Searchers' Assessments of Task Complexity for Web Searching

David J. Bell and Ian Ruthven[1]

Department of Computer and Information Sciences
University of Strathclyde, Glasgow, G1 1XH
dbell,ir@cis.strath.ac.uk

Abstract. The complexity of search tasks has been shown to be an important factor in searchers' ability to find relevant information and their satisfaction with the performance of search engines. In user evaluations of search engines an understanding of how task complexity affects search behaviour is important to properly understand the results of an evaluation. In this paper we examine the issue of search task complexity for the purposes of evaluation. In particular we concentrate on the searchers' ability to recognise the internal complexity of search tasks, how complexity is affected by task design, and how complexity affects the success of searching.

1 Introduction

User evaluations of search systems and interfaces attempt to assess the utility of search tools when used by human searchers. One of the main components in such evaluations are search tasks; descriptions of an information need that can used by searchers to formulate search statements and assess the relevance of retrieved documents.

In operational evaluations, such as the ones described in [1], the search tasks come from the searchers themselves. These search tasks are ones that the searcher has encountered independently of the evaluation and reflect a searcher's personal information need. This type of search task provides realistic search scenarios with which to assess a search system.

More commonly, search tasks are used within laboratory evaluations in which the experimental designer creates a number of search tasks for use within the experiment. This means that the same tasks can be used across a range of experimental subjects and systems thus allowing for a comparison of search success under different experimental conditions. A good example of created search tasks can be found within the interactive track of TREC [8]. Here the use of created search tasks allows cross-site evaluation of search systems. The nature of search tasks used in TREC changes from year to year to investigate different types of search tasks. Figure 1 gives examples of typical TREC search tasks.

[1] Corresponding author.

TREC 1999	**Aspectual recall task**	**Title**: Hubble Telescope Achievements **Description**: Identify positive accomplishments of the Hubble telescope since it was launched in 1991. **Instances**: In the time allotted, please find as many DIFFERENT positive accomplishments of the sort described above as you can. Please save at least one document for EACH such DIFFERENT accomplishment. If one document discusses several such accomplishments, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT accomplishments of the sort described above as possible.
TREC 2000	**Question answering task**	Do more people graduate with an MBA from Harvard Business School or MIT Sloan?

Fig. 1. Example TREC Interactive Track topics

Borlund [3, 4] has promoted the use of simulated work task situations in order to create more realistic search tasks. Simulated work task situations are short search narratives that describe not only the need for information but also the situation – the work task – that led to the need for information. An example, taken from [4] is shown in Figure 2. Simulated work task situations are intended to provide searchers with a search context against which the searchers can make personal assessments of relevance.

> After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Fig. 2. Example simulated work task situation from [4]

One important aspect in the creation of search tasks, whether through TREC-style topics or simulated work task situations, is how difficult it is to search using the task. Many factors can affect the difficulty of a search tasks, for example:

- *the difficulty of understanding what information is required*. Search tasks may require specialist knowledge about the task domain before starting a search, or the tasks may be too vaguely specified to allow the searcher to proceed with the search.

- *the difficulty of searching*. For some tasks it may be difficult to specify a search statement, or query, to submit to the retrieval system. For other tasks it may be difficult to find information because the collection contains little information on a given topic.
- *the difficulty of interpreting relevance*. Depending on the searcher's knowledge of a topic, or previous searching experience, it may be difficult for a searcher to decide when a document contains relevant information. For example, in the question answering tasks in Figure 1, it may be easy to assert a document contains *an* answer but not when it contains a *correct* answer. For other tasks, it may be more difficult to decide on whether a document is relevant without more information on the task area or the context of the search.

These three areas affect different parts of a search; the initial pre-search understanding of a search task, the conversion of this conceptual understanding to a search statement, and the process of assessing retrieved material. Task difficulty therefore affects the whole search process and consequently our evaluation of search systems. Tasks that are too easy may result in too little interaction for analysis; tasks that are too difficult could result in low user commitment to the experiment. It is important therefore to be able to distinguish tasks according to how difficult they may prove in an experimental setting. In this paper we explore the nature of task difficulty, in particular the nature of *task complexity*, where task complexity is a measure of the uncertainty within a search task. We carry out a study on search tasks of varying complexity within a web search environment to investigate searchers' perceptions of task complexity and how these perceptions relate to characteristics of the search tasks.

We present our methodology and components of the study in section 3, the main findings in section 4, and discuss the limitations and implications in section 5. Prior to this, in section 2, we present an overview of task complexity for information seeking.

2 Related Work

The notion of a *task* in information seeking covers a range of interrelated concepts. For example, the *work* task, e.g. preparing a research paper, relates to the activity that results in a need for information, [2, 9]. A work task may give rise to several *search* tasks, the specific search on which a user is engaged. Each individual search task involves a series of tasks and decisions relating to operating the system and assessing search results [5].

Several studies on the impact of tasks on information seeking have pointed to the importance of task complexity and the variables that can affect complexity. Kelliher, for example, relates complexity to the number of decisions to be made and indicates that, when faced with highly complex tasks, decision-makers attempt to reduce complexity by eliminating alternative actions or outcomes [10]. Vakkari [11] surveys task complexity as it has been investigated within information seeking and relates the notion of task complexity to important variables such as prior searcher knowledge, search strategies and relevance. He points to the fact that although we can categorise some of the factors that affect complexity, task complexity is not an objective

measure: personal factors can affect an individual's assessment of the complexity of a task.

Both Campbell [7] and Byström and Järvelin [6] have examined the factors that can make a task more or less complex. Campbell describes task complexity as a function of the psychological states of the task performer, the interaction between the task characteristics and the abilities of the task performer and the objective attributes of the task itself, such as the number of sub-tasks or the uncertainty of the task outcome. He proposes four attributes that can increase the complexity of a given task: multiple potential paths to a desired end-result, the presence of multiple desired outcomes, the presence of conflicting interdependencies between paths, and uncertainty regarding paths. These all can apply to information retrieval interaction; a searcher may obtain relevant information using different queries or search strategies (*multiple paths*), may require different pieces of information (*multiple outcomes*), paths may conflict (a searcher may have to split search tasks), and paths may be uncertain (the use of relevance feedback, for example, may have an unknown effect on the search). Based on the combination of these four attributes, Campbell proposed a categorization of 16 task types, e.g. simple tasks which contain none of the complexity-increasing attributes, and fuzzy tasks which contain both multiple end-states and multiple paths.

Byström and Järvelin also proposed a categorization of tasks, specifically related to information seeking and based on real-life information seeking situations [6]. This categorisation defines five levels of task complexity based on the *a priori determinability* of tasks. The *a priori determinability* is a measure of the extent to which the searcher can deduce the required task inputs (what information is necessary for searching), processes (how to find the required information) and outcomes (how to recognise the required information) based on the initial task statement. Increasing complexity is associated with increasing uncertainty regarding these factors, i.e. the less sure a searcher is about task inputs, search process or search outcomes, the more complex is the search task.

Byström and Järvelin's work was based on investigating real search behaviour in real work situations. As such it is wide in scope, incorporating aspects of the real work tasks as well as search tasks. The measure of complexity proposed in their study was based on retrospective analyses of the factors that increase or reduce the complexity of an information-seeking task.

In this study we use similar factors to test whether we can *predictively* influence the complexity of *artificial* search tasks; ones that may be applied to laboratory investigations. We also investigate how task complexity influences searchers' perceptions and satisfaction with the search process. As we discuss in section 5 the ability to manipulate and assess the complexity of search tasks can aid in the understanding and design of user evaluations.

3 Methodology of Study

In this study we create search tasks of varying complexity and use the tasks to analyse searchers' reactions to tasks of varying complexity. We use the search tasks within a

laboratory evaluation methodology, similar to those used in evaluations such as TREC, to compare the complexity of tasks within the environment in which they would typically be used. In this section we describe the main components of the study: the creation of the search tasks (section 3.1), the search systems used (section 3.2) and the participants (section 3.3). In section 3.4 we describe the methodology itself.

3.1 Search Tasks

Our model of task complexity is based on the classification proposed by Byström and Järvelin [6]. They define a five-level categorization of task complexity. We conflate this model into a three-level model to create a better separation between the complexity of tasks.

- **Complexity level 1**[2] are tasks where the tasks are almost completely *a priori* determinable. It is generally clear what information is required, how to find the information and how to assess relevance. However, some parts of the search process or information needed may be vague.
- **Complexity level 2**[3] are tasks in which the desired information may be clear, however the searcher must make case-by-case decisions regarding the inputs and search process.
- **Complexity level 3**[4] are the most complex tasks. In this type of task the whole search may be unclear from the start, i.e. it is unclear what information is being sought, how to obtain relevant information and how the searcher will know they have found relevant information.

In the study we created three groups of search task. Each of these task groups contains three variations of an individual search task, each variation reflecting a different level of complexity. An example is shown in Figure 3 for task group C. In this case, each of the three task variations is centred around the same information need – *information on changes to petrol prices.* Increasing task complexity is associated with manipulating the factors that affect the *a priori* determinability factors related to the tasks. The first of these factors involves the information input to the task. This was altered by changing the amount of information, provided by the task description, that the participant will be able to use within the search. Task C1, for example, restricts the search to the *price of petrol in the UK in recent years*, the inputs *price, recent* and *UK* provide information that the searcher can use to understand what information is being sought. Task C3 on the other hand, provides fewer clues about information can be used to search.

---

[2] Corresponds to the range of tasks between Byström and Järvelin's Automatic Information Processing Tasks and Normal Information Processing Tasks.

[3] Corresponds to Normal Decision Tasks.

[4] Corresponds to the range of tasks between Known Genuine Decision Tasks and – Genuine Decision Tasks.

Lowest complexity - complexity level 1 (Task C1)
While out for dinner one night, your friend complains about the rising price of petrol however as you have been driving for long, you are unaware of any major changes in price. You decide to find out how the price of petrol in the UK has changed in recent years.
Medium complexity - complexity level 2 (Task C2)
Whilst out for dinner one night, one of your friends' guests is complaining about the price of petrol and all the factors that cause it. Throughout the night they seem to complain about everything they can, reducing the credibility of their earlier statements so you decide to research which factors actually are important in deciding the price of petrol in the UK.
Highest complexity - complexity level 3 (Task C3)
Whilst having dinner with an American colleague, they comment on the high price of petrol in the UK compared to other countries, despite large volumes coming from the same sources. Unaware of any major differences, you decide to find out how and why petrol prices vary worldwide.

Fig. 3. Task group C

The second factor involves manipulating the process involved in finding the relevant information. A more complex task may involve comparing or analysing data from multiple sources. For the task group shown in Figure 3, the least complex task involves finding data related to petrol prices within the UK, the most complex task involves finding data related to worldwide prices. The most complex task, therefore, may not be answered by a single source, and the process of finding information becomes less clear from the start.

The final factor relates to the requested information output of the task – what information is required to complete the search task. This can be manipulated in two ways, by the amount of data required and the type of data required. For the tasks in Figure 3, the least complex tasks limits the amount of data applicable (by requesting only recent information), the UK restriction means relevant data will likely only refer to certain units (currency and volumes) that are applicable to UK petrol prices. In contrast the most complex task asks for worldwide factors that influence prices increasing the amount of data that is applicable, and, as different factors may be important in different countries, increasing the type of factors that are applicable.

The investigation therefore contrasts increasing complexity across versions of the same search task. An alternative would have been to create unique tasks of varying complexity. However it can be difficult to assess the relative complexity of tasks on different topics. Our methodology reduces the overall number of search topics to be created, and allows comparison between different versions of the same core task. The tasks were framed within simulated work task situations to encourage personalised searching by the participants.

In pilot testing we created several task groups. The three search groups that displayed the best variation in task complexity, as assessed by participants in the pilot study, were chosen for the final study. The three search tasks will be referred to as groups A-C[5], within each group the individual search tasks are numbered from 1-3

5 Task groups A and B are given in the Appendix.

with 1 reflecting the task with the lowest complexity, e.g. A1 is the task in group A with the lowest complexity.

3.2 Search Systems

In the study we asked the participants to search using the search tasks. We used two search interfaces. Both systems were interfaces to the WiseNut[6] internet search engine. Two search interfaces were employed in the study to be able to generalise searchers' assessment of search task complexity beyond the interface itself, i.e. so that the measurement of complexity is not solely a factor of the individual interface used. The interfaces are described in detail in [12, 13][7] in this section we shall only describe the main features of the two interfaces used. Screen-shots of the two interfaces are given in Appendix A.

The first interface, **Sum-Int**, is a summarisation interface [12], Appendix Figure A.1. Titles of retrieved web pages are shown in groups of ten and moving the mouse over the title of a retrieved page will displayed a short summary of the web page to the searcher. The summaries themselves are composed of the top four sentences in the web page that are the best match to the searcher's query.

The second interface, **TRS-Int**, also offers a summary of retrieved documents, Appendix Figure A.2. This interface also displays to the searcher a list of sentences taken from the top 30 retrieved documents, the *top-ranking sentences*, ranked in order of how well the sentence matches the searcher's query. The intention behind this feature is to help the searcher locate relevant information regardless of which document contains the information. This has previously been shown to be useful in helping the searcher identify relevant material [13]. In TRS-Int, each the title of each retrieved page is associated with a check-box. By clicking on the check-box the searcher can indicate to the system that the retrieved page contains useful information. If the searcher clicks on a check-box the contents of the page's summary is used to modify the searcher's query and the list of top-ranking sentences is updated to reflect the new query. This form of relevance feedback is intended to keep the most useful sentences at the top of the list of sentences.

3.3 Participants and Methodology

30 people participated in the main study: 9 female and 21 male. All participants were aged between 18 and 25 years and were university students from a variety of academic disciplines. Each participant was asked to search on three search tasks, one from each of the three search groups (A-C) and were given 5 minutes to search on each task. The time restriction was based on pilot testing which indicated that 5 minutes was sufficient time for most participants to complete most of the tasks. In presenting the tasks to the participants the order of search task *topic* was held constant (the participants received a task from group A followed by one from group B, finally

6 http://www.wisenut.com/

7 We gratefully acknowledge the support of Ryen White of the University of Glasgow in providing these interfaces.

a task from group C), however the complexity of the search tasks were rotated using a Greco-Latin square design, e.g. participant 1 received tasks A1, B2, C3, participant 2 received tasks A2, B3, C1, etc. None of the participants had previously used either search interface. Each participant searched only on one of the search interfaces to avoid the participants having to cope with two novel search interfaces.

4 Results

In this section we present the main results of this investigation. Our analysis is focussed on the three main aspects of the investigation: the participants' ability to recognise task complexity, the factors that affect complexity and the relationship between complexity and the participants' interest in the tasks. In each of the following sections we will develop the main research hypotheses being investigated.

4.1 Participants' Perceptions of Complexity

In this section we investigate our core hypothesis, namely that by modifying the search tasks in the manner described in section 3.1 we create search tasks that have recognisably different levels of complexity. In one sense, this is a validation test for our approach to manipulating task complexity: if there is no difference between reported assessments of task complexity then it may that searchers *can* recognise task complexity but our method of creating complex tasks is poor. One the other hand, if the participants report clear differences in task complexity then we can conclude that task complexity can be recognised and that our method creates tasks of varying complexity. Our research hypothesis is, therefore, that participants can differentiate the complexity of the employed search tasks.

To investigate this, after each search, participants were asked to record the overall complexity of the search task on a 5-point scale in which a value of 1 reflected a task with little complexity and a value of 5 indicated a highly complex task. Table 1 (row 2) summarises the results from the participants' assessments of the tasks' complexity. As can be seen, for all task groups, the participants' rating of task complexity increases according to the predicted complexity of the task. This provides an initial validation of the method of varying task complexity. The responses for the tasks A1, and C3 were significantly different[8] from the other tasks in the task group and all tasks in group B were significantly different from each other.

[8] Using a one-tailed Mann-Whitney test for independent samples, $p<0.05$

Table 1. Average rating of task complexity

	A1	A2	A3	B1	B2	B3	C1	C2	C3
Complexity	2.2	2.9	3.5	1.6	2.5	2.8	2.2	2.4	3.8
Completion	3.1	2.8	2.3	3.6	3.4	2.9	3.2	2.5	2.3
Process	2.4	2.8	2.7	3.7	3.7	2.8	3.2	3.0	2.1

Following from this initial hypothesis we investigate two possible related aspects; perceived task completion Table 1 (row 3) and ease of finding information Table 1 (row 4). In particular we measured responses to the degree to which the participants felt they had completed the task and how simple they felt it to find information (process). Both are measured on a 5-point scale in which a value of 5 reflects greater sense of task completion or a simpler process of finding information.

Generally the participants' assessment of task completion was inversely correlated with their assessments of task complexity; the more complex a task was rated, the less likely the participants were to feel that they had completed the task.[9] The actual correlation figures are discussed in section 4.4.

For task groups B and C there was also an inverse correlation with the participants' assessments of how simple was the process of finding information: the more difficult was the process of finding information the more complex the task was perceived as being.[10] However, this is does not hold for task group A. There is, therefore, some support for the difficulty of finding information, while not a complete determinant in the assessment of complexity, playing a part in complexity. In the next section we examine what causes the difference in complexity assessment.

4.2 Factors Affecting Complexity

The factors that were used to differentiate between the tasks in each task group were related to the *a priori* determinability of the search task; based on the task description how easy was it for the searcher to elicit useful information from the task description on what information was required, how easy was it to recognise relevant information and how clear was it to decide how relevant information was to be found.

To investigate which of these factors affected complexity, the participants were asked to rate the tasks according to three questions, again using a 5-point scale with 5 reflecting highest level of agreement: '*Useful information was provided by the task*', '*The type of information to be retrieved was clear*' and '*The amount of information to be retrieved was clear*'. Table 2 summarises the participants' responses. Generally we would predict that the values would decrease from left to right, i.e. as less useful information is provided, or less information on the type or amount of information required is given, then task complexity would increase. Even though the differences between the scores for utility of information were slight, this relationship generally

[9] Significance testing showed significant differences between the scores for tasks A1/A2, A1/A3, B1/B3 and C1/C3.

[10] Significance testing showed significant differences between the scores for tasks B1/B3, B2/B3, C1/C3 and C2/C3.

holds across the tasks with the higher complex tasks receiving scores less than or equal to the less complex tasks.[11] Therefore as the task expresses less useful information on what information is required, or less information on the type or amount of information to be retrieved, the participants perceive the task to be more complex.

Table 2. Participant responses to complexity increasing factors

	A1	A2	A3	B1	B2	B3	C1	C2	C3
Useful information was provided	3.3	2.5	2.4	3.7	3.1	3.1	3.4	2.8	2.8
Information type was clear	4.2	3.0	2.5	4.3	3.9	3.3	4.1	3.6	2.9
Information amount was clear	4.2	3.6	2.4	4.2	3.5	2.1	3.3	3.2	2.2

4.3 Personal Reactions to the Search Tasks

As mentioned previously, a searcher's estimate of task complexity can be influenced by subjective factors such as how much knowledge the searcher has about the task. In this section we examine the participants' reactions to the assertions '*This task was easy to understand*', '*The task was interesting*' and '*The task was relevant to me*'. In Table 3 we summarise the participants' responses. Answers are given on a 5-point scale, with a value of 5 reflecting the highest level of agreement.

Table 3. Participant responses to personal reactions

	A1	A2	A3	B1	B2	B3	C1	C2	C3
Easy to understand	3.5	2.9	3.2	4.1	3.8	3.1	3.8	3.1	3.0
Task was interesting	3.1	2.7	2.9	3.9	3.5	3.3	2.8	2.3	3.0
Task was relevant to me	3.6	3.4	3.5	3.8	3.2	3.1	3.1	2.5	2.7

There were few patterns regarding the latter two aspects (task interest and task relevance) except that tasks that had a lower complexity were more likely to be judged as more interesting or relevant than more complex tasks. However, there were no significant differences found regarding these two aspects. Within each group the search tasks were based on the same core topic, e.g. changes in petrol prices, therefore we might not expect any differences between the responses *within* a task group. That is, we might not expect a participant to be more *interested* in the topic of petrol prices whether the topic is placed within a highly complex or less complex search task. The lack of significant differences across task groups might simply reflect the individual differences in topic interest among our subjects. On the other hand, it may also reflect the fact that searchers who are closer to task completion, those who are searching on less complex tasks, are more likely to have obtained interesting information earlier in the search.

[11] Significant differences between comparisons A1/A2, A1/3, C1/C2 for utility of information, A1/A2, A1/A3, B1/B3, B2/B3, C1/C3 for type of information required, and A1/A2, A1/3, B1/B2, B1/B3, B2/B3, C1/C3, C2/C3 for amount of information required.

A similar pattern arises regarding how easy it was to understand the task with the lowest level complexity being rated as easier to understand on all groups. Here there were stronger differences, with, B3 significantly lower than B1, and C1 significantly higher than C2 or C3. So the ability to understand the task is related to the assessed complexity. From discussions with the participants, this was related to the *a priori* determinability: the participants' ability to understand what was required from reading the search task.

4.4 Correlation analyses

In this section we examine the correlation of participants assessments of complexity against the various aspects described in sections 4.1 to 4.3 to compare the relative importance of each aspect. In Table 4 we show the results of applying Spearman's Rank Correlation Coefficient to the participants responses.

Table 4. Correlation of questionnaire responses with assessments of task complexity

	Process	**Completion**	**Useful**	**Type**	**Amount**	**Understanding**	**Interesting**	**Relevant**
A	-0.24	-0.22	-0.33	-0.65	-0.66	-0.12	0.00	0.34
B	-0.73	-0.59	-0.40	-0.68	-0.79	-0.71	-0.56	-0.31
C	-0.53	-0.53	-0.48	-0.74	-0.69	-0.49	0.05	-0.12

Across the task groups there was a constant relatively high inverse correlation of complexity with the type and amount of information required being clear from the task. Indeed, for each task group the strongest correlation was with the amount of information required. There was generally little correlation, however, with how interesting or relevant the task was to the searcher although the ability to understand the task set was important in task groups B and C. In this study, therefore, the information requirements of the task – how much information is required and what type of information – and the searcher's ability to understand these requirements appear to be more strongly linked to complexity than issues such as interest or relevance.

In task groups B and C the complexity was inversely related to task completion and the reported simplicity of the information-seeking process. This demonstrates the importance of assessing complexity when assessing the results of user evaluations.

4.5 Cross-system Comparisons

As mentioned in section 3.2 we used two search interfaces to be able to generalise the results beyond a single interface. We compared the results of the questionnaires for each task when performed on the two interfaces using a two-tailed Mann-Whitney test for independent samples, $p<0.05$. Although the numbers of responses used for each comparison are small, there were no significant differences found with the exception of responses to the assertion '*The task was relevant to me*' which were significantly higher for task C3 on the TRS-Int than on the Sum-Int.

5 Discussion

This study examined the impact of search task complexity on web searching. We created sets of search tasks using Byström and Järvelin's five-level model as the basis of our characterisation of task complexity. Using the created search tasks we examined whether web searchers could recognise task complexity and how this impacted on issues such as search success and searcher satisfaction.

There are several limitations to the study. For example, our study is limited in only examining search tasks rather than the whole work task that promotes individual search tasks. Also, our subjects only experienced one task from each complexity level rather than running several tasks from each level. Finally, the differentiation between the complexity of individual search tasks may not have been sufficiently great to properly determine the effect of complexity on other factors such as searcher understanding. Creating search tasks itself can be a difficult task as tasks can vary along other dimensions as well as complexity and these dimensions can interact. For example, one repeated comment was that some tasks were more complex because there was limited information available.

Our main aim is to promote task complexity as a factor in designing and interpreting user evaluations. In such evaluations, e.g. [12], questionnaire results on aspects such as searcher satisfaction, task completion, etc. are used in a comparative situation, e.g. System A leads to greater satisfaction than System B. However, it is not only the relative findings that are important; the absolute scores given to questionnaire responses are also important. If few searchers report reasonable search satisfaction, or task completion then the evaluation itself may be flawed. Assessing task complexity in pilot or pre-testing can be a useful method of determining whether search tasks are appropriate for individual evaluations. The method of using the same basic task, but varying the complexity, can elicit which version of a task is most appropriate for a given experimental study.

The *a priori* determinability can be used as a simple means of initially varying the tasks for pre-testing but the actual task complexity can be amplified or reduced by other factors such as the searcher's interest in the topic. This also provides additional support for Borlund's assertion that search tasks should be tailored to the experimental subject group [3].

Finally, we should recognise that task complexity is a dynamic entity. Tasks that offer little complexity may be answered quickly and by one document. However, tasks that are more complex may require several searches and aggregation of information from several documents or several search iterations. In user evaluations it is common to allow searchers only a certain time-frame in which to complete a search. This allows for stricter comparison between searches by different people or searches on different search systems. However, if the time given is too short then the searcher may not move from the stage of collecting information to the process of deciding on relevance and completing the search task. Therefore tasks may seem more complex at earlier search stages, when the searcher is collecting information, than in later search stages. In evaluations we should select appropriately complex tasks for the time we give to searchers or, alternatively, use measures of complexity to decide how much time we should allow searchers to complete a search task.

In summary, our findings validate Byström and Järvelin's model of task complexity and propose this model as a means of predicting and manipulating task complexity. The findings also indicate that task complexity should be seen as an integral part of designing and interpreting user evaluation results.

Acknowledgements

We would like to acknowledge the contribution of our participants to the work described in this paper and the many helpful comments from the anonymous reviewers.

References

1. M. Beaulieu. Experiments with interfaces to support query expansion. Journal of Documentation. 53. 1. pp 8-19. 1997.
2. N. J. Belkin, H. M. Brooks and R. N. Oddy. Ask for information retrieval. Journal of Documentation. 38. 2. pp 61-71. 1982.
3. P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation. 56. 1. pp 71-90. 2000.
4. P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research. 8. 3. 2003.
5. M. K. Buckland, and D. Florian. Expertise, task complexity, and artificial intelligence: A conceptual framework. Journal of the American Society for Information Science. 42. 9. pp 635-643. 1991.
6. K. Byström and K. Järvelin. Task complexity affects information seeking and use. Information Processing and Management. 31. 2. pp 191-213. 1995.
7. D. Campbell. Task complexity: a review and analysis. Academy of Management Review. 13. pp 40-52. 1988.
8. W. Hersh and P. Over. The TREC 2001 interactive track NIST Special Publication 500-250: The Tenth Text Retrieval Conference (TREC 2001). p 38. 2002.
9. P. Ingwersen. Information retrieval interaction. Taylor Graham. London. 1992.
10. C. Kelliher. An empirical investigation of the effects of personality type and variation in information load on the information search strategies employed by decision-makers. Texas A&M University Ph.D. 1990.
11. P. Vakkari. Task complexity, information types, search strategies and relevance: integrating studies on information seeking and retrieval. Information Processing and Management. 35. 6. pp 819-837. 1999.
12. R. W. White, J. M. Jose and I. Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. Information Processing and Management. 39. 5. pp 707-733. 2003.
13. R. White, I. Ruthven and J. Jose. Finding relevant documents using top-ranking sentences: an evaluation of two alternative schemes. Proceedings of the 25th Annual International ACM SIGIR Conference (SIGIR 2002). Tampere. pp 57-64. 2002.

Appendix A

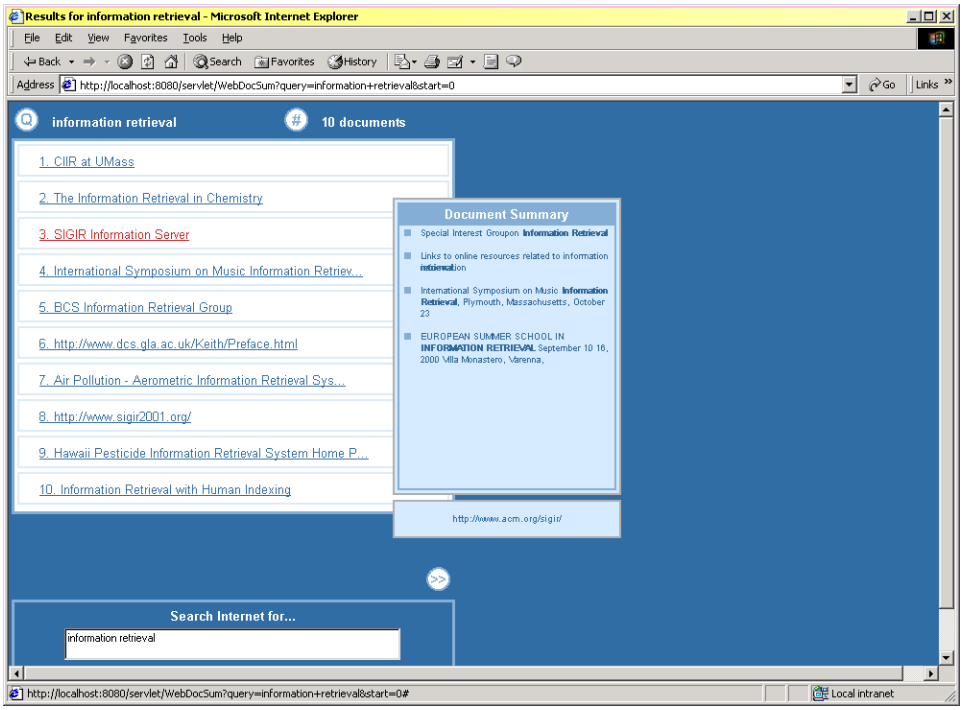

Fig. A.1. Interface one

Table A.1. Task groups A

Lowest complexity (Task A1)
A friend has recently been applying to various universities and courses but has been complaining that they are finding it difficult to attain a place due to the rising numbers of students. You were unsure if their assessment was correct so you have decided to find out how the size of the student population changed over the last 5 years and how it is expected to change over the coming 5 years.
Medium complexity (Task A2)
A friend has recently been applying to various universities and courses but has been complaining that they are finding it difficult to attain a place due as a much larger and varied number of people are attending university. You were unaware if their assessment was correct so you have decided to find out how the composition of the student population has changed over the last 5 years.
Highest complexity (Task A3)
A friend who has been attempting to gain a university place has been complaining that there are too many people attending university today, you were unsure if this assessment was correct and have decided to find out what changes there have been in the student population in recent times.

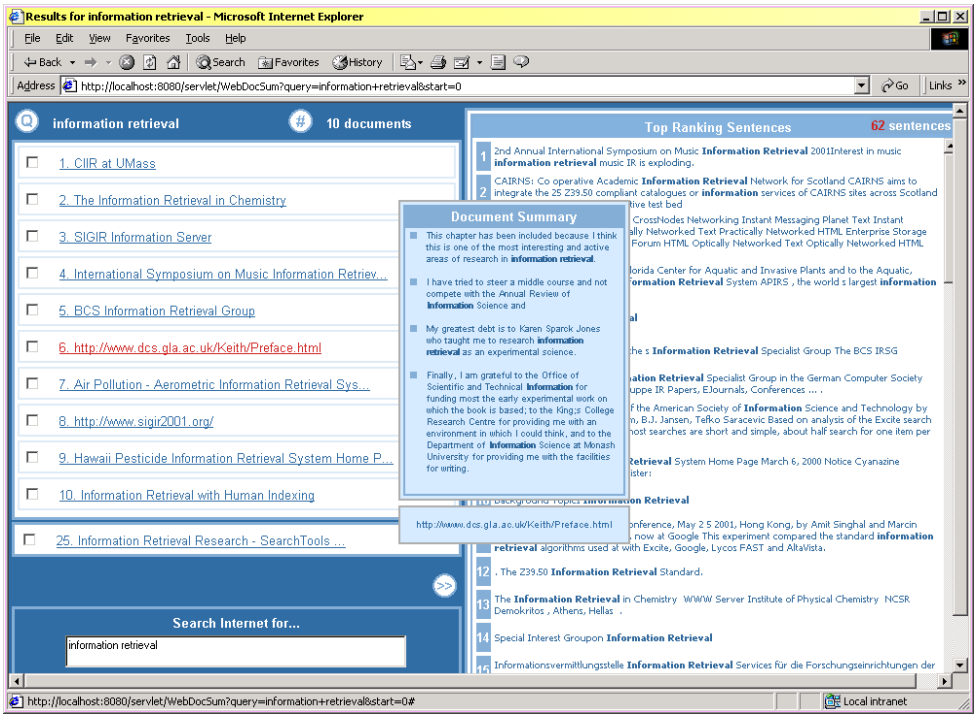

Fig. A.2. Interface two

Table A.2. Task groups B

Lowest complexity (Task B1)
Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to wait until 3G or 3rd Generation phones are available before purchasing a new one. The staff are looking for a quick sale and don't seem to be very forthcoming with information on this technology so you decide to find out yourself what special features will be available on 3G or 3rd Generation mobile phones before making a decision.
Medium complexity (Task B2)
Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to wait until 3rd Generation phones are available before purchasing a new one. The staff are looking for a quick sale and don't seem to be very forthcoming with information on this technology so you decide to find out yourself what special features will be available on 3rd Generation mobile phones before making a decision.
Highest complexity (Task B3)
Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to wait to buy a 3rd Generation phone. Your friend didn't want to be sucked into buying something that may soon be obsolete so has asked you to explain 3rd Generation mobile phone technology to them.