An Ostensive Browsing and Searching on the Web*

Hideo Joho, Robert D. Birbeck, and Joemon M. Jose

{hideo,birbecrd,jj}@dcs.gla.ac.uk
Department of Computing Science, University of Glasgow
17 Lilybank Gardens, Glasgow G12 8QQ, UK.

Abstract. The ostensive model assumes that a user's information need is dynamic and developing, thus, a recently accessed object can be seen as more indicative to the current information need. The model has been proved to be effective in image retrieval. This paper investigates the effectiveness of an ostensive model applied to web retrieval, where query-biased sentences are used to implicitly capture an underlying information need and to support a user's browsing of search results. Our study suggests that the sentence-based approach to an ostensive browsing is promising to facilitate an effective exploration of search results.

1 Introduction

Relevance feedback is one of the critical components in information retrieval (IR) systems. Leveraging a searcher's feedback to improve retrieval effectiveness is a form of system's adaptation to an underlying information need. A criticism of the existing relevance feedback models such as [1] is that they often assume that the underlying information need is static during the search session. Bates [2] and Kuhlthau [3] argue that this does not always represent the searching behaviour of real searchers. They suggest that information needs and search goals are often dynamic and developing during the search. In addition, Pharo and Järvelin [4] suggest that the searching behaviour can be irrational when the searchers face a complex problem. Several models have been proposed by researchers, where the dynamic nature of information needs was taken into account in one way or another [5–8]. Of those, the *ostensive model* (OM) proposed by Campbell and Van Rijsbergen [5] is particularly interesting because it offers a simple but effective way of capturing the developing information need for relevance feedback. The OM has been applied to image retrieval [9, 10]. The model's success in image retrieval appears, partly, to be due to the representation of information objects (e.g., thumbnail image) used in the search result presentation. The representation of objects is important in the OM since it is used by the searcher to interact with the search interface, and since it is used by the system to capture relevance feedback implicitly.

In this paper, we present an application of the ostensive model in Web retrieval, where the top ranking sentences (TRS) [11] are used as the primary representation of information objects for the browsing of search results. There are several motivations for using TRS in our application. First, TRS is a query-biased summary of a document [12], thus, it can be a promising representation for an application of the adaptive models

* This work was supported by EPSRC (Ref: EP/C004108/1).

Fig. 1. A screenshot of an ostensive browsing interface

such as the OM. Second, the generation of effective TRS has been established by a series of studies [11, 13–15]. In the existing studies, however, the TRS was presented in a static manner. In this study, the sentences were dynamically ranked by an ostensive model to help searchers find relevant information. The rest of this paper is structured as follows. Section 2 presents the interface of a sentence-based ostensive browsing. Section 3 describes the experimental design of our user study. Section 4 presents the results of the evaluation. Finally, Section 5 discusses our findings and future work.

2 Sentence-based ostensive browsing

Our approach to an ostensive browsing was based on query-biased sentences [12]. For each record of the URLs retrieved by Google, up to three sentences were extracted from the document using a version of the software originally developed by [13]. The software extracted candidate sentences from retrieved documents and ranked them based on a mixture of factors such as the frequency of query terms, document location, and HTML tags. In our interface, the sentences were then appended to the search result, as shown in Fig.1. In the interface, the words from click-through sentences were used to implicitly capture a user's underlying information need. More specifically, when a set of URLs were retrieved in response to a query, the sentences were extracted from the URLs, and content-bearing words were stored in a document-term matrix. The words were given an initial weight based on *TF*IDF* within the set of all top ranking sentences (as opposed to a full-text). When a sentence was accessed in the result, the weight of the words that appeared in the sentence was updated. A new set of sentences were then ranked by the current weight of words and presented to a user. The weight of words was consistently updated as the user interacted with the sentences. A higher weight was given to the words that occurred in a more recently accessed sentence. More specifically, the initial weight was updated by a linear combination with the sum of *ostensive relevance value* [5], defined as $\frac{1}{2^k}$, where k was the distance from the latest interaction. While a more sophisticated function can be used to update the weight [14], we decided to keep it simple since it was not our aim to investigate an optimal ostensive function.

The effectiveness of TRS has been studied in a series of experiments conducted by White, et al. [11, 13–15]. Compared to their system, our interface was intentionally designed to be a simple extension of an existing search engine's result presentation. However, our interface enabled users to browse the retrieved documents via an ostensive

presentation of TRS. The main objective of our study is to investigate the effects of a sentence-based ostensive browsing devised for the effective exploration of retrieved documents in a user's information searching behaviour. The next section describes our experimental design to address the research objective.

3 Experiment

A repeated measures within-subject design was used for our experiment, where the independent variables were the system and subject group (see below). The experiment contained a range of dependent variables due to our holistic approach to user-centred evaluation in IIR. Yet, they were largely grouped into participants' browsing of search results, query re/formulation, and their overall task performance. The dependent variables were measured by the post-search questionnaires as well as user interactions with the interfaces recorded by the system. This section presents the details of our experimental design.

Participants A total of 24 participants were recruited for our experiment. The recruitment was carried out by our call for participation distributed to the mailing lists of the University of Glasgow and in a subsequent word-of-mouth fashion. Participants were divided into two groups (twelve each) based on their background. The first group consisted of the undergraduate and postgraduate students in Computer Science (CS) fields who tended to have more search experience than the second group. The second group consisted of the people from various backgrounds (but not CS) who tended to have less search experience than the first group. In this paper, the first group is called *More Experienced* group and denoted as G_1 while the second group is called *Less Experienced* group and denoted as G_2. The entry questionnaire established that the age of our participants ranged from 19 to 50 with an average of 27.8. The average age of the More Experienced and Less Experienced Group was 21.1 and 34.5, respectively. The More Experienced group had on average 7.9 years of search experience (standard deviation: $\sigma = 1.4$) while the Less Experienced group had on average 4.4 years of search experience ($\sigma = 2.0$).

Systems Three systems were devised for our experiment. All systems presented the 10 retrieved records per result page. The first system (System 1, denoted as S_1) was a control system where up to three TRS were appended to the existing document surrogate (title, snippet, url, size, etc.) of individual retrieved records. The presentation of TRS in System 1 was static and no further browsing was available. The second system (System 2, or S_2) was the same as System 1 except that the ostensive presentation of TRS was implemented as discussed in Section 2. When a user *hovered* the mouse pointer on a TRS of retrieved records, three new TRS were extracted from other retrieved records and presented in a cascading menu style. After some informal experimentation on the visualisation, we decided to present up to three levels of menus since it appeared to provide reasonable readability of TRS without cluttering the screen. A more detail measure of appropriate levels for the TRS presentation is beyond the scope of this experiment. When a TRS was *clicked* from the cascading menu, a new window was opened to show

the contents of the page where the TRS was extracted. The top 30 retrieved URLs were used to extract and rank TRS for the ostensive browsing. The third system (System 3, or S_3) was the same as System 2 except that query terms were suggested based on user's browsing of TRS. The words appeared in the browsed TRS were recorded and ranked by the OM function. The top six words[1] (except stopwords) were suggested to user by updating the query box in the system interface. We did not include an interface that had no TRS in our experiment because past work (e.g., [13]) has already demonstrated the benefits of TRS compared to such an interface.

Tasks Participants were asked to carry out three search tasks in the experiment. One of our research interests was to evaluate the effectiveness of the proposed interfaces based on a range of search tasks. The tasks were designed based on the simulated work task situation framework [16]. The framework described a task as a form of short scenario. The scenario explained the contexts and motivation of the search with sufficient information about the relevance of pages. An overview of the tasks used in our experiment is as follows.

Task 1: Background search task. This task asked participants to find general background information on a topic. In our experiment, participants were asked to find the pages which provide information about the recent change of student populations.

Task 2: Decision-making task. This task asked participants to make a decision about a topic. In our experiment, participants were asked to find the best Hi-Fi speakers available in a target price. Participants were encouraged to compare the speakers' details in the decision making process. Task 1 and 2 were based on the descriptions originally proposed by [15].

Task 3: Many items task. This task asked participants to find as many items as they feel necessary about a certain topic. In this experiment, the task involved finding out interesting things to do at the city of Kyoto in Japan for a free weekend there. This task was a variant of aspectual search devised in the Interactive Track of TREC [17].

Procedure The user study was carried out in the following manner. At arrival time participants were asked to read an information sheet which described an overview of the experiment and guideline for the participation. Upon the agreement of participation, participants were asked to fill in an entry questionnaire to indicate their background information. Then they were presented with a training topic and explained the nature of simulated-work task. They were given approximately 10 minutes to familiarise with the search interfaces and task activity. During the training session, the three systems were introduced to participants and questions regarding the interface and tasks were answered. During the tasks, participants were asked to bookmark the pages when relevant information was found. However, no explicit instruction was given to participants regarding the number of bookmarks required to complete the tasks. All participants have used the bookmarking function of web browsers in the past and they did not express any difficulty of bookmarking during the experiment. Participants were given 15 minutes to complete a task, but were allowed to end it when they felt they had completed

[1] This size was selected based on a study of a TRS-based system [13].

the tasks. After the first task was completed, participants were asked to fill in a post-search questionnaire to provide subjective assessments about their search. A new task was then given to them and the change of system was informed. The same procedure was repeated three times. Each participant carried out all three tasks using a different order of the three systems. To reduce the bias of system, participants were systematically assigned to one of the following orders of the system: S_1-S_2-S_3, S_1-S_3-S_2, S_2-S_1-S_3, S_2-S_3-S_1, S_3-S_1-S_2, and S_3-S_2-S_1. Since the type and domain of search tasks used in our experiment were different, the order of tasks remained consistent across participants. When the three tasks were completed, participants were asked to fill in an exit questionnaire to indicate their overall preference of system, followed by an open-ended interview to capture their feedback and comments about the result presentation and experiment. The whole session tended to take between 1.5 to 2.5 hours. Participants were rewarded with £5 for their participation.

4 Results and analysis

This section presents the experimental results of our study based on 72 searches carried out by 24 participants. The results presented in this section, unless otherwise stated, is the mean value of 12 and 24 searches for G_1/G_2 and G_{1+2}, respectably, across the systems. The standard deviation of the mean values are given in the brackets. As for the statistical tests, we opted for the non-parametric tests due to the lack of the normal distribution assumed in our data set [18]. The Friedman Test was run to establish the statistical significance ($p \leq .05$) of the differences observed among the three systems (S_1, S_2, and S_3). When a difference was found to be significant, the post hoc test (Wilcoxon Signed Ranks Test) was carried out to find a significant pair(s) through the multiple pairwise comparisons of the three systems. To take an appropriate control of Type I errors, the significance level was set to $p \leq .0167$[2] in the post hoc tests, based on the Bonferroni correction [19]. The same procedure was applied to the results based on all participants (denoted as G_{1+2}). Furthermore, the Mann-Whitney U test was used to establish the statistical significance ($p \leq .05$) of the differences observed between the two subject groups (G_1 and G_2).

This section is structured as follows. Firstly, the experimental results that are related to the browsing of search results are presented. Secondly, we present the results regarding participants' query re/formulation process. Finally, participants' perceptions on the search tasks and their overall task performance are presented, followed by their system preference.

4.1 Browsing of search results

All the systems evaluated in this study presented up to three top ranking sentences (TRS) in the individual retrieved records, in addition to the existing surrogate components such as the title, snippet, URL, and file size. The difference between System 1 and System 2/3 was the functionality of the ostensive presentation of TRS, which was

[2] That is .05 divided by 3 pairwise comparisons.

Table 1. Ease of browsing and finding rel docs (Range: 1-7, Lower = Easier)

Ease of browsing search results	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	2.1 (0.9)	2.1 (0.8)	1.8 (0.6)
Less Experienced (G_2)	2.5 (1.5)	2.4 (1.6)	1.7 (1.4)
All participants (G_{1+2})	2.3 (1.2)	2.3 (1.3)	**1.7** (1.0)
Ease of identifying relevant docs	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	3.3 (1.1)	2.3 (1.3)	2.0 (0.9)
Less Experienced (G_2)	2.5 (1.5)	2.4 (1.6)	**1.4** (0.7)
All participants (G_{1+2})	2.9 (1.4)	2.3 (1.4)	**1.7** (0.8)

designed to facilitate a user's browsing of search results using a set of query-biased sentences. Therefore, this section investigates the effect of the new presentation for the browsing of search results.

Table 1 shows participants' subjective assessment on the ease of browsing the search results during the tasks. Participants were asked to indicate their assessment by the question "*How easy was it to browse the search results and find the relevant information?*". The assessment was captured by a 7 point scale where a low score represented a more positive perception in the analysis. As can be seen, the difference between System 1 (S_1) and 2 (S_2) was small, while System 3 (S_3) tended to have a more positive score than the other two systems. The result seems to be consistent across the subject groups. The Friedman Tests show that the differences are significant in G_{1+2} ($\chi^2(2) = 7.682$, $p = .022$) but not in the individual subject groups. The post hoc tests show that the difference between S_1 and S_3 in G_{1+2} is statistically significant ($Z = -2.412$, $p = .010$). This suggests that, in overall, participants found S_3 easier to browse the search results and find relevant information than S_1. However, since we did not find a significant difference between S_1 and S_2, the query suggestion offered in S_3 appeared to influence their assessment in this question.

Table 1 also shows participants' assessment on the ease of identifying perceived relevant documents. Participants were asked to indicate their assessment by the question "*How easy was it to identify a relevant document from the results presented?*". As such, this question focused on the relevance assessments on the search results. Participants' perceptions were captured in the same manner as the previous question. As can be seen, both S_2 and S_3 tended to have a more positive score than S_1 in the More Experienced group (G_1). In the Less Experienced group (G_2), on the other hand, the difference between S_1 and S_2 was small but S_3 tended to have a more positive score than the other two systems. The Friedman Tests show that the differences are significant in G_2 ($\chi^2(2) = 10.231$, $p = .003$) and G_{1+2} ($\chi^2(2) = 14.381$, $p = .000$). The post hoc tests show that the difference between S_1 and S_3 is significant in G_2 ($Z = -2.410$, $p = .008$) and G_{1+2} ($Z = -3.388$, $p = .000$). Since the p value in G_1 (.053) was close to .05, we also ran the post hoc test in G_1. The difference was found to be significant between S_1 and S_3 ($Z = -2.570$, $p = .006$), but this should be taken as a tentative result. Overall, these results suggest that participants found S_3 easier to identify relevant documents from the search results compared to S_1. The results also suggest that this trend can be more evident for participants in G_2 than G_1.

Table 2. Number of result pages viewed

	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	5.5 (2.8)	5.4 (3.9)	4.8 (2.1)
Less Experienced (G_2)	5.6 (3.1)	4.3 (2.5)	3.9 (2.0)
All participants (G_{1+2})	5.5 (2.9)	4.8 (3.3)	4.4 (2.1)

Table 3. Contribution of layout features (Range: 1-7, Lower = Stronger)

	Title	Snippet	TRS	URL	Size	File Type
System 1 (S_1)	2.0 (1.0)	2.5 (1.0)	3.0 (1.7)	4.5 (2.3)	6.0 (1.6)	5.4 (1.7)
System 2 (S_2)	1.8 (0.9)	2.0 (1.2)	**1.9** (1.0)	4.2 (2.1)	6.0 (1.7)	5.8 (1.6)
System 3 (S_3)	1.6 (0.9)	1.8 (1.0)	**1.6** (0.8)	4.5 (2.1)	6.1 (1.6)	6.0 (1.6)

$N = 24$

Table 2 shows the number of result pages viewed by participants during the tasks. In the experiment, all systems displayed 10 records per result page. However, the ostensive presentation offered in System 2 and 3 allowed participants to access the top 30 records through TRS. Therefore, it was anticipated that the number of result pages which participants viewed to complete the tasks should be reduced. This appeared to be the case, as suggested at the bottom row (All participants) of Table 2. In both subject groups, participants tended to view fewer pages in S_2 and S_3 compared to S_1. However, the Friedman Tests show that the differences among the three systems are not significant. Therefore, while participants' perceptions on the browsing and relevance assessments tended to be more positive when the ostensive browsing and query suggestion were offered in the interface, no conclusive evidence was found for their benefit in the reduction of the number of result pages viewed by participants.

We further investigated the contribution of the individual interface features to participants' decisions of visiting URLs from the search results. The features examined were the title, snippet, TRS, URL, size, and file type of retrieved records. Participants were asked to indicate how strongly each feature contributed to their decision of visiting URLs in the search results. Table 3 shows the result of the analysis. An interesting trend was that while the contribution of the URL, size, and file type tended to remain similar across the systems, TRS' contribution appeared to be increased when the ostensive browsing was available in the interface (i.e., S_2 and S_3). The Friedman Test shows that the difference of TRS is significant across the systems ($\chi^2(2) = 21.031, p = .000$). The post hoc tests shows that the difference between S_1 and S_2 ($Z = -3.157$, $p = .001$) and between S_1 and S_3 ($Z = -3.558$, $p = .000$) are significant. This suggests that participants tended to rely more on TRS to access the URLs from the search results when the ostensive browsing was available, compared to the static presentation in S_1. We also noted that participants gave a more positive score to the title and snippet in S_2 and S_3 compared to S_1. The Spearman's correlation coefficients show that the contributions of TRS and snippet are significantly correlated ($rho = .335$, $p = .004$), and so are the contributions of snippet and title ($rho = .627$, $p = .000$). This suggests that the ostensive browsing had an effect of increasing participants' awareness of the other features of document surrogates during the tasks.

Summary: This section has presented the experimental results regarding the browsing and relevance assessments of the search results. The results show that participants often found the systems with the ostensive presentation easier to browse the search results and identify relevant documents. The results also suggest that the ostensive browsing can lead to an increased level of awareness for the other components of document surrogates.

4.2 Query formulation

Formulating an effective query is often a difficult task for searchers [20]. It has been suggested that a variety of information can be used as the source of a searcher's query re/formulation [21]. In our experiment, the expansion terms were suggested in S_3 based on the interaction with the ostensive browsing of TRS. This section presents the experimental results regarding the query re/formulation.

Table 4 shows participants' subjective assessments on the support of formulating queries offered by the interfaces. Participants were asked to indicate their assessment by the question "*Did the interface increase your ability to formulate relevant queries?*". As can be seen, S_3 were given the most positive score among the three systems in both of the subject groups. The Friedman Tests show that the differences among the three systems are significant in G_1, G_2, and G_{1+2} ($G_1 : \chi^2(2) = 7.171$, $p = .027$. $G_2 : \chi^2(2) = 9.829$, $p = .004$. $G_{1+2} : \chi^2(2) = 16.763$, $p = .000$). The post-hoc tests show that the differences between S_1 and S_2 ($Z = -2.574$, $p = .005$) and between S_1 and S_3 ($Z = -2.257$, $p = .011$) are statistically significant in G_{1+2}. This and the relatively close score between S_2 and S_3 suggest that participants tended to find it easier to formulate queries based not only on the term suggestion function offered in S_3, but also on the overall ostensive browsing that were offered in S_2 and S_3. This also indicates that there is room for improving the way in which suggested terms were presented to participants in S_3. We will elaborate this aspect in Section 5.

Table 5 shows the results of participants' query re/formulation process recorded during the tasks. It presents the number of queries submitted to the interface, unique words used during a task, and average query length. Due to the space limit, it only shows the result in G_{1+2} and no significant difference was found between subject groups. The results show that participants tended to submit a fewer number of queries in S_2 and S_3 compared to S_1. However, the number of unique words and average query length in S_3 appeared to be larger/longer than S_1. The Friedman Tests show that the differences among the three systems are not significant for the number of queries, unique words, nor query length. While the difference was not significant, the task breakdown of the results shows that the number of unique words submitted to S_3 was consistently larger

Table 4. Support of formulating queries (Range: 1-7, Lower = Better)

	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	4.3 (1.8)	3.1 (1.4)	2.7 (2.0)
Less Experienced (G_2)	2.8 (1.5)	2.2 (0.6)	1.9 (1.7)
All participants (G_{1+2})	3.5 (1.8)	**2.6** (1.1)	**2.3** (1.9)

Table 5. Number of queries, unique words, and query length (G_{1+2} only)

	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
Queries	5.0 (3.0)	4.0 (3.0)	3.8 (2.1)
Unique words	7.2 (3.4)	6.2 (2.9)	8.6 (4.5)
Query length	4.4 (1.8)	3.7 (1.0)	4.8 (2.1)

$N = 24$

than S_1 in all three tasks. On the other hand, the number of queries submitted to S_3 was smaller than S_1 in Task 2 and 3, and similar (4.9 in S_1 vs. 5.0 in S_3) in Task 1. Therefore, the query suggestion offered in S_3 appeared to help participants diversify search vocabulary to complete a task without increasing the number of queries.

Summary: This section has presented the results regarding the query re/formulation performed by participants during the tasks. While the effect of the ostensive presentation was not always evident in the system logs, there was some indication that suggested that the effort of manually formulating queries can be reduced when the ostensive browsing was available in the interface. This was also partly supported by participants' subjective assessment on the interface's support to query re/formulation.

4.3 Task perceptions and performance

We have discussed the effects of the ostensive presentation of TRS on the browsing of search results and query re/formulation process. This section investigates how these effects influence participants' perceptions on the tasks they carried out. The overall task performance is also analysed in relation to the perceptions.

Table 6 shows participants' subjective assessments on the search tasks they carried out (G_{1+2} only). In particular, the perceptions on the satisfaction of the task outcomes and on the complexity of tasks were investigated. For the satisfaction, the question "*How satisfied are you with the results of this search?*" was asked and the answer was captured by a 7-point scale as before (i.e., *Very (1)* to *Not at all (7)*). For the complexity, participants were asked to indicate a degree of agreement with the following statement "*The search task we asked you to perform was: Very Complex (7) to Very Simple (1)*". As can be seen, the overall difference between the three systems regarding the satisfaction of task outcomes appeared to be small. However, the standard deviation indicates that the assessments on S_1 is likely to be more consistent across participants compared to S_2 or S_3. On the other hand, participants appeared to find the tasks less complex when S_2 or S_3 were used compared to S_1. The trend was consistent across the subject groups. However, the standard deviation on S_2/S_3 was again higher than S_1. The task

Table 6. Participants' perceptions on the tasks (G_{1+2} only)

	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
Satisfaction of search outcomes	2.8 (1.2)	2.9 (1.7)	2.9 (2.2)
Task complexity	3.7 (1.3)	3.3 (1.7)	3.4 (2.3)

N = 24; Range: 1-7; Lower = Better (Satisfaction); Lower = Simpler (Complexity).

Table 7. Number of bookmarked pages and task completion time

Number of bookmarked pages	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	3.1 (1.2)	2.5 (2.0)	2.5 (1.7)
Less Experienced (G_2)	4.4 (1.7)	4.2 (2.1)	4.6 (2.8)
All participants (G_{1+2})	3.8 (1.6)	3.3 (2.2)	3.5 (2.5)
Time taken to complete the tasks (sec)	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	666 (150)	715 (151)	609 (162)
Less Experienced (G_2)	724 (77)	724 (139)	716 (145)
All participants (G_{1+2})	695 (120)	719 (142)	662 (160)

breakdown of the results show that the perception on the task complexity in S_2 was consistently better than S_1 in all three tasks while the difference was small. S_3 was given a more noticeably positive assessment in Task 2 and 3 compared to S_1, but it was given a noticeably worse assessment on Task 1. However, the Friedman Tests show that the differences among the three systems are not significant in all results. We also looked at an interaction effect between system and task, and no significant effect was found.

Table 7 shows the number of pages bookmarked by participants when perceived relevant information was found. As can be seen, participants in G_1 appeared to bookmark more pages in S_1 than S_2/S_3. In G_2, participants appeared to bookmark a comparable number of pages between S_1 and S_3. However, the Friedman Tests show that the differences among the three systems are not significant. An interesting point was that G_1 tended to bookmark fewer pages than G_2. The Mann-Whitney U Tests show that the difference between the subject groups was significant in all systems. This suggests that More Experienced group tended to complete the tasks with fewer pages than Less Experienced group across the systems. Table 7 also shows the time taken to complete the tasks in seconds. Overall, participants appeared to complete the tasks faster with S_3 compared to S_1 or S_2 in both subject groups. However, the Friedman Tests show that the differences among the three systems are not significant.

Summary: This section has presented the results regarding participants' perceptions on the tasks they carried out, and their overall task performance. In summary, we did not find much evidence which suggested that the ostensive presentation of TRS had an significant effect on participants' perceptions on the search tasks. While the number of pages bookmarked to complete the tasks can be different across the subject groups, no significant difference was found among the systems regarding the overall task performance.

4.4 System preference

At the end of three tasks, participants were asked to indicate the preference of the systems based on the experience of the searches they carried out. The result is shown in Table 8. As can be seen, participants in both subject groups appeared to prefer S_3 most followed by S_2. Friedman Tests show that the differences among the three systems are significant ($G_1 : \chi^2(2) = 13.500$, $p = .000$. $G_2 : \chi^2(2) = 16.667$, $p = .000$. $G_{1+2} : \chi^2(2) = 30.083$, $p = .000$). The post hoc tests show that the difference between S_1 and S_3 is significant in G_1 ($Z = -2.973$, $p = .001$), the differences between all

Table 8. Participants' system preference (Lower = Better)

	System 1 (S_1)	System 2 (S_2)	System 3 (S_3)
More Experienced (G_1)	2.8 (0.6)	2.0 (0.6)	**1.3** (0.5)
Less Experienced (G_2)	2.8 (0.6)	**2.0** (0.0)	**1.2** (0.6)
All participants (G_{1+2})	2.8 (0.6)	**2.0** (0.4)	**1.2** (0.5)

three systems are significant in G_2 ($Z = -2.887$, $p = .003$), and G_{1+2} ($Z \leq -4.090$, $p = .000$). The Mann-Whitney U test show that the differences between the subject groups are not significant in all systems. While we do not exclude a possibility of participants giving a more positive assessment on S_2 and S_3 just because the interfaces were new to them, these results suggest that participants found a case where the functionality provided by S_2 and S_3 was useful during the tasks.

5 Conclusive discussion

This paper presented a sentence-based approach to the ostensive browsing and searching on the web. A user study with 24 participants was carried out to investigate the effectiveness of our approach. The experimental results have the implications on the effects of the ostensive presentation in the searching process. First, the ostensive presentation can facilitate the effective browsing of retrieved documents. Participants often found the system with the ostensive presentation easier to browse the search results and find relevant information, compared to the static presentation of TRS. With the ostensive browsing available in the interface, participants tended to rely more on them to make a decision of which URLs to visit. An interesting effect we found was that the ostensive presentation appeared to increase a level of participants' awareness on other components of document surrogate. Therefore, the interaction design proposed in this work can be an interesting alternative to the existing TRS presentation such as [14].

Second, the ostensive browsing appears to have a positive effect on a user's formulation of effective queries. Participants tended to submit a fewer number of queries to complete the tasks in System 2 and 3 compared to System 1. While the difference was not statistically significant, participants found System 2 and 3 more supportive of their query re/formulation. The close assessment between System 2 and 3 leads us to believe that the active interaction with TRS had a positive effect on participant's query re/formulation process. However, this also suggests that the way in which the suggested terms are presented should be improved. In the current implementation, the query box was updated with the suggested terms when TRS was accessed. Participants sometimes accepted all suggested terms or delete them all to submit a new query. A better control on the selection of suggested terms should be devised for future system.

In conclusion, our study suggests that the sentence-based approach to an ostensive browsing is promising to facilitate an effective exploration of search results, and further investigation should be carried out.

References

1. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science **41**(4) (1990) 288–297
2. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Review **13** (1989) 407–424
3. Kuhlthau, C.C.: Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science **42**(5) (1991) 361–371
4. Pharo, N., Järvelin, K.: "Irrational" searchers and IR-Rational researchers. Journal of the American Society for Information Science and Technology **57**(2) (2006) 222–232
5. Campbell, I., van Rijsbergen, K.: The ostensive model of developing information needs. In: Proceedings of the COLIS 2., Copenhagen, Denmark (1996) 251–268
6. Chalmers, M., Rodden, K., Brodbeck, D.: The order of things: Activity-centred information access. In: Proceedings of the 7th WWW Conference, Brisbane, Australia (1998) 359–367
7. Chi, E.H., Pirolli, P., Chen, K., Pitkow, J.: Using information scent to model user information needs and actions and the web. In: Proceedings of the CHI conference, Seattle, WA, ACM Press (2001) 490–497
8. Das-Neves, F., Fox, E.A., Yu, X.: Connecting topics in document collections with stepping stones and pathways. In: Proceedings of the 14th CIKM, Bremen, Germany, ACM Press (2005) 91–98
9. Campbell, I.: Interactive evaluation of the ostensive model, using a new test-collection of images with multiple relevance assessments. Journal of Information Retrieval **2**(1) (2000) 87–114
10. Urban, J., Jose, J.M., van Rijsbergen, C.J.: An adaptive technique for content-based image retrieval. Multimedia Tools and Applications **31**(1) (2006) 1–28
11. White, R.W., Jose, J.M., Ruthven, I.: A granular approach to web search result presentation. In: Proceedings of the INTERACT 2003, Zürich, Switzerland, IOS Press (2003) 220–227
12. Tombros, A., Sanderson, M.: Advantages of query-biased summaries in information retrieval. In: Proceedings of the 21st SIGIR Conference, Melbourne, Australia, ACM (1998) 2–10
13. White, R.: Implicit Feedback for Interactive Information Retrieval. Phd thesis, Department of Computing Science, University of Glasgow, Glasgow, UK (2004)
14. White, R.W., Jose, J.M., van Rijsbergen, C.J., Ruthven, I.: A simulated study of implicit feedback models. In: Proceedings of the 26th ECIR, Sunderland, UK (2004) 311–326
15. White, R., Jose, J.M., Ruthven, I.: Using top-ranking sentences to facilitate effective information access. Journal of the American Society for Information Science and Technology **56**(10) (2005) 1113–1125
16. Borlund, P.: Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation **56**(1) (2000) 71–90
17. Over, P.: Trec-7 interactive track report. In Voorheer, E.M., Harman, D., eds.: NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7), Gaithersburg, MD, NIST (1998) 33–39
18. Hull, D.: Using statistical testing in the evaluation of retrieval experiment. In: Proceedings of the 16th SIGIR Conference, Pittsburgh, PA, ACM (1993) 329–338
19. Siegel, S., Castellan, J.N.: Nonparametric Statistics for the Behavioral Sciences. 2nd edn. McGraw-Hill (1988)
20. Belkin, N.J.: Helping people find what they don't know. Communications of the ACM **43**(8) (2000) 58–61
21. Spink, A.: Term relevance feedback and query expansion: Relation to design. In: Proceedings of the 17th SIGIR Conference, Berlin, Germany, ACM (1994) 81–90