A Task-Based Evaluation of an Aggregated Search Interface

Shanu Sushmita, Hideo Joho, and Mounia Lalmas

Department of Computing Science, University of Glasgow

Abstract. This paper presents a user study that evaluated the effectiveness of an aggregated search interface in the context of non-navigational search tasks. An experimental system was developed to present search results aggregated from multiple information sources, and compared to a conventional tabbed interface. Sixteen participants were recruited to evaluate the performance of the two interfaces. Our results suggest that the aggregated search interface is a promising way of supporting non-navigational search tasks. The quantity and diversity of the retrieved items which participants accessed to complete a task, increased in the aggregated interface. Participants also found the aggregated presentation easier to access to retrieved items and to find relevant information, compared to the conventional interface.

1 Introduction

A recent study reported that 80% of queries submitted to search engines are non-navigational [14]; people are often seeking general information on a broad topic such as "global warming" or "nutrition". Information needs behind such non-navigational queries are often satisfied by relevant information collected from multiple documents in different genres. Due to the increased quantity and diversity of multimedia contents available on the web, images, audio, movies are also becoming relevant to many queries. A conventional way of gathering relevant information from several *information sources* (e.g., web, image, news, wiki) is to browse the search results of individual sources separately available in search engines.

However, a new paradigm of search result presentation has been emerging; aggregated search interfaces. An aggregated search interface is designed to integrate retrieval results from different information sources into a single result page. In this paradigm, users do not have to visit separate pages to browse the search results to access a range of retrieved items. There appears to be two types of integration; blended and non-blended. A blended integration tends to present a single ranked list based on multiple sources, while a non-blended integration tends to present multiple sources in a separate panel in the same page.

Although a log analysis suggested a potential need of aggregated search interfaces [19], there are many unexplored research questions in this paradigm. One such question is the effectiveness of aggregated search interfaces in supporting non-navigation search tasks. In this paper, we present a task-based user

study which compares the performance of an aggregated search interface to a conventional interface.

The outline of the paper is as follows. Section 2 discusses background and related work. Our experimental design is described in Section 3. Section 4 presents the results of our study, and their analysis. Finally, Section 5 discusses our findings and future work.

2 Background and Related work

The search interfaces such as Grouper [5] and Flamenco [3] are some of the conventional ways of organizing retrieved documents or an entire document collection. The former is based on the clustering approach whereas the later follows the faceted browsing approach. Clustering aims to group similar documents together so that users can see multiple aspects from a set of retrieved documents. Whereas, the faceted browsing approach enables users to navigate along the structure of the collection, for example, according to the age, style or school, and creator for an art gallery collection [3]. Users can submit a query but also can browse other items via related facets. Although these approaches are useful for getting multiple aspects of a given query, they are typically single source applications.

Federated search, distributed information retrieval, and metasearch engines are the techniques that aim at providing results from various sources. With the former two, a broker receives the query from the user and selects a relevant sub-set of collections for that query. The top ranked results returned from the selected collections are merged into a single list. Current collection selection methods compare the query with the summary of each collection (term statistics [11] or sample documents [17, 16]) and rank collections accordingly.

A metasearch engine sends a user query to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source. Metasearch engines enable users to enter search criteria once and access several search engines simultaneously. They operate on the premise that the web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining results from several search engines. This also may save the user from having to use multiple search engines separately.

An aggregated search can be seen as an extension of metasearch as it also provides information from different sources. However, the distinction of information sources is more apparent in aggregated search interfaces since the individual information sources retrieve items from very different collections. Yahoo! alpha[1] and Naver[2] are an example of such aggregation approach. These two systems use the non-blended integration where individual sources are presented in a dedicated panel within a single result page, while other search engines adapt a blended integration.

[1] http://au.alpha.yahoo.com/
[2] http://www.naver.com/

In order to support users with a broad query or ambiguous information need, providing diverse information to users has become necessary. More attention is now being paid towards providing diverse results to the users (see e.g. [1, 2]). For example, a study to measure the diversity within image search results can be seen in [4].

Aggregated search also attempts to achieve diversity by presenting results from different information sources (image, video, web, news, etc) on one result page. Here, the aim is to provide diversity across information sources. However, evaluation outcomes regarding the effectiveness and usefulness of aggregated search have been limited in the literature, which we intent to remedy with our work. In this paper, we describe a task-based evaluation of an aggregated search interface.

3 Experimental Design

A within subject experiment design was used in our study, where two search interfaces (controlled and experimental) were tested by sixteen participants, performing two search tasks with each interface.

In the following subsections we define the research hypotheses of this study and discuss the experiment designed to investigate the hypotheses.

3.1 Research Hypotheses

The overall hypothesis of our study is that *an aggregated presentation can facilitate non-navigational search tasks by offering diversified search results*. More specifically, we formulate the following sub-hypotheses to investigate:

H1 An aggregated presentation can increase the quantity and diversity of documents viewed by users to complete a task.
H2 An aggregated presentation can increase the quantity and diversity of relevant information collected by users to complete a task.
H3 An aggregated presentation can improve users' perceptions on the search system.

While an increased number of clicks can be seen as a sign of confusion in navigational queries, informational search tasks often require to view a range of documents to complete the task. Therefore, an effective interface should be able to facilitate the browsing of retrieved documents (**H1**). This should also affect the relevant information collected to complete a task (**H2**). Finally, participants were expected to have a positive perception on the system that enabled them to perform a task successfully (**H3**).

3.2 Search interfaces

Two search interfaces, called DIGEST system, were devised to address our research hypotheses. Both interfaces used the same back-end search engine (Yahoo!

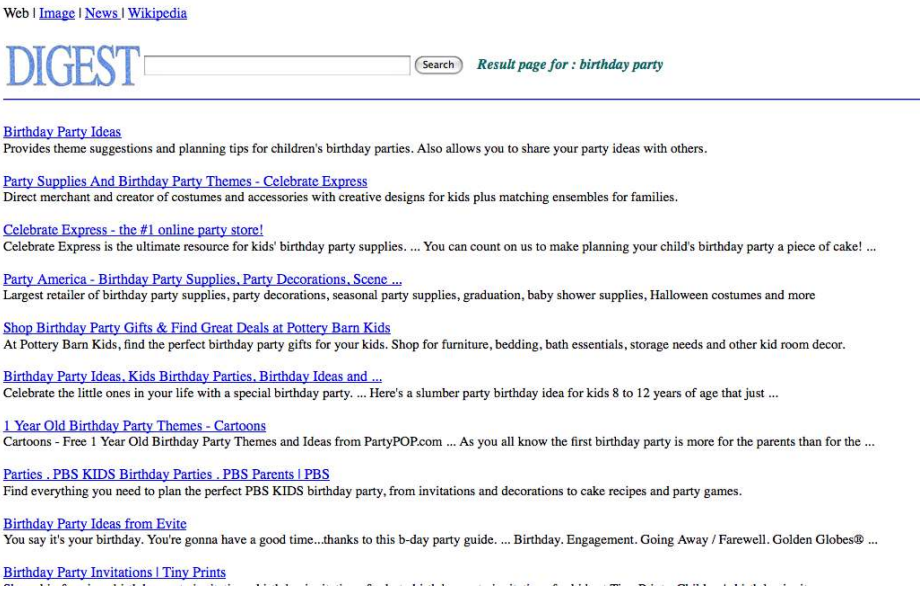


Fig. 1. Controlled System (Tabbed)

search API). For a given query, the API was set to retrieve the top 30 items from four information sources, in this paper, Web, Image, News, and Wiki. The difference between the two interfaces was the presentation of retrieved items.

Figure 1 shows the controlled system where the results from the four sources were presented in a separate tab. The default source was set to the Web tab, and users can click other tabs at the top of the interface to view the results from other sources. This represented a conventional vertical presentation of search results available in major search engines. The controlled system presented the first 10 results for every selected information source with an option of "more results" at the bottom to view the remaining 20 results (in chunks of 10).

Figure 2 presents the experimental system where the results from the four sources were integrated into a single page. This represented an aggregated presentation of search results. The first 10 web results, 12 image results, 10 wiki results and 5 news results, were shown, in each corresponding panel. Every information source on the experimental system also had an option of "more results" (similar to the controlled system) in order to view the remaining results. The layout of the four sources was arbitrarily designed and fixed throughout the experiment. A formal study to determine an optimal layout is left for future work.

3.3 Task

Participants of our user study were asked to perform non-navigational search tasks using the interfaces described above. Each search task was based on the

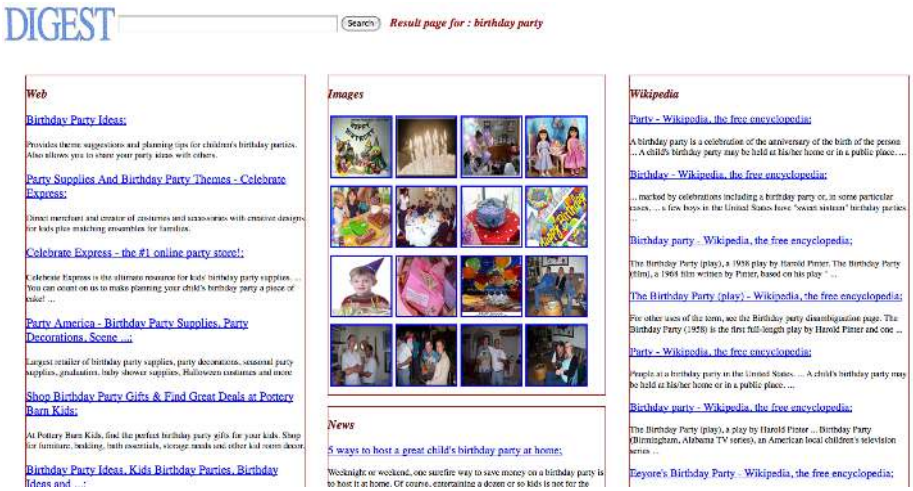


Fig. 2. Experimental System (Aggregated)

simulated work task situation framework proposed by [10]. The framework was designed to encourage participants to engage with an artificial task by giving a situational background scenario of the task. Figure 3 shows an example of the search scenario. As can be seen, our search tasks required to browse several documents and collect relevant information from multiple sources. Participants were asked to copy and paste relevant texts, URLs, and images to a word processing software during the task. We used the software as an electronic notebook. Examples of notebooks made by participants are shown in Figure 4.

We prepared 6 search scenarios so that participants could choose the scenarios based on their interest. This design aimed to facilitate participants engagement with the artificial search tasks. Participants were given 15 minutes to complete each task. Each participant performed four search tasks, two with the experimental system and two with the controlled system. The order of the systems was rotated to reduce learning effects.

3.4 Participants

The experiment was carried out with 11 males and 5 females from our university. Out of 16 participants, 7 were undergraduate students, 2 postgraduate students, 3 PhD students, and 4 were research staff members. The participants were from various educational fields, namely, computing, business management, arts and commerce. The participants were recruited through our call for participation email distributed to several lists. An entry questionnaire established that 82% of participants stated that they had accessed more than one information source to complete a search task. Therefore, our participants were not totally unfamiliar with search tasks that require multiple sources. However, none had used our interfaces or tasks before.

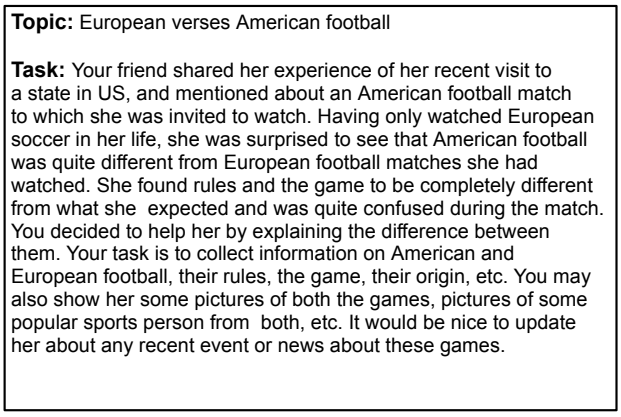


Fig. 3. An example of a simulated task

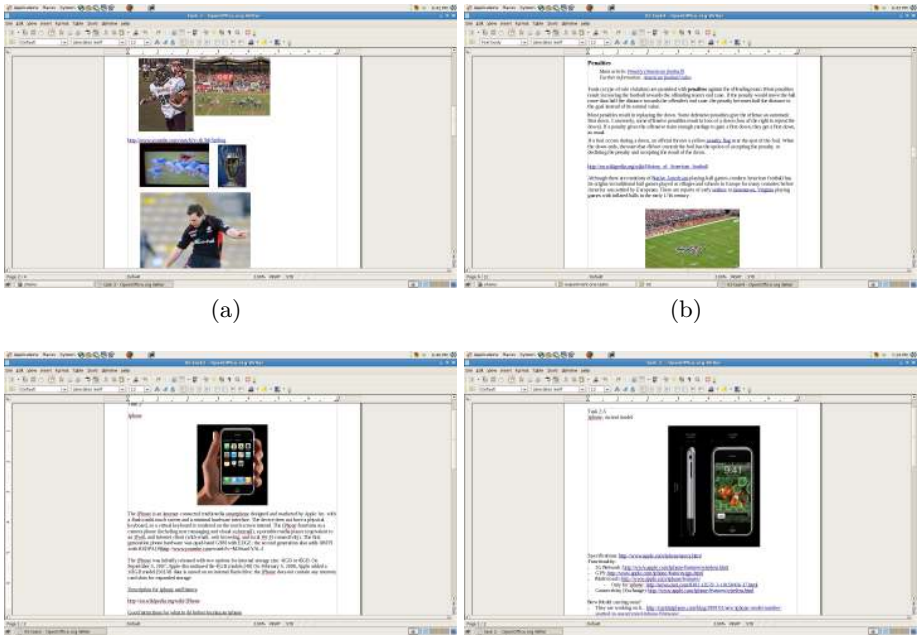
3.5 Procedure

For each participant, the experiment was performed in the following manner. When they arrived at the experiment site, they were welcomed and explained the overall aim of the experiment. When they agreed to participate, a consent form was signed. Then, they were asked to fill in an entry questionnaire to capture their profile and search background. Next, they had a training session with both interfaces using a sample search task. The training session typically lasted for five minutes.

Then, they were asked to perform the first search task by selecting the most interesting scenario from the six scenarios. During the task, the system automatically logged participants' interaction with the interface. When the first task was completed, they were asked to fill in a post-task questionnaire to capture their subjective assessments on the system and task. Then, participants were informed of the change of the interface, and the second scenario was selected. This was repeated four times. After the completion of the four tasks, they were asked to fill in an exit questionnaire to capture their perceptions of systems and tasks as a whole. Participants were rewarded fifteen pounds for their participation after the experiment.

4 Results

This section presents the results of our experiment based on the research hypotheses stated in Section 3.1. We had a total of 32 search sessions per system in the analysis. To measure the statistical significance of the results, we applied both t-test (parametric) and Wilcoxon signed rank test (non-parametric) to the difference between the controlled and experimental systems. All tests were paired and two-sided, and critical value was set to 0.05, unless otherwise stated.



(a) (b)

(c) (d)

Fig. 4. Sample information collected during search tasks by participants

4.1 Quantity and diversity of documents viewed

The first hypothesis **H1** looked at the effect of an aggregated presentation on the quantity and diversity of documents participants viewed to complete a task. To examine this hypothesis, we first analysed participants' click-through data on different information sources. The results are shown in Table 1.

The bottom row of the table shows the average number of retrieved items viewed to complete a task. As can be seen, participants viewed a significantly larger number of items in the experimental system when compared to the controlled system. The breakdown of the information sources suggests that the difference was due to the significantly different frequency in the Wiki and Image sources. These results provide a support for that the aggregated presentation increased the quantity of retrieved items viewed.

We also looked at the combination of information sources accessed by participants to complete a task. The results are shown in Table 2. As can be seen, in five more sessions, participants accessed all four information sources in the experimental system when compared to the controlled system. Also, more sessions were completed by a single source (Web) in the controlled system. This suggests that the aggregated presentation encouraged participants to view more diversified sources from search results. We also noticed that the frequent source was different in the two systems. When we looked at the diversity score 3, the sources

Table 1. Frequency of participants' clicks per information sources (N=32)

Source	Controlled system		Experimental system		T-Test	Wilcoxon-Test
	Mean	SD	Mean	SD	p-value	p-value
Web	7.7	6.7	8.6	5.9	.3696	.1847
News	1.7	1.7	1.0	1.2	.0663	.1039
Wiki	1.0	1.5	2.7	2.2	**.0002**	**.0005**
Image	2.4	3.2	6.8	7.7	**.0013**	**.0004**
All	12.8	8.8	19.1	12.1	**.0002**	**.0024**

Table 2. Combination of information sources, where W=web, I= image, N=news and Wi= wiki

Diversity	Sources	Controlled system	Experimental system
1	W	4	0
2	W+I	2	3
2	W+N	1	2
2	W+Wi	1	2
2	I+Wi	0	1
3	W+I+N	12	1
3	W+N+Wi	3	1
3	W+I+Wi	1	9
4	W+I+N+Wi	8	13
	Total	32	32

of Web, Image, and News was the most popular combination in the controlled system while the Web, Image, and Wiki were the most common combination in the experimental system. We will discuss this aspect later.

Overall, our results provided some evidence to support **H1**.

4.2 Quantity and diversity of relevant information collected

The second hypothesis examined whether or not an aggregated presentation increased the quantity and diversity of relevant information collected by participants to complete a task. To answer this hypothesis, we performed a similar analysis to the previous section but on the number of texts, images, and URLs collected in the notebook. The number of texts was counted based on the number of paragraphs. The results of the analysis are shown in Table 3.

Again, the bottom row of the table shows the average number of collected items to complete a task. As can be seen, participants collected five more items in the experimental system when compared to the controlled system. The difference was found to be significant by the Wilcoxon test. The breakdown of collected items shows that participants tended to collect more items in all three types (Texts, Images, and URLs) when they used the experimental system. However, no difference was found to be significant.

Table 4 shows the combination of the collected items. As can be seen, the number of sessions where all three types were collected (diversity score 3) was

Table 3. Information collection using Controlled & Experimental systems

	Controlled system		Experimental system		T-Test	Wilcoxon-Test
	Mean	SD	Mean	SD	p-value	p-value
Text	7.8	13.2	10.8	21.4	.3657	.3211
Images	3.3	2.8	4.6	3.4	.1140	.0815
URLs	6.1	5.3	7.4	7.1	.1956	.3250
All	17.3	12.7	22.7	18.6	.1173	**.0409**

Table 4. Information collected using Controlled & Experimental systems for text, image and ulr combinations. Here, I=image, T= text and U = url

Diversity	Information Type	Controlled system	Experimental system
1	I	0	2
1	U	2	0
2	I+T	9	6
2	I+U	11	1
2	T+U	0	12
3	I+T+U	10	11
	Total	32	32

similar across the systems. The frequency in the other two diversity scores (diversity score 1 and 2) was also found to be comparable. However, there was some noticeable difference in the combinations. More specifically, the combination of Image and Text (I+U) and combination of Text and URLs (T+U) had a very different frequency across the systems. The cause of this difference is not entirely clear to us. We are currently examining the log files to get further insight into this phenomenon.

To summarise, our results provided partial evidence to support the quantity aspect of **H2**, but no obvious evidence was found to support the diversity aspect of the hypothesis.

4.3 User perceptions

The last hypothesis looked at the effect of the aggregated presentation on participants' perceptions of the systems. To answer this hypothesis, we analysed participants' subjective assessments on the systems, which were captured by a 5-point Likert scale in the exit questionnaire. More specifically, we asked their agreement on the two following statements for each of the two systems.

Q1 The system was useful to complete my search tasks (1 = Strongly agree; 5 = Strongly disagree).

Q2 It was easy to find relevant information with the system (1 = Strongly agree; 5 = Strongly disagree).

Since our hypotheses expected the experimental system to have a better assessment than the controlled system, the statistical tests were applied with

Table 5. Users' perceptions on the systems (N=16).

	Controlled system		Experimental system		T-Test	Wicoxon-Test
	Mean	SD	Mean	SD	p-value	p-value
Q1	2.4	1.1	1.9	1.1	.1311	.1771
Q2	2.4	1.0	1.8	0.9	**.0430**	**.0466**

paired but one-tailed where an alternative was set to be greater. Note that a lower value represented a higher degree of agreement in our analysis. The results are shown in Table 5. As can be seen, participants tended to find the experimental system easier to find relevant information to complete a task. Although participants tended to give a better score on the experimental system regarding the usefulness, the difference was not found to be significant.

We also asked participants which system was easier to access search results in the exit questionnaire. 75% of participants selected the experimental system for the question. Overall, these results provide partial evidence to support **H3**.

5 Discussion and future work

Aggregation is an emerging paradigm of the search result presentation. There are many unexplored questions in this area. In this paper, we performed a task-based user study to compare the effectiveness of an aggregated presentation to a conventional presentation. In particular, we investigated the effect of the aggregated presentation on the quantity and diversity of information objects accessed by users in non-navigational search tasks. This section first discusses the limitation of our study, followed by the implications of our results on the design of aggregated search interfaces.

There are some limitations in our study. First, we used only one back-end search engine to test the effectiveness of the interfaces. Although this made the comparison fair, the implication of our results is limited to this particular engine. Second, we tested the systems with a small number of topics compared to a system-centred evaluation. Other types of tasks such as a decision-making task will also give us a better understanding of the effect of aggregated presentation. Third, the collected items were based on perceived relevance and the quality of collected items was not assessed. Finally, the layout of aggregation was fixed in our experiment. This seems to have an implication on participants' information seeking behaviour, which will be discussed next.

Beaulieu [9] observed the trade-off between the complexity of search interfaces and cognitive load of the users. This applies to the design of aggregated search interfaces, too. Our experimental system used a more complex presentation than the controlled system to integrate multiple information sources in a single page. Therefore, the aggregated interface could increase the cognitive load of the end-users. However, our experimental results suggested that participants were capable of interacting with an aggregated presentation, and tended to find the experimental system easier to find relevant information when compared to

the controlled system. This might be due to the fact that the controlled system still required extra effort to select information sources to access a range of retrieved items.

Another implication was that the layout of aggregation was likely to affect people's selection of information sources. In Section 4.1, we found that the combination of the Web, Image, and News was the most common selection in the controlled system while the Web, Image, and Wiki were the popular selection in the experimental system. They were exactly the same order of the sources in the interfaces. The tab on the top of the controlled interface listed the sources in the order of Web, Image, News, and Wiki. The top three panels of the aggregated interface were the Web, Image, and Wiki. This suggests that people's browsing of information sources can be sequential, and their attention moves horizontally rather than scrolling down the result page vertically. This also implies that an aggregated search interface might be able to offer an effective support by optimising the order of information sources for different tasks or queries.

The last point leads us to formulate our future work which will be addressing research questions such as "Is there an optimal combination and order of information sources?", "How can we model the optimal combination and order of information sources for a given query or task?", "Is the effect of layout strong enough to affect task performance?"

In conclusion, our study provided empirical evidence to support that an aggregated presentation of information sources can increase the quantity and diversify of the retrieved items accessed to complete non-navigational search tasks. Participants tended to find the aggregated presentation easier to access retrieved items and to find relevant information. Although these positive effects were not strong enough to increase the number of relevant information collected, we speculate that an intelligent way of organising information sources is a key to achieve such a goal.

Acknowledgements This work was carried out in the context of research partly funded by a Yahoo! Research Alliance Gift and the MIAUCE project (Ref: IST-033715). Mounia Lalmas is currently funded by Microsoft Research/Royal Academy of Engineering.

References

1. F. Radlinski and S. Dumais. Improving personalized web search using result diversification. *ACM SIGIR conference on Research and development in information retrieval*, pages 691–692, 2006.
2. M. Coyle and B. Smyth. On the importance of being diverse: analysing similarity and diversity in web search. *Source Intelligent Information Processing II*, pages 341–350, 2004.
3. K. P. Yee, K. Swearingen, K. Li and M. Hearst. Faceted metadata for image search and browsing. *SIGCHI conference on Human factors in computing systems*, pages 401–408, 2003.

4. M. Sanderson, J. Tang, T. Arni and P. Clough. What Else Is There? Search Diversity Examined. *European Conference on Information Retrieval (ECIR)*, pages 562–569, 2009
5. O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to Web search results. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 1999.
6. T. Heimonen, A. Aula, H. Hutchinson and L. Granka. Comparing the User Experience of Search User Interface Designs. *CHI 2008 Workshop on User Experience Evaluation Methods in Product Development*, 2008.
7. Wilson, M.L. Wilson, M.C. Schraefel and R.W. White. Evaluating Advanced Search Interfaces using Established Information-Seeking Models *Journal of the American Society for Information Science and Technology*, 2009.
8. O. Hoeber, X.D. Yang. User-Oriented Evaluation Methods for Interactive Web Search Interfaces. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 239–243, 2007.
9. M. Beaulieu. Experiments with interfaces to support query expansion. *Journal of Documentation*, 53(1):8–19, 1997.
10. P. Borlund. Experimental components for the evaluations of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.
11. J. Callan. Distributed Information Retrieval, Chapter 5 of *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.
12. F. Diaz. Integration of news content into web results. *ACM International Conference on Web Search and Data Mining*, pages 182–191, 2009
13. S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. *SIGCHI Conference on Human Factors in Computing Systems*, pages 277–284, 2001.
14. B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, 2008.
15. Y. Kural, S. Robertson, and S. Jones. Deciphering cluster representations. *Inf. Process. Manage.*, 37(4):593–601, 2001.
16. M. Shokouhi, M. Baillie, and L. Azzopardi. Updating collection representations for federated search. *ACM SIGIR conference on Research and development in information retrieval*, pages 511–518, 2007.
17. L. Si and J. Callan. Relevant document distribution estimation method for resource selection. *ACM SIGIR conference on Research and development in informaion retrieval*, pages 298–305, 2003.
18. L. Si, J. Lu, and J. Callan. Distributed information retrieval with skewed database size distributions. *dg.o '03: Annual national conference on Digital government research*, pages 1–6. Digital Government Society of North America, 2003.
19. S. Sushmita, H. Joho, M. Lalmas and J. M. Jose. Understanding domain "relevance" in web search. *WWW 2009 Workshop on Web Search Result Summarization and Presentation*, 2009.