, L. (2001). Exploring the Web with Reconnaissance Agents. *Communications of the ACM*. **44**(7), 69-75.

[7] Maes, P. (1994). Agents that Reduce Work and Information Overload. *Communications of the ACM*. **37**(7), 30-40.

[8] Mitchell, T., Caruana, R., Freitag, D., McDermott, J. and Zabowski, D. (1994). Experience with a learning personal assistant. *Communications of the ACM*. **37**(7), 81-91.

[9] White, R.W., Jose, J.M. and Ruthven, I. 'A task-oriented study on the influencing effects of query-biased summarisation in web searching'. *Information Processing and Management*. 2003. in press.

[10] Wolber, D., Kepe, M. and Ranitovic, I. (2002). Exposing Document Context in the Personal Web. *Proc. of the Annual ACM Conference on Intelligent User Interfaces*, 151-158.