Do Users Find Looking at Text More Useful than Visual Representations? A Comparison of Three Search Result Interfaces

Hilal Al Maqbali *Falk Scholer* *James A. Thom* *Mingfang Wu*

School of Computer Science and Information Technology
RMIT University
GPO Box 2476, Melbourne 3001
Victoria, Australia

h.almaqbali@student.rmit.edu.au,{falk.scholer,james.thom,mingfang.wu}@rmit.edu.au

Abstract

The organisation, content and presentation of document surrogates has a substantial impact on the effectiveness of web search result interfaces. Most interfaces include textual information, including for example the document title, URL, and a short query-biased summary of the content. Other interfaces include additional browsing features, such as topic clustering, or thumbnails of the web pages. In this study we analyse three search interfaces, and compare the effectiveness of textual information and additional browsing features. Our analysis indicates that most users spend a substantially larger proportion of time looking at text information, and that those interfaces that focus on text-based representations of document content tend to lead to quicker task completion times for named-page finding search tasks.

Keywords Information Retrieval, User Studies Involving Documents, Web Documents, Eye Tracking

1 Introduction

Search engines are a key tool for supporting users in finding information on the world wide web. These information retrieval systems aim to find relevant documents in response to a user query. While the performance of the underlying ranking function – responsible for identifying good answer resources – is clearly of great importance, the organisation, content and presentation of document surrogates in the search results interface can also have a substantial impact on overall search effectiveness.

One recent study [9] found that only 21% of users found relevant results when querying a search engine, and that 75% were disappointed with the results returned. The way users interact with the search result interface may be one factor in the poor user experience.

This paper analyses three search interfaces that make use of different features including text summaries, clustering information, and visual thumbnail images:

C Carrot2 (`http://www.carrot2.org`),

M Middlespot (`http://www.middlespot.com`), and

N Nexplore (`http://www.nexplore.com`)

A preliminary analysis of overall task completion time with the different interfaces was presented in a previous paper [1]. In this paper, we investigate how much time users spent looking at different regions of the screen, in particular comparing the time spent looking at text surrogates of the result pages with time spent looking at more visual representations. Our analysis indicates that most users find text surrogates to be more useful.

The remainder of this paper is organised as follows. Some related work is presented in Section 2; our experiment design, including the different search interfaces, users, and topics used, is described in Section 3; the results of the experiment are analysed in Section 4; and discussion and conclusions are given in Section 5.

2 Related Work

The presentation of search results influences users' assimilation and guides users to look for the information that is relevant to them. In the past, quite a few studies explored visual presentation of search results [3, 10, 11]. A proper visual representation can communicate some kinds of information much more rapidly and effectively than textual representation. However, visualisation of textually represented information is difficult and challenging [6].

The effectiveness of visual representations largely depends on whether the representation is highly coupled with a search task and on the inherent structure of documents to be presented. Joho and Jose [8] investigated how textual and visual forms of information enabled users to more effectively interact with search answer interfaces in undertaking relevance assessments and reformulating queries.

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

Cutrell and Guan [4] found that adding extra contextual information to the document surrogates can improve the effectiveness on search answer interfaces for informational tasks. Hearst and Pedersen [7] is one of many studies that has investigated the effectiveness of clustering search results. Compared with the results of the study described in this paper, where we find the Carrot2 interface that supported clustering to be ineffective when undertaking a navigational task searching for a single correct answer, they found clustering of answers was effective in supporting a user's task that involved finding a set of relevant answer documents.

A previous study by Dziadosz and Chandrasekar [5] had found that the combination of thumbnails and text summary to be more effective for users than either thumbnails or text summaries alone. However, our study suggests that the combination of thumbnails and text is only effective when they are not large, since users mostly look at the text summaries and it is not effective to use too much of the screen real estate on the images of answer pages.

3 Experimental methodology

To investigate the relative attention that users pay to different interface components, we conducted a user study that involved carrying out a series of named-page finding search tasks using a variety of search interfaces.

3.1 User study

Our study was carried out at RMIT University Open Day in August 2009. Subjects participated in the experiment were mostly high school students with an interest in computer science who were visitors to our laboratory. Participants were given a plain language statement outlining the goals of the experiment, the types of tasks to be undertaken, and the data that would be collected. Based on this information, 35 volunteers chose to participate in the experiments. No training was given with the different search interfaces.

Each participant undertook three navigational search tasks (described below), using different search interfaces. Information about visual attention given to the different screen components was collected using a Tobii T60 eye tracker. This non-intrusive device records the gaze position, providing information on fixations and saccades (brief rapid eye movements).

3.2 Search interface features

Our experiment involved users using three different search result interfaces that contained different amounts of surrogate text and visual browse features about answer documents on the result pages.

The three interfaces were selected because they provide a variety of additional novel features, not just a ranked list of text extracts. Carrot2 does not present visual features, however it clusters its search results. In Middlespot, screenshots are presented for the retrieved documents. Nexplore has more visual features such as highlighting of query terms, thumbnails, background colour and highlighting the abstract of the retrieved document when the mouse is moved over it.

Interface Features	C	M	N
Text features	66%	17%	56%
Browse features	19%	75%	7%
Other regions	16%	8%	37%

Table 1: The distribution of interface features.

In this paper, we consider the following areas within each interface page displaying the ranked list of answers:

Surrogate text: Search engines provide surrogates for answer page in the ranked list of answers. This surrogate text may include the URL of the answer, as well as text from the answer web page title, and a synopsis of the answer web page. The surrogate text for the answer documents is in each of the regions marked (1) on the respective answer interfaces: Figure 1 for Carrot2 (accounting for approximately 66% of the screen), Figure 2 for Middlespot (17%), and Figure 3 for Nexplore (56%).

Browse features: The visual browse features for the answer documents are in each of the regions marked (2) on respective answer interfaces. Figure 1 shows the clustering area in Carrot2 which occupies approximately 19% of the screen. Figure 2 shows large images of the answer pages that are displayed in Middlespot and occupying approximately 75% of the screen. Figure 3 shows a much smaller region, approximately 7% of the screen, containing the thumbnails displayed by Nexplore.

Other regions: Each interface also had some other regions, such as banners and the surrounding screen, including a region at the bottom of the screen (not shown in the figures) that contained the topic and some instructions, This accounted for approximately 16% of the screen with the Carrot2 interface, 8% with Middlespot interface, and 37% with Nexplore (since this last interface included a separate area for Wiki Search).

As summarised in Table 1, significant portions of the Carrot2 and Nexplore interfaces are given to surrogate text. The great majority of the Middlespot interface, on the other hand, is occupied by visual browse features.

3.3 Topics

One taxonomic study [2] shows that web search tasks can be classified as informational, transactional or navigational. Navigational tasks are used in our study because we assume that users become more interested in

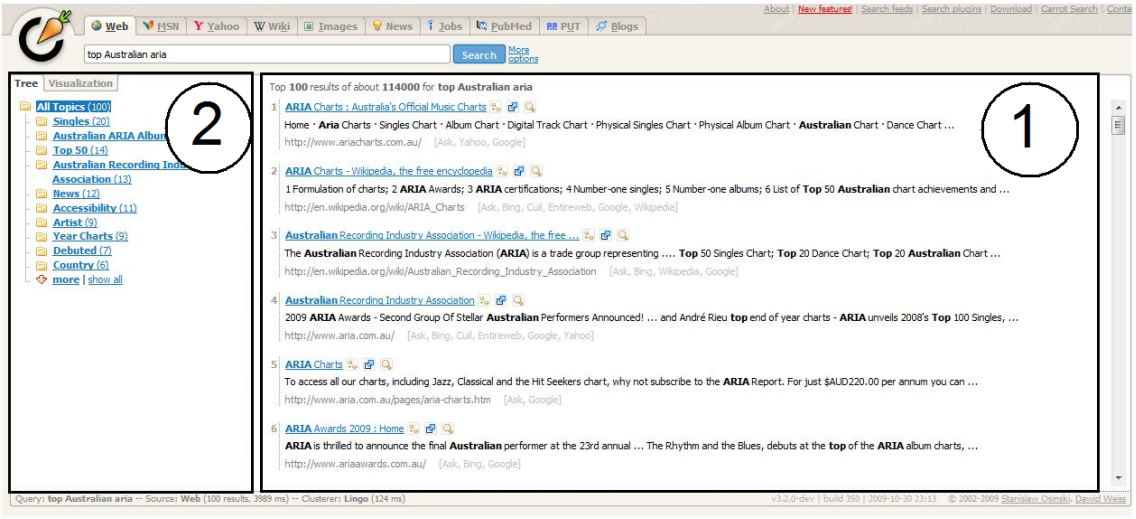

Figure 1: Carrot2 interface (`www.carrot2.org`). Areas marked 1 and 2 indicate Text and Browse features, respectively. Descriptions of the features are provided in the main text.

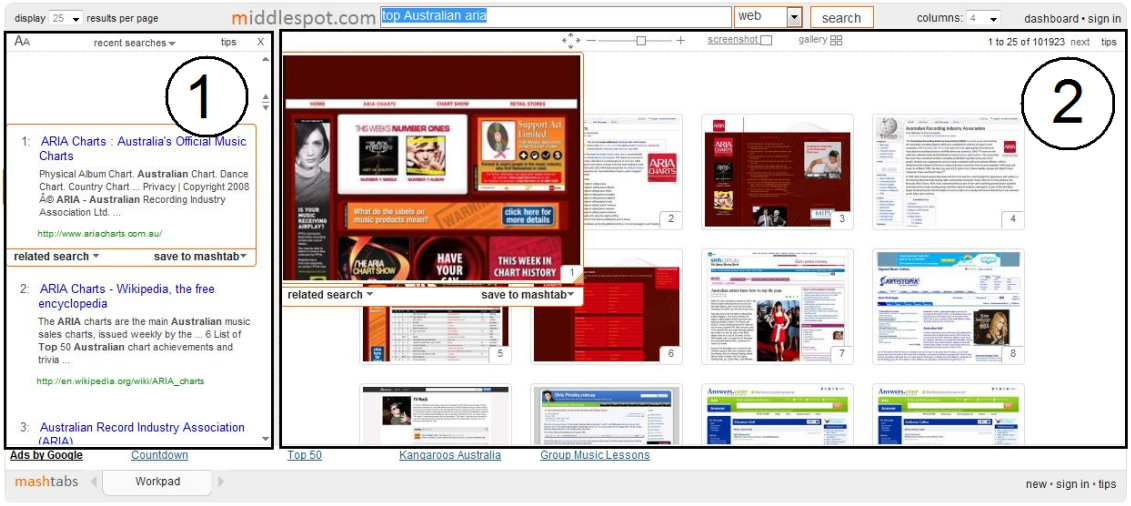

Figure 2: Middlespot interface (`www.middlespot.com`). Descriptions of features are provided in the main text.

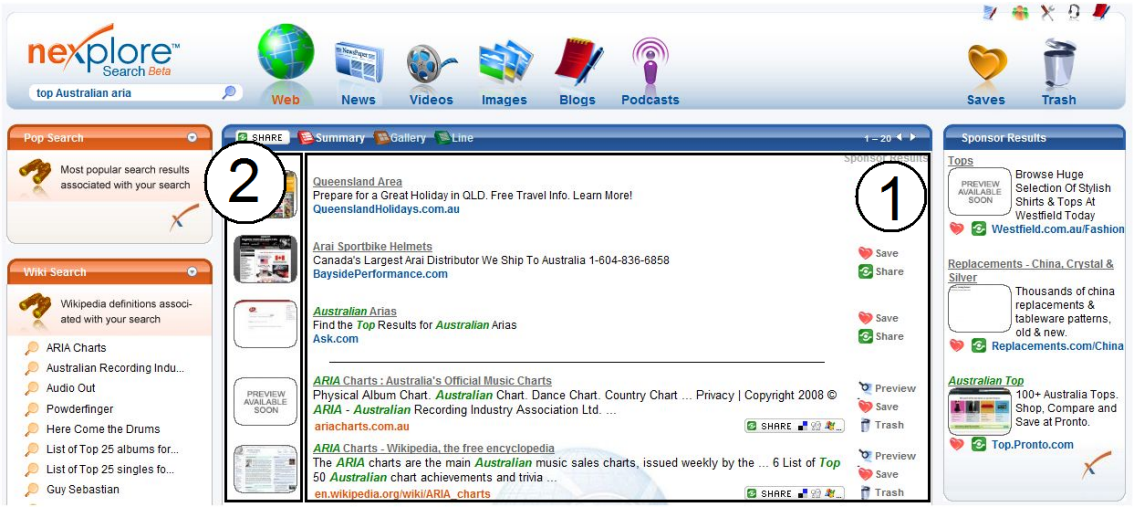

Figure 3: Nexplore interface (`www.nexplore.com`). Descriptions of features are provided in the main text.

Trial	1st task	2nd task	3rd task
1	M- H (4)	C- G (4)	N- A (3)
2	M- G (4)	C- A (4)	N- H (4)
3	M- A (4)	C- H (4)	N- G (3)
4	C- G (3)	N- A (3)	M- H (3)
5	C- A (5)	N- H (5)	M- G (5)
6	C- H (4)	N- G (4)	M- A (3)
7	N- A (3)	M- H (3)	C- G (3)
8	N- H (3)	M- G (3)	C- A (3)
9	N- G (3)	M- A (3)	C- H (3)

Table 2: Experimental design.

using additional web search interface features to get their desired information.

For each interface, users were given a navigational search task, for which they were asked to find a specific single correct answer page for the given topic. The topics were chosen to cover areas that were likely to be of interest to young searchers, and where searchers were unlikely to be hindered due to lack of general knowledge about the domain. The three topics were:

A: Find the ARIA chart of the top 50 music singles in Australia (query terms: `top australia aria`)

G: Find the MSN games website (query term: `msn`)

H: Find the official homepage of the 2009 movie Harry Potter (query terms: `magical potter`)

These topics, and their corresponding answer documents, represent different aspects of navigational searches: the answer for the first topic is a single web page presenting the required (named) information; the second is the hub page for a prime sub-part of the overall MSN website; and, the third is the home page (or index) of an overall website.

After reading a topic, the user would click a "start" button to load the results of issuing the predefined query terms (as indicated above) into one of the three search interfaces. The user could then interact with the search result screen however they wanted to.

We used a latin square experiment design with a block of nine trials varying the order in which topics and interfaces were presented to users, each user was presented with one topic for each interface. Due to some interruptions and other problems, not all combinations were completed exactly the same number of times. Table 2 shows the number of times (in parentheses) each of the different combinations of interface (C, M, N) and topic (A, G, H) were completed as the first, second or third task undertaken by one of the users.

4 Results

We analyse user behaviour when carrying out the three search tasks using the Carrot2, Middlespot and Nexplore interfaces based on the relative attention paid to different interface features, and task completion time.

4.1 Interface features

Different search interface features attract highly variable amounts of user attention. Figure 4 shows the proportions of total viewing time that users spent looking at text, browse and other features for each trial (that is, over all search interfaces and all users). The solid line shows the median time, while the boxes show the 25th to 75th percentiles. Whiskers show the range of the data, with outliers (observations more extreme than 1.5 times the interquartile range). Since the time data is not normally distributed (Shapiro-Wilk, $p < 0.0001$), we analyse multi-level factors using the Kruskal-Wallis test, a non-parametric alternative to ANOVA. Pairswise comparisons are made using the Wilcoxon signed-rank test. The relative times for the different features vary significantly (Kruskal-Wallis, $p < 0.0001$). In particular, users spend significantly more time viewing text features compared to browse features (Wilcoxon, $p < 0.0001$) and other ($p < 0.0001$). The difference in viewing patterns between browse and other is not significant ($p = 0.6504$).

Figure 5 shows the median time (over all search answer interfaces) users spent looking at different regions of the screen, broken down by cases where users identified the correct or incorrect answer document for each search trial. The text region was the area of the screen that users spent most of their time looking at, users found slightly more correct answers if they spent a bit more time in this area; while when users spent more time looking at the visual browse regions these were not effective and could often lead users to the incorrect answers rather than correct answers. Time spent looking at both text and browse regions is significantly different between correct and incorrect answers (Wilcoxon, p = 0.0060 for text regions and p = 0.0303 for browse regions) while the difference is not significant for other areas of the screen (p = 0.7669).

Figure 6 shows the distribution of the proportion of time that users spent viewing different features, split by the three interfaces. For the Carrot2 and Nexplore interfaces, users spent substantially more time viewing the text features. However, for the Middlespot interface, the browse features (in this case, the screenshots of web pages) attracted the greatest proportion of viewing time.

4.2 Task completion time

User task completion performance is evaluated by measuring the time taken to carry out a search task to the user's satisfaction. That is, we measure the time from when the search results screen is displayed to the user, until the time that they indicate that they have found a desired answer (generally, by clicking on the hyperlink in the search results list that they chose as their final answer). This is in contrast to our previous analysis [1], where task completion time was measured by taking the time that the user chose to exit the task (by explicitly pressing F10) as the endpoint. This adds additional variation to the results, since some users spend addi-

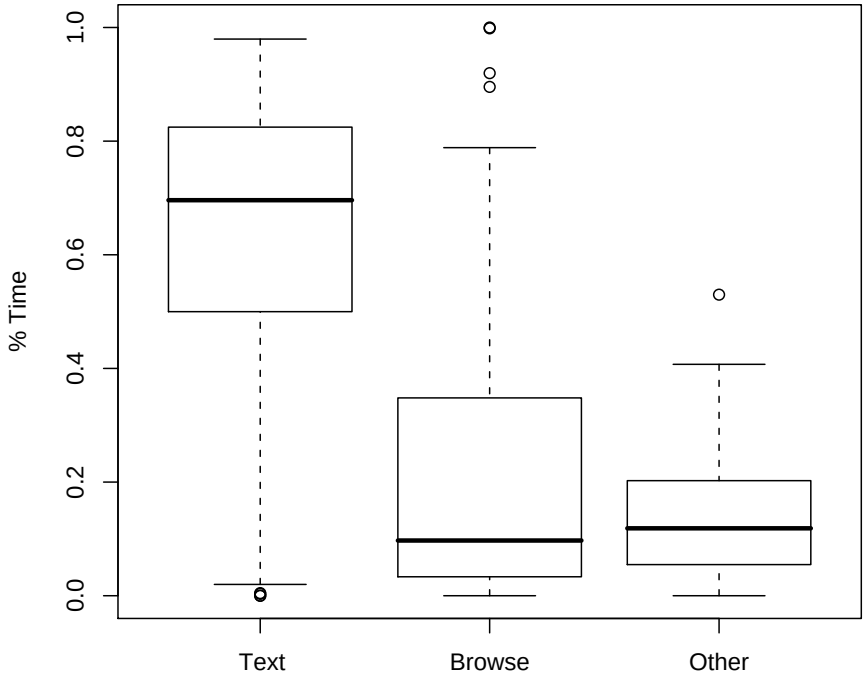

Figure 4: Relative time spent viewing different interface regions.

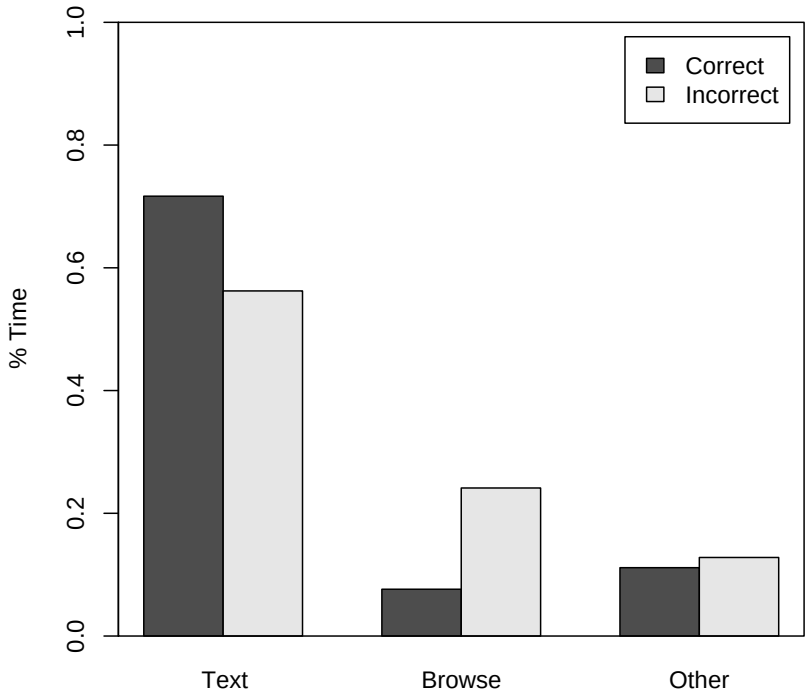

Figure 5: Median proportion of time spent viewing different regions when users found a correct or incorrect answer.

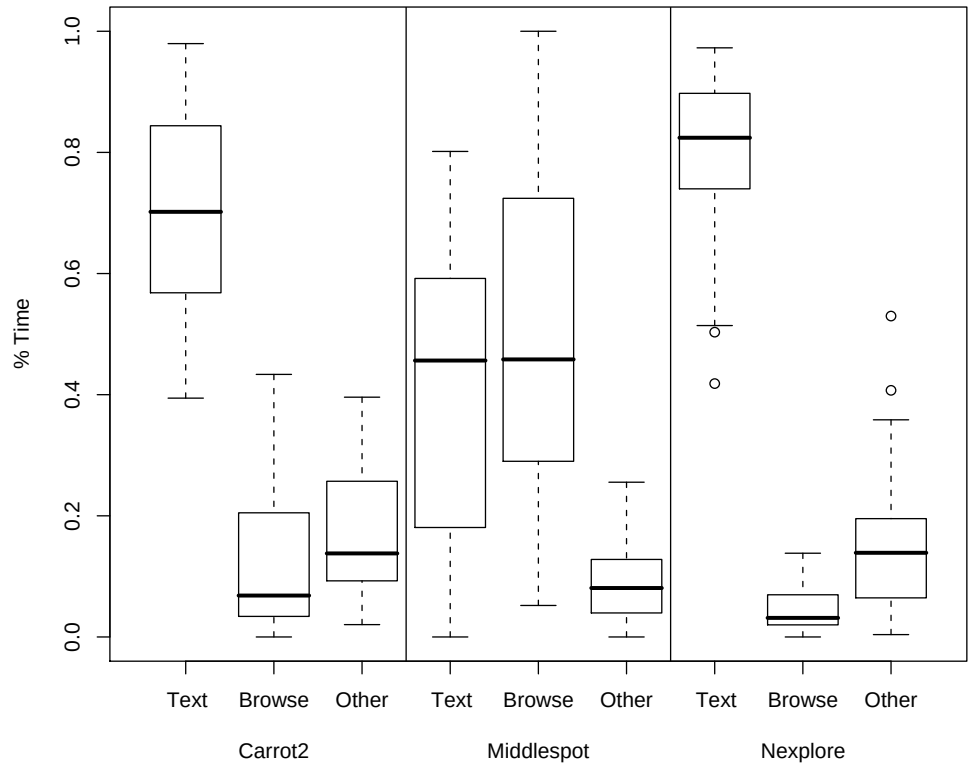

Figure 6: Proportion of time spent viewing different components, by interface.

tional time viewing their chosen answer page, before indicating task completion.

Figure 7 shows the time taken to find an answer, in seconds, for each of the three interfaces. The differences are weakly significant (Kruskal-Wallis, $p = 0.0604$). In particular, the Middlespot and Nexplore differ significantly (Wilcoxon, $p = 0.0129$), while the other pairs do not (Middlespot and Carrot 2, $p = 0.2474$; Nexplore and Carrot2, $p = 0.3225$).

Variation can also be introduced by other sources. The effect of using different search topics was significant (Kruskal-Wallis, $p = 0.0330$). Moreover, because we used real search interfaces and live search results, the rank of the correct answer items in the search results lists of the different interfaces varied somewhat. Although the ranks were similar on average (rank 7.6 for Carrot2, 6.3 for Middlespot, and 6.0 for Nexplore) this did have a significant effect on task completion time (Kruskal-Wallis, $p = 0.0048$). The different users participating in the experiment were not a significant source of variation (Kruskal-Wallis, $p = 0.1227$).

However, this analysis includes all user responses, irrespective of whether the user actually found the correct answer required for the query. We investigate this next.

Answer	Carrot2	Middlespot	Nexplore
Correct	24	18	24
Incorrect	9	14	7

Table 3: Distribution of correct answers by interface.

4.3 Search success

Users were asked to indicate when they felt that they had found the correct answer to the query. However, in many cases users did not in fact identify the correct resource. Table 3 shows the number of incorrect and correct answers found, split by the interface used. The results are strongly indicative of higher success rates with both the Carrot2 and Nexplore interfaces (72.7% and 77.4% of answers are correct, compared to 56.2% for Middlespot). However, the differences are not statistically significant (Fisher, $p = 0.1746$).

We re-analyse the time taken for task completion, using only those trials for which users identified the correct resource in response to the information need. For these responses, the difference between interfaces is greater, and statistically significant (Kruskal-Wallis, $p = 0.0077$). Differences between the interfaces on a pairwise basis are also more pronounced: the median task completion time with Middlespot at 23.71 seconds is significantly longer than that for Carrot2 at 12.81

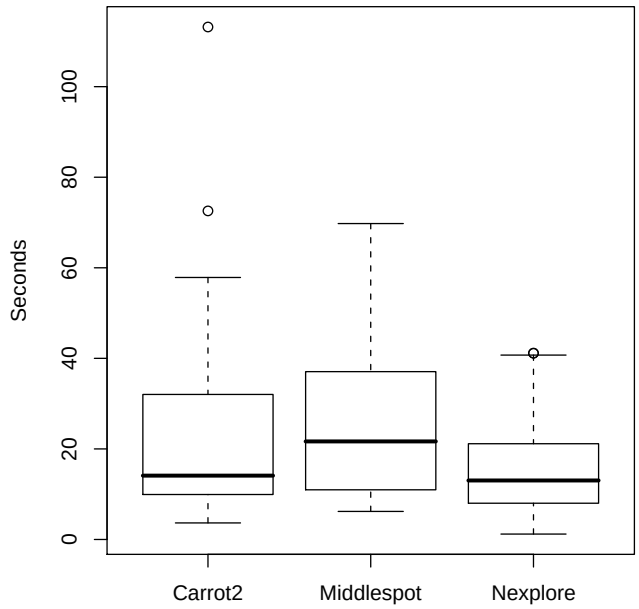

Figure 7: Task completion times by interface.

seconds (Wilcoxon, $p = 0.0112$) and for Nexplore at 12.21 seconds (Wilcoxon, $p = 0.0027$). The difference between Carrot2 and Nexplore is not significant (Wilcoxon, $p = 0.7360$).

Moreover, when considering only those results where users successfully identified correct answers, the effects from topic and user variation are not significant (Kruskal-Wallis, $p = 0.3445$ and 0.2743, respectively). The rank of the answer item only has a weakly significant effect (Kruskal-Wallis, $p = 0.0619$).

5 Discussion and Conclusions

Search result interfaces are an important component of information retrieval systems, and can have substantial impact on overall search task performance. In this paper, we have analysed three publicly available search interfaces, and examined how user attention is split between various features that the search providers make available.

Our analysis has shown that users spend significantly different proportions of time interacting with text, browse and other components of the interfaces. Not surprisingly, these proportions differ between the three interfaces; for Nexplore and Carrot2, text is preferred, while for Middlespot (which presents much less text to the user) browsing features are viewed more.

We have also analysed how task completion time differs between the interfacts, and success rates in identifying correct answers for given informatinon needs. The results show that users spent significantly longer time to interact with the Middlespot interface but found the fewest correct answers. We conclude that, for the navigational search tasks, text features are important in guiding users to finding correct answers quickly.

For the small sample of named-resource finding search tasks, it appears that text information can be vital in supporting users to find the answers that they need. Whether this would also apply to other search tasks, such as informational tasks, will be the subject of future research.

In future work we plan to conduct further user studies over a wider range of tasks. We also plan to investigate the effect of the proportion of screen space that is given over to browsing features as a controlled variable (that is, systematically controlling the proportion).

References

[1] H. Ali [Al Maqbali], F. Scholer, J. A. Thom and M. Wu. User interaction with novel web search interfaces. In *21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group (CHISIG) of the Human Factors and Ergonomics Society of Australia (HFESA)*, 2009.

[2] A. Border. A taxonomy of web search. *ACM SIGIR Forum*, Volume 36, Number 2, pages 3–10, 2002.

[3] S. K. Card, J. D. Mackinlay and B. Shneideman. Morgan Kaufmann Publishers, 1999.

[4] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416, San Jose, California, USA, April–May 2007.

[5] S. Dziadosz and R. Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference*

on Research and development in information retrieval, pages 365–366, Tampere, Finland, August 2002.

[6] M. Hearst. *User Interfaces and Visualisation*. Addison-Wesley, 1999.

[7] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 76–84, Zurich, Switzerland, August 1996.

[8] H. Joho and J. M. Jose. A comparative study of the effectiveness of search result presentation on the web. In *Advances in Information Retrieval, Proceedings of 28th European Conference on IR Research*, pages 302–313, April 2006.

[9] R. S. Rele and A. T. Duchowski. Using eye tracking to evaluate alternative search results interfaces. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, pages 1459–1463, 2005.

[10] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages*, 1996.

[11] B. Shneiderman. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of 2008 ACM SIGMOD International Conference on Management of Data*, pages 3–12, Vancouver, Canada, 2008.