

Title: Mediated Web information retrieval for a complex searching task

Corresponding Author: Hyuk-Jin Lee, Ph. D., Assistant Professor

Affiliation & Mailing address: Texas Woman's University, School of Library and Information Studies, P.O. Box 425438, Denton, TX 76204-5438, 940-898-2187 (phone), 940-898-2611 (Fax)

Email: hlee@mail.twu.edu

Second (Co) Author: Gheorghe Muresan, Ph. D.

Affiliation & Mailing address: Live Search, Microsoft Corp, One Redmond Way, Redmond, WA 98052

Email: gmuresan@acm.org

Abstract: The goal of this study is to understand whether providing a search intermediary familiar with a problem domain and its topical structure would support a user's Web searching tasks, especially complicated tasks with multifaceted topics, and whether the order of searching tasks or system usage influences their successful completion. This study investigates the effect of two factors, the interaction mode and the display layout on the three main measures of the user's Web searching behaviors: effectiveness, efficiency, and usability. Two interaction modes are compared, mediation via a domain-specific document collection vs. non-mediated search, and two display layouts, a combination of browsing-supporting hierarchic display and ranked list of results vs. the simple linear list of search results. The results are analyzed in the Flow theory point of view; they were analyzed by order of the tasks and system usage order. The findings of this study contribute to a better understanding of how the mediation system and / or the combined display support a Web information user.

1. Introduction

The general problem in information retrieval (IR) is that users do not know what they need (Taylor, 1968), they do not know how to formulate their need into a query to the IR system (Belkin, 1982a), and they do not know how to reformulate their need into query after viewing the

results list (Spink, 1995). This situation frequently occurs when users use a large unorganized database such as the World Wide Web. Many findings in Web searching revealed the low query reformulation rate during query sessions (Spink, 2002a; 2002b). On the other hand, Spink and Jansen (2004) explained that despite the generally short nature of user Web queries and search sessions, their studies are also showing that some users are engaging in more complex Web search interactions. Their analyses of Web search logs revealed that some users are conducting more successive or multitasking searches over time and that Web search engines are being used to locate a broadening array of information. Jansen and Spink (2006) also indicated that findings from a study focusing on one Web search engine cannot be applied wholesale to all Web search engines. This showed the necessity of studies focusing on a special Web interface or system for a specific type of Web search task and a specific information domain. Therefore, better query construction with appropriate interaction in the Web searching is necessary, especially when the searching task is complicated.

Solutions to the problem center on the intermediary before electronic IR (i.e., the librarian is the intermediary), which since the electronic search and the Internet has centered on developing an IR system that performs the task of the intermediary, providing mediation between user and the document set in the database. Mediation means the interaction between a user and a system to support the user's exploration of information resources to solve his/her information need. Ingwersen (2005) explained that a group of IR expert system research focused on intermediary design connected to operational exact match IR systems, and by means of various knowledge-based techniques, many system prototypes attempted to cope with rather complex IR situations. In this article we focus on topic modeling (Muresan & Harper, 2004) which approximates the pre-electronic era library intermediary, and manipulating or organizing the results list to provide mediation for the search. There are many automated intermediary techniques applied to IR system; however, success has been more likely when the scope is limited

(Meadow et al., 2000). Interactive intermediary functions can be helpful in: selection of a database, (re) formulation of a query, interpretation of results, and help in understanding the retrieval system. As Meadow et al. (2000) indicated the heart of most automated intermediaries is formulation and editing of a query. Basically there are two kinds of approaches for assisting a user in clarifying his/her information needs for an improved query. There is the intermediary system that does not apply any previous knowledge of a specific user. On the other hand, several systems utilized the accumulated knowledge from previous users' information retrieval performances. In respect of relevance judgment, query reformulation is divided into two main streams: relevance feedback (RF) and automatic query reformulation (automatic QE). The major difference between the two is whether or not 'relevance judgment' is involved. While RF is the process where users identify relevant documents in an initial list of retrieved documents and the system creates a new query based on those sample relevant documents, automatic QE deals with the problem caused by vocabulary mismatch in IR systems. Related studies are introduced in the following subsection 1.1.

Also, intermediary system should help the user in the disambiguation process, for instance, by identifying, highlighting and suggesting different aspects of a topic or different topics represented in a set of retrieved documents. Thus, evaluating the Web retrieval results is another important related topic as a user tries out a query and then, after looking at the search results, reformulates query in light of evidence provided in the result list. There are several ways to manipulate the results list. Examples are clustering, classification, network display, or map display. We focus here on clustering in this study; clustering is a form of mediation to the user when he/she evaluates the results lists, allowing him/her to reformulate the query in a more effective, efficient, and usable manner by offering the similar relevant documents in a cluster. Few studies applied the structured display of search results as an "intermediary" to improve IR results in a large unorganized collection. Related studies are introduced in the following subsection 1.2.

We analyzed the results by Flow theory approach developed by Csikszentmihalyi (1990), which has been often applied to system use and design. Flow is defined as a state in which people are fully involved in an activity. The theory assumes that when a person is engaging in a new task, his/her skill level will be low and the challenge needed to engage the person will also be low. However, as his/her skill level and the level of challenge increases, he/she will improve one's activity by maintaining a sense of flow. As an example, Haynes et al. (1992) found a novice searching group to improve their search performance to the level of experienced searchers by their fourth on-line search when they use a medical database such as MEDLINE. The results of this study were, thus, analyzed by Task order (total 4 tasks) to see whether the order of the given tasks had any influence on the performances of the experiments, and were also analyzed by System order to see whether the order of a new and a default system influenced the performances. Detailed explanations are in section 6.1, Data analysis.

Therefore, the main issue this study investigates is how an automated intermediary could support the user's IR processes, especially for the searcher unfamiliar with the search domain trying to explore a large document collection in order to satisfy a multifaceted information need, the complex searching task that is concerned with more than one related topic. For this purpose, this study is evaluating the effectiveness of machine assistance by: 1) domain-specific mediation support, and 2) clustering display of search results. There has been no study investigating the possibility of combination of these two concepts, especially for information user behaviors in the Flow theory view.

1.1. Constructing the query

This study focuses on the idea of a simulation of domain knowledge mediation by the intermediary searcher. Various studies showed that searches conducted with intermediary support have a far higher success rate than the unmediated ones (Nordlie, 1996). Belkin et al. (1982a, 1982b) emphasized the intermediary's search function in improving a query, which becomes a surrogate for the original information need, by knowledge in the domain area of the query. Spink

(2002a, 2002b) also showed that the pre-search interview conducted by a human intermediary aided the users in defining their information problem more precisely. Spink (1995) explained that the value of professional search intermediaries lies not only in generating search terms from their domain knowledge, but also in identifying search terms from the user's initial question statement, facilitating the user's identification of additional search terms from user's domain knowledge during the verbal interaction, and identifying effective search terms during term relevance feedback (p. 170). In their study, Roussinov and Chen (2001) designed a mediation approach for the Web called "Adaptive Search," which is based on a novel use of clustering, summarization and user feedback. Adaptive search allows users to find information in the significant portion of the Web, and it acts as a layer between a user and a traditional query based on a search engine. Muresan and Harper (2004) suggested search mediation, in which the system builds a model of the user's interest based on observing the user's information-seeking interaction with a document sample that is representative enough and covers all the sub-domains, topics and concepts of the searcher's problem domain; the user model is subsequently used by the system to suggest good queries or to search the target collection on the user's behalf. Muresan and Harper (2002, 2004) showed in user simulations that mediated access through exploring a small, well-structured document collection could significantly improve retrieval effectiveness. This study has been directly inspired by these previous approaches.

Among many intermediary techniques, Relevance Feedback (RF) is the most popular query reformulation strategy (Efthimiadis, 1996). However, which relevant term suggestion technique among many RF techniques is constantly superior in any Web search circumstances is still not clear and is not agreed. Recent recognizable work is the study by Huang et al. (2003), which presented a novel query log-based approach for performing relevant term extraction and term suggestion, and for providing organized relevant terms and contextual information in users' query sessions. However, as Jansen et al. (1998) pointed out, the RF function was not used much by the users in the Web environment, compared to the one in an IR system with professional

assistance. Huang et al (2003) also acknowledged the limits of their approach: the number of extracted relevant terms from the log-based approach (11.8) was obviously less than that for the document-based method (89.5), and the proposed approach required a sufficiently large log. Mayr et al. (2008)'s recent study introduced a search term suggestion method in Web search, which is an aid for query reformulation and reconstruction that was adapted from human search intermediaries to the automatic version for a web-based information portal. However, this study has not been implemented yet. They also pointed out that the result sets of transformed or expanded queries in distributed collections were often very large and tests showed that the conventional web-based ranking methods were not appropriate for the to the user (Mayr et al., 2008). So far, no study for term suggestions specifically dealt with the multifaceted information search task, which requires not only a term suggestion supporter but also the domain knowledge that is required for generating the terms.

In this study, a particular term suggestion technique is not adopted because it would inhibit measuring the major axes of the study, the domain-specific mediation support and the display of the retrieval results, even though the application of such techniques could give benefit to a searcher in terms of query formulation support. Therefore, a research on combining such term suggestion techniques with the mediated system or the new display of the retrieval results could be a promising future research agenda.

1.2. Evaluating search results

Nordlie (1996) explained that, in mediated searches in a library, ambiguities were resolved and the users' information needs were determined not primarily through extensive direct elicitation by the intermediary, but while interacting with the material on the bookshelf. This means that the structure of the information sources is critical for an information user, at least as much as the support by an intermediary. Even though there were some successful cases on commercial Internet sites such as Google and Yahoo!, there have been several attempts to introduce some promising visualization schemes for the Web documents such as hierarchical

display, network display, or map display; most of such visualization schemes were focused on the domain experts rather than the novice IR users (Cole, Mandelblatt & Stevenson, 2002).

Clustering was adopted as a main method of structuring document collections for this study and it derived from automatic overviews, usually created by unsupervised clustering techniques on the text of documents, which attempt to extract overall characterizing themes from document collections. Clustering relies on the cluster hypothesis, which states that relevant documents tend to be more similar to each other than to non-relevant ones, and therefore tend to appear in the same cluster (Jardine & Van Rijsbergen, 1971). The implication is that a user might save time by looking at the contents of the cluster with the highest proportion of relevant documents, and also by avoiding those clusters with mainly non-relevant documents. There are some Web search engines using clustering interface. As an example, Grokker provides the clustered search results by grouping the results topically, and presents them in an interactive visual map. AquaBrowser visualizes concept graphs; this system specializes in showing large trees and graphs, which can be seen as light-weight taxonomies, and is mainly used to facilitate navigation of large structures.

Research results by Wu and colleagues (2001) based on 'aspectual' recall showed that the clustering approach was more effective than the linear ranked one. The linear ranking list means the traditional ranked list based on best-match searching. Users preferred a cluster-based interface to a list interface for the interactive aspectual retrieval tasks. The study showed that the assessment of the structured organizations based on the subjective judgment of the experiment subjects suggests that the structured organization can be more effective; however, assessment based on objective judgments showed mixed results. This study, therefore, showed that although most subjects preferred the cluster structure, the overall performance of using two structures was similar. Other studies (Hearst & Pedersen, 1996; Zamir & Etzioni, 1998) also showed mixed results; for example, an offline experiment (Hearst & Pedersen, 1996) showed that precision and recall were higher within the best cluster but not within the retrieval results as a whole. Many

scholars stated that the reason for the ineffective performance of the structured display of search results in comparison to the list display lies in the ineffective representation of the structured data (Shneiderman, 1998; Hearst & Pederson, 1996; Kural, Robertson, & Jones (2001). HuddleSearch (Osdin, Ounis & White, 2002) enabled the user to find the relevant documents by means of navigating within the concept of a cluster hierarchy by investigating the use of query-biased summarization, created on-the-fly at retrieval time. However, even though there have been many studies about what type of cluster or label is best for information retrieval, no final conclusions have been reached yet (Geraci et al., 2007; Leuski, 2001).

Unlike the previous attempts focusing on developing the power of a ranked list or of a clustering (or classification) approach¹ as a single display scheme, Leuski and Allan (1999) presented an information organization system that combines the two approaches for visualizing the retrieval results: the ranked list and the spring-embedding visualization of clustering. They argued that the search would show an average 20% more effectiveness in the clustering display than following the ranked list. Leuski and Allan (2002) extended their previous study and developed the Lighthouse system, an interactive IR system that mixes a ranked list, document clustering, and a visualization of inter-document similarities. However, as mentioned above, there was no study that investigated the possibility of combining the combined display and the intermediary concept for a specific type of Web task such as a multifaceted topic task.

2. Research framework and questions

The main purpose of this study is to investigate whether simulating a human intermediary, by providing specific domain knowledge and structure, would support a user's search. To simulate the intermediary, we applied Muresan and Harper (2004)'s source collection concept. A source collection for mediation is defined as a thematically focused and organized small

¹ Clustering attempts to group documents that are internally similar to each other and classification is grouping documents based on their similarity to some external set of criteria (Wu et al., 2001).

collection of a specific domain for aiding a user in obtaining appropriate knowledge about that particular domain with respect to his specific information need. The user's exploration of a small well-structured document collection covering a specific subject domain with multiple access points is expected to help the user better grasp the terminology and topical structure of the problem domain as well as clarify and refine his information needs before jumping to a huge target collection such as the Web. By comparing a baseline (non-mediated) system with an experimental (mediated) system, we can test the effectiveness, efficiency, and the usability that such mediation support provides for searching a large unorganized collection. In addition, this study investigates the effect of two different displays of search results. One is the traditional linear ranked list display, and the other is a combination of the linear ranked list supplemented by a topically structured display of search results. Structuring the search results is functionally equivalent to the intermediary's understanding of the structure of the domain. Thus, the combination of two independent variables (mediated vs. non-mediated search; linear vs. combined display) provides four conditions of the experiment in Table 1.

TABLE 1. Framework of the study.

Mediation Display	<i>Non-mediated</i>	<i>Mediated</i>
<i>Linear ranked list of search results</i>	NL: Linear ranked list (Target collection)	ML: Linear ranked list (Source and Target collection)
<i>Combination of Linear ranked and structured display of search results</i>	NC: Combination of linear and structured lists (Target collection)	MC: Combination of linear and structured lists (Source and Target collection)

- NL- Non-mediated and Linear, NC- Non-mediated and Combined
- ML- Mediated and Linear, MC- Mediated and Combined

Non-mediated IR means direct access to the Web, the target collection. Mediated IR means access to the target collection after examining the source collection. The combination of linear ranked list and a classified display is selected as an experimental display based on the previous researches that showed the promising results from this method (Anick & Vaithyanathan,

1997; Leuski & Allan, 1999, 2000). All the results are analyzed based on the Flow theory. We investigated if the order of tasks or the order of the system usage (Non-mediated and Mediated) has an effect on the results.

There are two research questions drawn from the general problem statement and the previous studies: Research Question 1: From the ‘mediation’ point of view, the study aims to find out whether there is a difference in the IR performance between the non-mediated and the mediated IR conditions; Research Question 2: The experiment investigates whether there is a difference in the IR performance between a linear ranked list of search results, and a combination of a linear ranked list and a structured display of search results.

3. Hypotheses

Two hypotheses were developed based on the research questions. The hypotheses predict that the experimental modes (the mediated IR and the combination of displays) would support information users engaging in a large unorganized online collection:

- H1. The mediated IR system is better than the non-mediated IR system for IR performance for a multifaceted topic task in a large unorganized collection².
- H2. The combination of a linear ranked and a clustered display of search results is better than the traditional linear ranked list for IR performance for a multifaceted topic task in a large unorganized collection³.

We investigated the two hypotheses in a multitude of ways, evaluating search performance and estimating success based on different measures: (a) search effectiveness; (b) search efficiency; and (c) usability. Moreover, these measures were evaluated in two distinct ways:

² As explained in the first hypotheses, from all the measurements we expected that mediation is better than non-mediation. However, we expected that mediation would have a negative effect on user’s evaluation of subjective usability, such as ease per system and user effort because mediation may require more cognitive effort from the user.

³ Similar to the H1, we expected that the combined display would have negative effect on user’s evaluation of subjective usability, such as user effort.

(i) based on objective measures such as relevance judgments, time to complete a task, number of saved documents/aspects, and number of queries; and (ii) based on the subjects' subjective/perceived success, preference/usefulness on the system, and ease in completing a task. Finally, we investigated if the order of tasks or the order of the system usage has an effect on the experiment results.

For assessing objective effectiveness, the aspectual recall was selected, and the user's perception of task performance was the measure of subjective effectiveness. Objective efficiency was measured by the user's actual searching time and subjective efficiency was gauged by the user's judgment on the time spent compared to his ordinary Web searching. Subjective usability was gauged by the degree of ease per topic, such as user's evaluation of topic easiness, the degree of ease per system, system usefulness, user effort to utilize a system, and preference of a system. Finally, we adopted objective usability to measure how a user uses a system for his searching task. This was measured by the number of the opened/saved/unsaved documents, and the number of changed queries and query terms.

We expected that a user would gain clearer knowledge of what he needed through browsing a relatively small structured source collection, and that this clarity would lead him to develop improved queries that would yield better search performance in many aspects compared to a user who does not use a mediated IR system. We also expected that the combined display would give the user a better understanding of the results of the target collection.

4. Experiment setting

The adoption of the problem domain and of the experimental topics (tasks) was based on two factors. First, we wanted to satisfy the conditions in which mediation is expected to help; some domain knowledge is needed in order to specify better queries, but the domain should be of interest to the general public, to warrant searches. Secondly, we needed access to an appropriate source collection in support of mediation.

The New Jersey Environment Digital Library (NJEDL) collection was adopted as a source collection. The NJEDL is an adequate source collection: it has approximately 1,300 documents, a specialized domain in environmental information of New Jersey, and a specialized level of contents. While only titles and abstracts were used for indexing, searching, and clustering the source collection, during the experiment the subjects were able to examine the full documents as reference materials. As a target collection, the World Wide Web was used due to the fact that it is a large, varied, and indisputably, the most frequently searched information target. A popular search engine, Yahoo!, was selected as a main search engine for the target collection (the Web).

There are the two alternatives for structuring the source collection, clustering and classification. After comparing both methods in four aspects: system applicability, class representative (label), independence, and availability, we selected the clustering for the following reasons. First, there is no difference between classification and clustering in their applicability for our system. Regardless of detailed system changes, there will be no difference in terms of the roles in the system between the two structuring methods. Second, in terms of labeling, a manually constructed classification has an advantage because the class labels may make more sense to the users than labels generated automatically via clustering. Third, while classification is very dependent on a collection's domain, clustering is independent of such conditions. Automatic clustering technique may contribute to constructing the mediated system more efficiently because it saves time and labor of manually classifying each source collection. Besides, unlike classification, the clusters are adapted to changeable information needs in a mediated IR environment. Finally, as the most important realistic point, there is rarely a well-structured collection with a fit classification scheme due to the amount of effort required to develop.

Several researchers report the strength of a clustering method in organizing the Web based collections. Palmer et al. (2001) argued that where digital libraries are growing in size, and querying their contents becomes as frustrating as querying the Web, one remedy is to hierarchically cluster the results that are returned by searching a digital library. The study by

Zamir & Etzioni (1998) also supported the reason for considering clustering to be a feasible method of presenting the results of Web search engines, and suggested good methods for applying it to such a large collection. Many document clustering algorithms rely on clustering the entire document collection, but the Web search engines' collections are too large and fluid to allow off-line clustering. We used an off-line clustering method, Complete-Link clustering, for the source collection that has a reasonable document size ($N = 1,300$), and an on-the-fly clustering method, Carpineto and Romano (2004)'s Conceptual REorganization of DOcuments (CREDO⁴), for clustering the top ranked one hundred documents retrieved from the target collection. Therefore, the clustered structure of the source collection is static but the clustered structure of the target collection changes according to user's queries.

5. System Design

The design of the interface follows the metaphor and functionality of the ClusterBook (Muresan et al., 2001): access to information is afforded via direct search (as in a books index) or via browsing the structure of the domain, represented by the cluster hierarchy (as in a table of content). Different views of the document space are synchronized: when browsing the structure, the user can focus on each cluster's documents, at different depths of the hierarchy; conversely, when selecting a document in the raked list of search results, that document and its cluster are highlighted in the cluster structure, so that the user can explore its topical neighborhood. Thus, a user can benefit from searching or browsing the entire NJEDL collection (the N of documents = 1,300) with the static cluster structure where similar topics are tightly bound, or searching a relevant Web document from the clusters developed from a small retrieved Web document set ($N = 100$). We adopted the design of the tab-based interfaces for the NJEDL and the Web in the mediation mode (Belkin et. al., 2003). This design gives a user the full freedom to reiterate his usages between the NJEDL and the Web (Figure 3).

⁴ A Web-search engine that uses this clustering method is available at <http://www.fub.it>.

The top part of the user interface, used for the workflow of the experiment and the user input, is the same for all subjects, for consistency. The document visualization panels in the lower part, however, depend on the experimental group to which each subject is assigned. Each panel is designed and implemented as a separate module and has a certain visualization and exploration function: exploration of a ranked list of documents or exploration of the cluster structure.

In the non-mediated IR condition, the users directly search the target collection, the Web, to find the relevant information about the given topic. The subjects assigned to the NL group can only see the ranked list panel, as is the case in a typical Web search (Figure 1). Subjects in the NC group can scan the ranked list of results, or explore the topical structure obtained by structuring the search results, or may choose to combine the use of the two panels (Figure 2). When a document is selected, it appears in full-text form in a separate window. If the user regards the document as relevant, he saves it by clicking the Save button in the document window. When a document is saved, a dialog box invites the user to specify the reasons for selecting a particular document as relevant by specifying the aspects of the topic of interest addressed by that document (e.g., the aspects can refer to different instances in which an event can occur, or to different kinds of treatments, etc.). While specifying the aspects covered by each document directly supports the evaluation of the experiment, it also forces the user to analyze their information need and to ponder which aspects have been resolved. A surrogate of the saved document is now inserted in the saved document panel at the upper right hand side of the interface; a selected document can be re-opened in a viewer, e.g. for a side-by-side comparison with a new candidate to be saved, and can be “un-saved” if the user wishes to do so. The user repeats all these processes with new queries until either he is satisfied or the given time is up.

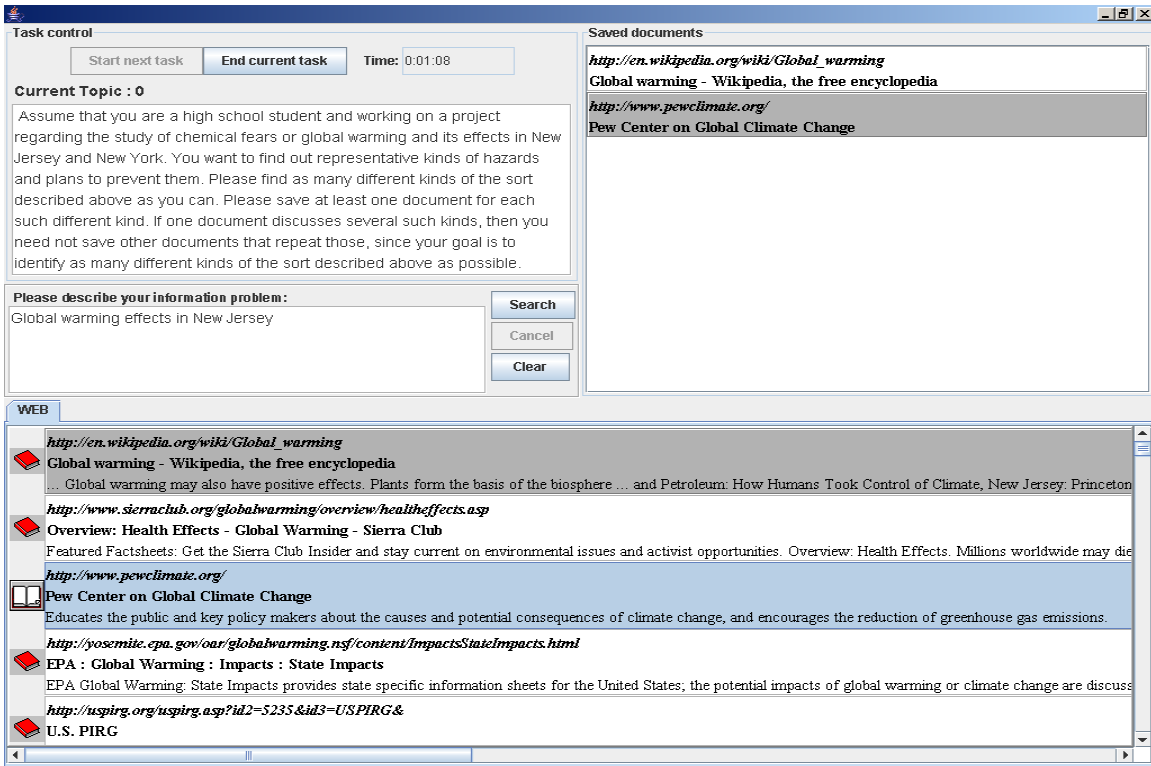


FIG 1. Non-mediated Linear list System (NL mode).

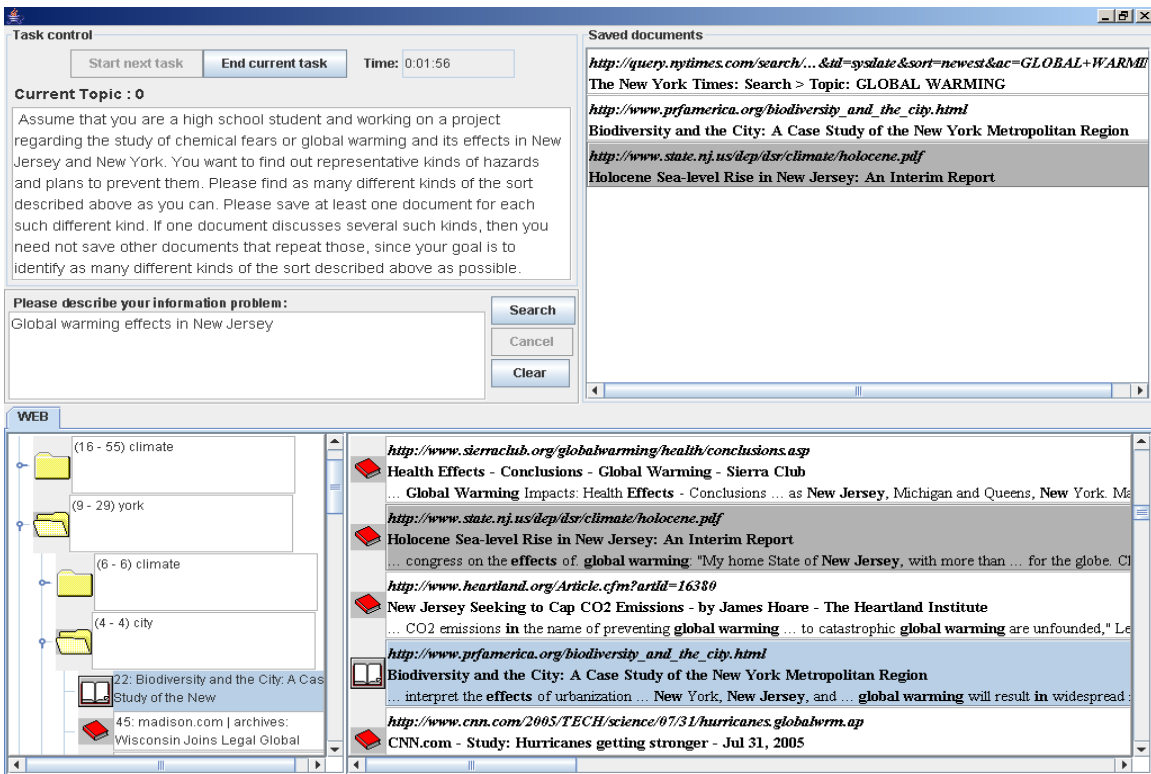


FIG 2. Non-mediated Combination display System (NC mode).

In the mediated IR condition, a tabbed panel allows the subjects to access either the source or the target collection. The recommended, but not enforced, flow of the interaction is as follows: the user, starting in the mediation tab, explores the domain of interest represented by the structured source collection, the NJEDL, selects and examines documents or cluster representatives for a certain information need, edits the query if necessary, then switches to the Web tab and submits the so-called “mediated query” to the search tool on the target collection, the Web (Figure 3). The user can explore the source collection either by submitting a search query, or by browsing the clusters, or by a combination of these.

The difference between the two mediation groups is similar to that between the non-mediation groups: while interacting with the target collection, in the Web tab, the subjects in the ML group can only scan the ranked list of results (similar to Figure 1), while the subjects in the MC group get the additional benefit of a clustered view of the results (Figure 4).

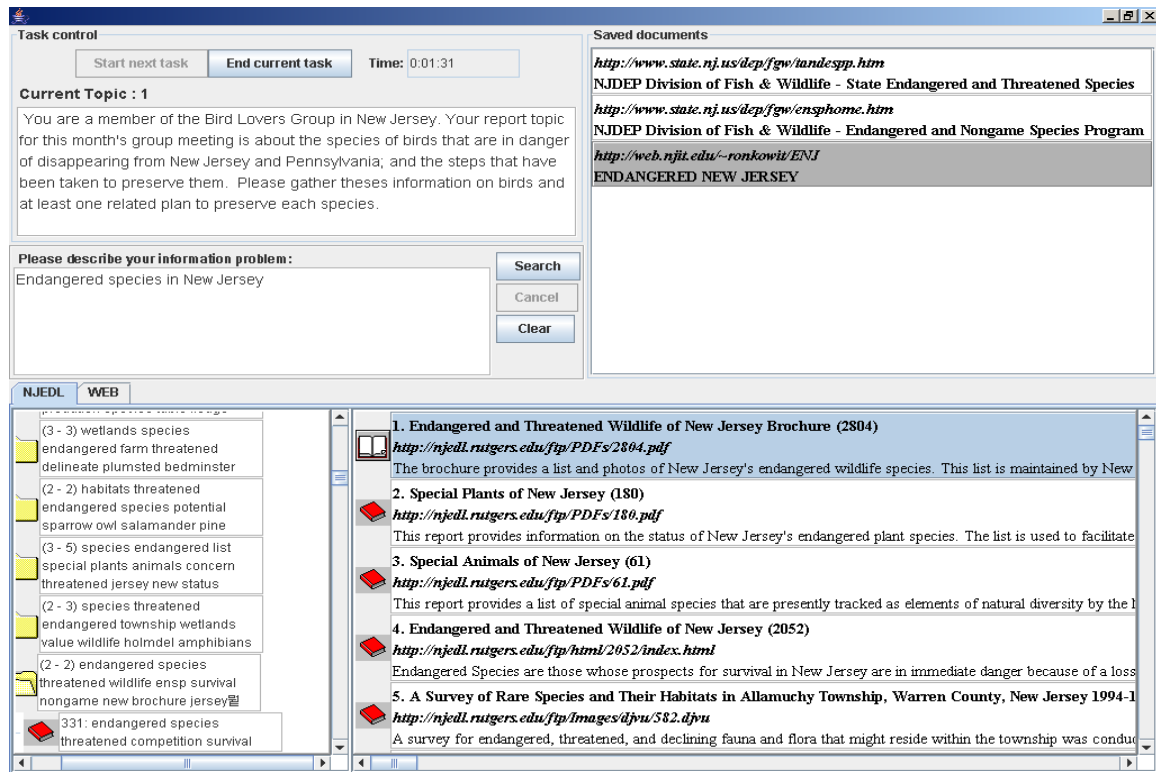


FIG 3. Source collection (NJEDL) of the Mediated Linear display System (ML mode).

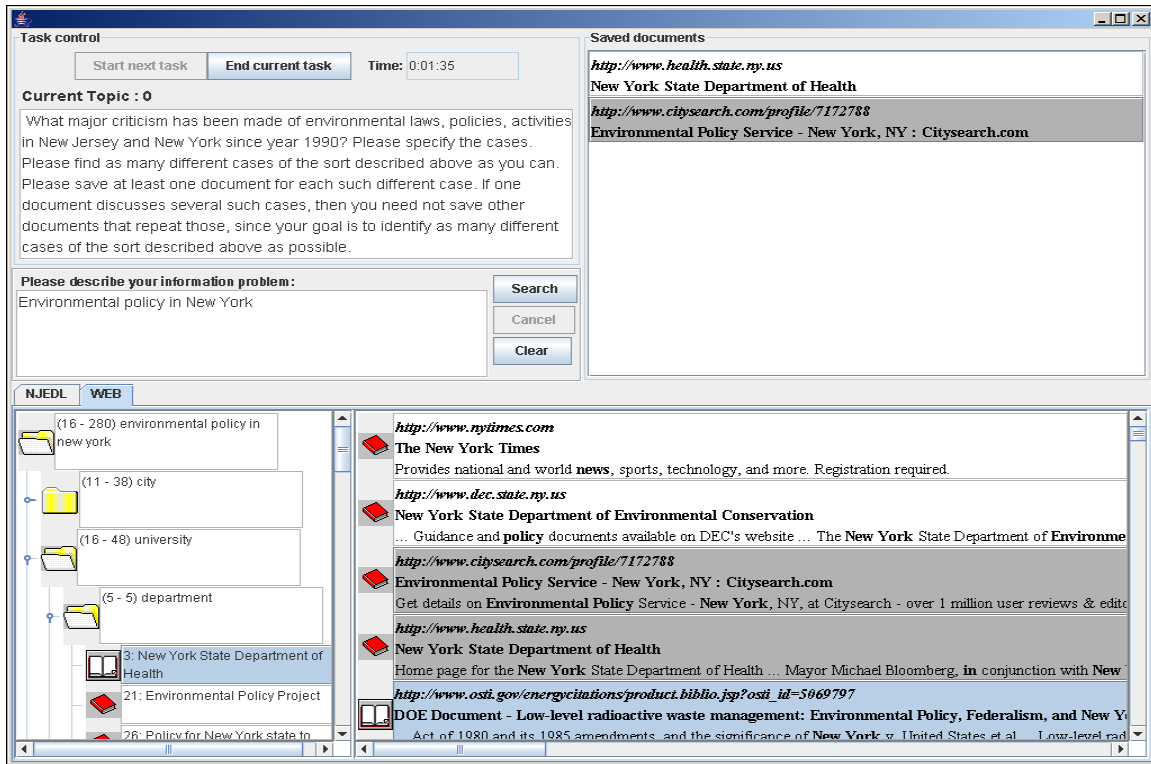


FIG 4. Target collection (Web) of the Mediated Combined display System (MC mode).

6. Research Methodology

All 32 subjects that we recruited to participate in this study were from the population of masters or doctoral students at Rutgers University. 7 Likert scale was used in our questionnaires (see Appendix B). Subjects' domain experience (3.22) and expertise about the environmental topics (2.88) were below average but their interest on the domain was average level (4.00). Aspectual retrieval (Over, 1997) was adopted as an appropriate evaluation measure for a task that requires the intermediary's support. As Koshman et al. (2006) explained, many Web users need to pool topics together and interact with the Web on more than one related topic, and in such multifaceted topic tasks they need the intermediary's support. Each topic describes an information need with many aspects - an aspect being roughly one of many possible answers to a question. Throughout the structured source collection, our system supports a user to investigate different sub-domains, and consequently, different aspects of a topic. Four topics regarding environmental

issues were developed for this study (see Appendix A). The task of the subject was to save relevant documents from the target collection which, taken together, cover as many different aspects of the topic as possible in the 20 minutes allowed per search. When saving a document, the subject was invited to record the relevant aspects covered by the document. While such action is not common search behavior, it has the advantage that it forced the subject to take the experiment seriously and to conduct at least a superficial analysis of the content of the documents read, and thus to save only relevant documents. Moreover, it helps in the experiment evaluation by providing the relevant aspects perceived by the subject; the alternative would be for the researchers to analyze each saved document, extract all candidate aspects, and assume that the subject perceived all these aspects as useful. The experiment was conducted in a computer lab at Rutgers University and approximately two hours were spent per search experiment; subjects completed the experiment one by one under supervision. All 32 subjects were randomly divided in half and each group (16 subjects) was assigned to one of the two displays of the IR search results conditions: a linear ranked interface, and respectively a combination of a linear ranked list and a clustered display of search results. Each subject was allocated 4 different topic searches: two of them with a non-mediated system, and two with a mediated system. Therefore, the first hypothesis (Mediated vs. Non-mediated) is a within-subject experiment, which is based on the belief that a user has to have a chance to experience both (Non-mediated and Mediated) IR conditions; and the second hypothesis (Linear vs. Combined) is the between-subject experiment. Rotation techniques were applied for the experiment design of both the systems and topics in order to avoid an order effect.

Experimental procedure

Each subject completed a questionnaire in which he was asked to fill in his demographic data. Then he conducted the first training session for the first system. Then the subject was given a sheet of paper which had a description of the first topic, and a brief pre-search questionnaire to evaluate their interest, experience and knowledge of the topic (see Appendix B). Important factors

for analyzing search behavior and evaluating the IR performance, such as task completion time, queries, aspects, opening/saving/unsaving documents, were logged during searching. The subject was asked to fill out a post-search questionnaire after finishing up each topic to answer questions about his searching experience (see Appendix B). After the subject completed the first two topics for one IR condition, he was given a system questionnaire (see Appendix B). The subject conducted a training session and started to use the second IR system for the other two topics. Then the subject filled out the second system questionnaire. After all four topic tasks were completed, the exit questionnaire was given to the subject.

Measures

The three main measures (effectiveness, efficiency, and usability) were divided into objective and subjective measures. Aspectual recall measures the objective effectiveness. Aspectual recall is the fraction of total aspects (as determined by the experimenter) for the topics that are covered by the pooled submitted documents. We categorized the aspectual recall into two different measurements in this study: aspectual recall from the saved documents (by dividing, for each topic, the total number of relevant aspects identified by the experimenter in the relevant saved documents to the total number of normalized aspects for that topic), and aspectual recall from the saved aspects (by dividing the total number of relevant aspects identified and recorded by the subjects to the total number of normalized aspects). In order for us to obtain these measurements, the experimenter predetermined a set of relevant aspects per topic through Web searching. Then, the experimenter identified relevant documents from the pooled submitted (saved) documents per topic after the experiments. We only considered a document relevant if it met all the conditions proposed in a given topic. Then, we combined the relevant aspects from the pooled saved 'relevant' document(s) per topic with the set of predetermined aspects. From the questionnaire, we gathered information on the user's perception of the task performance, which is our subjective measurement for the IR effectiveness: user's satisfaction and the time effectiveness on the search results. As the objective measurement, the elapsed time for the completion of the

task was measured. The amount of time a subject spent on this experiment compared to his ordinary Web searching was the determinant of subjective IR efficiency. Four measurements determine subjective usability: ease (ease of starting on a search, topic easiness for searching, ease of learning each of the two systems, Non-mediated and Mediated, and ease of using each of the two systems), usefulness (usefulness of the system for the tasks, usefulness of the combined user interface, and support from the source collection), degree of user's effort (difficulty of learning the system, difficulty of using the system, understanding of the way to use the system, and difficulty of understanding the structure of the display of (Web) search results (only for the combined display group), and preference (better supporter for searching between the two systems (Non-mediated and Mediated), and preference between the two systems). Finally, objective usability could be gathered via the logged data such as the number of the opened document, the number of saved/unsaved documents, the number of queries and query terms, and the query length.

6.1. Data analysis

Data analyses for the two hypotheses were conducted. We analyzed the effectiveness, efficiency, and usability by different units of the data: by each Linear and Combined group; by each Non-mediated and Mediated group; per each task order (total 4 tasks (4 topics, Appendix A)); and in two different subject groups based on their system usage order (Non-mediated & Mediated order group vs. Mediated & Non-mediated order group).

In order to find out if subject's experience of the tasks or the systems influenced his/her IR performance; data analyses per task were conducted to find out whether the order of the given tasks had any influence on the performances of the experiments, and data analyses per system usage order were conducted to observe whether the order of the usage of the two mediation systems had influenced any information search performances. We named the process in which a user becomes fully involved in an activity by gaining the experience and the skills as an adaptation process. Before discussing the findings, the definition of the abbreviated letters in the

tables and the figures are as follows. “N” indicates the non-mediated condition, “M” the mediated condition, “L” the traditional ranked linear list display, and “C” the combined display of the clustered structure and the ranked linear list. “NM” means the order of system usage is from the non-mediated system to the mediated, and “MN” means the opposite order.

7. Findings and Discussion

First, findings per task order and system usage order are reported. Then, the results about the two hypotheses are explained as the two subsections.

All the measures were analyzed according to the order of the four given tasks and system usage order between the two mediation conditions. All three measures mentioned above (effectiveness, efficiency, and usability) had interesting results. First, the order of tasks had an effect on all three main measures. In terms of objective effectiveness, both aspectual recalls from the saved documents and the saved aspects increased consistently in latter tasks, and for subjective effectiveness the user’s perception of task performance (both the satisfaction with the search results and the perception of the effective search time) also increased as more tasks were performed (Figure 5, scale; the first bar graph shows the percentage and the second the 7 Likert scale). In terms of efficiency, as the subjects conducted more tasks, they could finish the task faster (Figure 6, scale in seconds). It was also similar in several usability measures; for example, the subjects felt easier in starting their search tasks as they conducted more tasks and increasingly felt their search topics to be easier as they conducted more tasks (Figure 7, the 7 Likert scale). This shows that the subjects conducted better information searching performances as their searching experience with a complicated multifaceted task accumulated during the experiment.

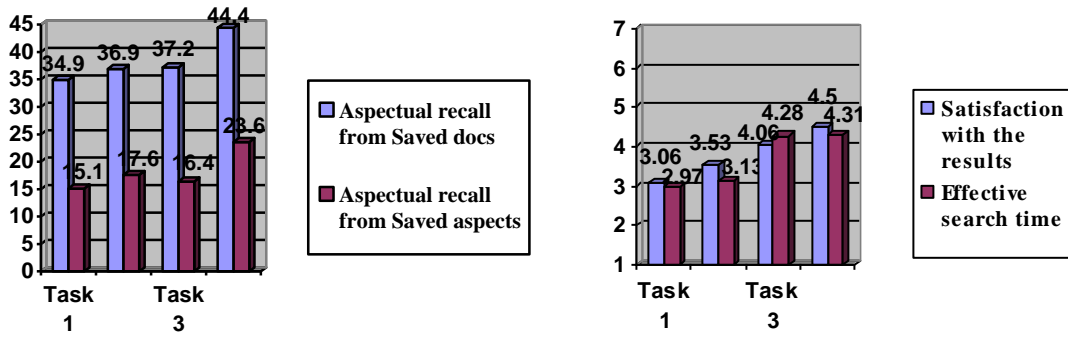


FIG 5. Objective and Subjective Effectiveness (per Task order).

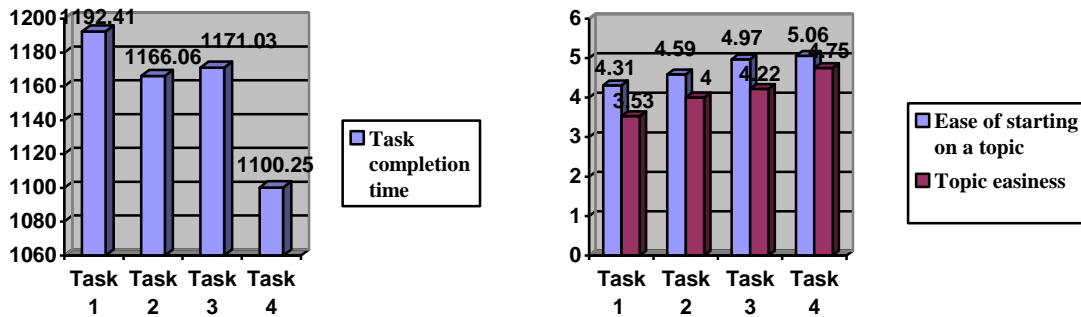


FIG 6. Objective Efficiency (per Task order). FIG 7. Subjective Usability (per Task order).

The usage order of the two mediation systems played an important role in every aspect of information searching as well. The subjects that conducted the non-mediated condition system first and then the mediated one (NM system usage order) showed better results in almost every measurement of the effectiveness, efficiency, and usability, compared to the ones that conducted the mediated condition system prior to the non-mediated one (MN system usage order). For example, aspectual recall from the saved aspects was better when a subject experienced the non-mediated condition before using the mediated condition system compared to the aspectual recall in the opposite order of usage ($\underline{M} = .20$ vs. $\underline{M} = .16$). Comparing the scores that the subjects gave when they used the mediated condition prior to the non-mediated condition with those of the opposite order, the subjects gave higher scores in the 7 Likert scale on both the satisfaction with search results ($\underline{M} = 4.17$, $\underline{SD} = 1.92$ (NM order) vs. $\underline{M} = 3.41$, $\underline{SD} = 1.80$ (MN order), $t(126) =$

2.32, $p < .05$) and the perception of the effective search time ($\underline{M} = 3.98$, $\underline{SD} = 1.88$ (NM order) vs. $\underline{M} = 3.36$, $\underline{SD} = 1.92$ (MN order)) when they used the non-mediated condition before using the mediated condition system. In efficiency, by measuring the second, subjects could finish their searches faster when they used the non-mediated condition before using the mediated condition ($\underline{M} = 1140.73$, $\underline{SD} = 148.35$ (NM order)) than in the opposite order of the system usage ($\underline{M} = 1174.14$, $\underline{SD} = 78.24$ (MN order)). On the other hand, subjects answered that they spent rather less time compared to the ordinary Web searching in the 7 Likert scale when they used the non-mediated condition prior to the mediated condition ($\underline{M} = 3.78$, $\underline{SD} = 1.31$ (NM order) vs. $\underline{M} = 4.53$, $\underline{SD} = 1.24$ (MN order), $t(62) = -2.35$, $p < .05$). Finally, in terms of usability, we found that the subjects in the NM order perceived easier in starting a search on a topic by answering the 7 Likert scale ($\underline{M} = 5.38$, $\underline{SD} = 1.38$ (NM order) vs. $\underline{M} = 4.09$, $\underline{SD} = 1.61$ (MN order), $t(126) = 4.82$, $p < .01$) and searching their topic tasks in the 7 Likert scale ($\underline{M} = 4.55$, $\underline{SD} = 1.58$ (NM order) vs. $\underline{M} = 3.70$, $\underline{SD} = 1.72$ (MN order), $t(126) = 2.88$, $p < .01$) than the ones in the MN order. It was same for the source collection and the combined display support; the subjects answered that the combined display was significantly less difficult in the NM order than in the MN order ($\underline{M} = 4.50$, $\underline{SD} = 1.31$ (NM order) vs. $\underline{M} = 2.63$, $\underline{SD} = 1.51$ (MN order), $t(64) = 2.66$, $p < .05$), and they answered the source collection supported their searching significantly more when they were in the NM order than in the MN order ($\underline{M} = 4.19$, $\underline{SD} = 1.42$ (NM order) vs. $\underline{M} = 3.31$, $\underline{SD} = 1.41$ (MN order), $t(64) = 2.38$, $p < .05$). The number of queries was also significantly smaller in the NM order ($\underline{M} = 5.89$, $\underline{SD} = 3.45$ (NM order) vs. $\underline{M} = 7.39$, $\underline{SD} = 4.02$ (MN order), $t(126) = -2.26$, $p < .05$). Overall, the subjects perceived that the systems that they used were significantly more useful in the NM order than in the MN order ($\underline{M} = 4.78$, $\underline{SD} = 1.24$ (NM order) vs. $\underline{M} = 3.81$, $\underline{SD} = 1.78$ (MN order), $t(126) = 3.21$, $p < .01$). The results emphasize the importance of an adaptation process when a subject faces a complicated multifaceted topic task and a new information retrieval system.

7.1. Mediated vs. Non-mediated

With regard to the two mediation conditions, the orders of the tasks and the system usage order were found to be strongly related to the measures. First, the findings indicate that the more information searching tasks a subject conducts, the better the results the mediated condition produces in every aspect of Web searching than the non-mediated one. For example, in their final task (task 4), subjects did not only score the best aspectual recall from the saved aspects in the mediated condition, but the difference between the non-mediated and the mediated conditions was also the largest among all the tasks (Table 2). The subjects increasingly became more satisfied with the mediated condition than with the non-mediated condition as they conducted more tasks (Table 2). In addition, as subjects conducted more tasks, they perceived that the mediated condition was better than the non-mediated one with regard to time effectiveness (Table 2). Also in objective efficiency, in the final task (task 4) the average task completion time in the mediated condition was 1 minute and 34 seconds shorter than the one in the non-mediated condition, which was the biggest difference among all four tasks (Table 3). In subjective efficiency, the subjects in the first two tasks (task 1 and 2) answered that they spent more time in their searching than in their usual Web searching in the mediated condition ($\underline{M} = 4.69$, $\underline{SD} = 1.45$) than in the non-mediated condition ($\underline{M} = 3.88$, $\underline{SD} = 1.26$). However, it was the opposite in the latter two tasks (task 3 and 4) – interestingly, they answered that they spent more time in the non-mediated condition ($\underline{M} = 4.38$, $\underline{SD} = 1.03$) than in the mediated condition ($\underline{M} = 3.69$, $\underline{SD} = 1.40$). Most usability measures indicated similar cases. For example, in subjective usability, subjects gave significantly higher scores in the 7 Likert scale for easiness of starting on a topic task to the non-mediated condition in the first two tasks⁵ (Table 4). However, it was the opposite case in the latter two tasks; especially, subjects gave significantly higher scores to the mediated condition in their last task⁶ (Table 4). In addition, they answered that the topic was easier when they used the mediated condition than in the non-mediated one as they conducted more tasks (Table 4). In

⁵ Task 1: Non-mediated vs. Mediated ($M = 5.19^{**}$, $SD = 1.42$ vs. $M = 3.44^{**}$, $SD = 1.50$), $t(30) = 3.38$, $p < .01$

Task 2: Non-mediated vs. Mediated ($M = 5.29^*$, $SD = 1.48$ vs. $M = 3.94^*$, $SD = 1.48$), $t(30) = 2.50$, $p < .05$

⁶ Task 4: Non-mediated vs. Mediated ($\underline{M} = 4.44^*$, $\underline{SD} = 1.67$ vs. $\underline{M} = 5.69^*$, $\underline{SD} = 1.40$), $t(30) = -2.29$, $p < .05$

objective usability, the number of queries was larger in the mediated condition than in the non-mediated condition across all the tasks. In the first two tasks, the difference was statistically significant⁷; however, in the final task (task 4), it was almost equal (Table 5). Not just the number of queries but also the number of query terms in the mediated condition was the smallest in the final task as well (Table 5).

TABLE 2. Effectiveness- Objective and Subjective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
	M	N	M	N	M	N	M	N
Aspectual recall for saved aspects (percentage)	.12 (.17)	.18 (.17)	.18 (.21)	.17 (.20)	.17 (.19)	.16 (.19)	.28 (.24)	.20 (.21)
Satisfaction with the search results (7 Likert scales)	2.19 (1.27)	3.94 (1.73)	3.31 (1.53)	3.75 (1.98)	4.00 (2.16)	4.12 (1.99)	5.00 (1.71)	4.00 (1.78)
Perception of the effective search time (7 Likert scales)	2.31 (1.49)	3.63 (1.66)	2.81 (1.68)	3.44 (1.63)	4.25 (2.11)	4.31 (1.92)	4.63 (1.99)	4.00 (1.96)

TABLE 3. Efficiency- Objective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
	M	N	M	N	M	N	M	N
Task completion time (seconds)	1190.63 (36.71)	1194.19 (16.53)	1188.56 (41.19)	1143.56 (169.63)	1172.00 (66.22)	1170.06 (70.71)	1053.19 (216.15)	1147.31 (127.95)

TABLE 4. Usability- Subjective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
	M	N	M	N	M	N	M	N
Ease of starting on a topic (7 Likert scales)	3.44 (1.50)	5.19 (1.42)	3.94 (1.48)	5.25 (1.48)	5.37 (1.31)	4.56 (1.67)	5.69 (1.40)	4.44 (1.67)
Topic easiness (7 Likert scales)	2.75 (1.44)	4.31 (1.30)	3.63 (1.50)	4.37 (1.63)	4.44 (1.71)	4.00 (1.79)	5.06 (1.69)	4.44 (1.83)

TABLE 5. Usability- Objective (per Task Order).

Task order	Task 1	Task 2	Task 3	Task 4
------------	--------	--------	--------	--------

⁷ Task 1: Non-mediated vs. Mediated (\underline{M} = 4.31*, \underline{SD} = 2.87 vs. \underline{M} = 7.94*, \underline{SD} = 4.88), $t(30) = -2.56$, $p < .05$

Task 2: Non-mediated vs. Mediated (\underline{M} = 5.06**, \underline{SD} = 1.69 vs. \underline{M} = 9.69**, \underline{SD} = 4.24), $t(30) = -3.68$, $p < .01$

Mediation condition	M	N	M	N	M	N	M	N
Number of queries (frequency)	7.94 (4.88)	4.31 (2.87)	9.69 (4.24)	5.06 (.2.69)	8.06 (4.22)	6.06 (2.69)	6.13 (2.89)	5.88 (2.92)
Number of query terms (frequency)	46.13 (31.00)	25.56 (21.36)	54.00 (29.05)	26.63 (14.10)	57.81 (72.12)	34.44 (17.01)	33.00 (16.12)	33.63 (18.30)

Second, it was found that the subjects that used the non-mediated condition system prior to the mediated one were able to experience an adaptation process before being exposed to a new IR system. For example, when a subject experienced the non-mediated condition first, the aspectual recall from both the saved documents and aspects in the mediated condition was higher than the one in the non-mediated condition (Table 6). Interestingly, it was exactly the opposite when a subject first used the mediated condition and then the non-mediated condition, as the aspectual recall from both the saved documents and aspects in the mediated condition was lower than the one in the non-mediated condition (Table 6). In addition, when the subjects used the non-mediated first, they were more satisfied with the results and significantly felt better about the time effectiveness of their searches in the mediated condition ($F = 4.05, p < .05$) than in the non-mediated condition (Table 6). However, in the opposite order of the system usage, satisfaction with the results ($F = 9.49, p < .01$) and time effectiveness of the searches in the mediated condition ($F = 13.33, p < .01$) was significantly lower than the one in the non-mediated condition (Table 6). Also, the task completion time was shorter in the mediated condition than in the non-mediated one when the non-mediated condition was offered prior to the mediated condition, and vice-versa when the mediated condition was offered prior to the non-mediated condition (Table 7). Subjects who used the non-mediated condition prior to the mediated one answered that they spent slightly more time in the non-mediated than in the mediated condition, which was opposite when they used the mediated condition prior to the non-mediated one ($t(62) = -1.01, p < .05$) (Table 7).

In addition, in usability, while there were no differences with regard to ease of starting on a topic and topic easiness between the two mediation conditions in the NM system usage order, the subjects in the MN system usage order answered that the non-mediated condition was

significantly better in starting on a topic ($t(62) = 2.07, p < .05$) and also in topic easiness ($t(62) = 2.49, p < .05$) (Table 8). The subjects' attitudes towards the ease of system were interesting as well. In the NM system usage order, there was no difference in how easy a subject perceived the learning of the two mediation condition systems⁸; and interestingly, over twice the number of the subjects answered that the mediated condition was easier than the non-mediated condition in terms of using a system⁹. In contrast, in the MN system usage order, significantly more subjects felt that the non-mediated condition system was easier to learn ($\chi^2 = 9.31 (df = 1), p < .01$)⁸ and use ($\chi^2 = 6.23 (df = 2), p < .05$)⁹. In addition, the subjects perceived that the mediated condition was more useful than the non-mediated condition in the NM system usage order ($M = 4.56, SD = 1.21$ (Non-mediated) vs. $M = 5.00, SD = 1.27$ (Mediated)); on the contrary, the usefulness of the non-mediated condition was slightly higher than the mediated condition in the MN system usage order ($M = 3.87, SD = 1.36$ (Non-mediated) vs. $M = 3.75, SD = 1.00$ (Mediated)). Finally, significantly more subjects in the NM system usage order answered that the mediated condition system to be a better supporter in searching than the non-mediated one ($\chi^2 = 4.00 (df = 1), p < .05$)¹⁰ and also preferred the mediated condition system ($\chi^2 = 5.33 (df = 1), p < .05$)¹¹; however, the preference towards the two mediation conditions was reversed in the MN system usage order^{10 11}.

TABLE 6. Effectiveness- Objective and Subjective (per Order of System Usage).

Order of system usage Mediation condition	Mediated & Non-mediated		Non-mediated & Mediated	
	M	N	M	N
Aspectual recall for saved documents (percentage)	.36 (.28)	.43 (.28)	.38 (.26)	.35 (.22)
Aspectual recall for saved aspects (percentage)	.15 (.19)	.18 (.20)	.22 (.22)	.18 (.18)
Satisfaction with the search results (7)	2.75	4.06	4.50 (1.98)	3.84 (1.83)

⁸ Ease of learning between the two systems: Non-mediated: 4, Mediated: 3, NA :9 (NM system usage order) vs. Non-mediated: 12, Mediated: 1, NA :3 (MN system usage order)

⁹ Ease of using between the two systems: Non-mediated: 3, Mediated: 6, NA :7 (NM system usage order) vs. Non-mediated: 11, Mediated: 2, NA :3 (MN system usage order)

¹⁰ Better supporter for searching between the two systems: Non-mediated: 0, Mediated: 12, NA :4 (NM system usage order) vs. Non-mediated: 6, Mediated: 5, NA :5 (MN system usage order)

¹¹ Preference between the two systems: Non-mediated: 2, Mediated: 10, NA :4 (NM system usage order) vs. Non-mediated: 9, Mediated: 4, NA :3 (MN system usage order)

Likert scales)	(1.50)**	(1.87)**		
Perception of the effective search time (7 Likert scales)	2.56 (1.58)**	4.16 (1.92)**	4.44 (2.03)*	3.53 (1.63)*

* p > 0.05; ** p > 0.01

TABLE 7. Efficiency- Objective and Subjective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Mediation condition	M	N	M	N
Task completion time (seconds)	1189.59 (38.40)	1158.69 (102.34)	1112.59 (168.44)	1168.88 (121.31)
Spent time on searching task compared to the ordinary Web searching (7 Likert scales)	4.69 (1.45)*	4.38 (1.02)*	3.69 (1.40)	3.88 (1.26)

* p > 0.05

TABLE 8. Usability- Subjective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Mediation condition	M	N	M	N
<i>Degree of ease per topic</i>				
Ease of starting on a topic (7 Likert scales)	3.69 (1.49)*	4.50 (1.64)*	5.53 (1.34)	5.22 (1.43)
Topic easiness (7 Likert scales)	3.19 (1.51)*	4.22 (1.79)*	4.75 (1.70)	4.34 (1.45)

* p > 0.05

Consequently, the order of system usage had an important impact on the relationship between the two mediation conditions across all three measures: the effectiveness, efficiency, and usability. When a subject had the adaptation process with a rather familiar information system, such as the non-mediated one, most of the results showed that the mediated condition was better with regard to the measures.

Overall, these results lead us to the conclusion that the mediated system is more effective, efficient, and useful than the non-mediated system as a subject conducts more multifaceted information searching tasks, and when he is provided with a preliminary process with a familiar IR system before using the mediated IR system. More work need to be conducted to understand the potential of other factors that influence the mediation condition for Web searching.

7.2. Linear vs. Combined

Similar to the results regarding the mediation, the effectiveness and usability with regard to the display condition had interesting results via the analyses in two important factors: the order

of the four given tasks and the system usage order between the two mediation conditions. Several results of the effectiveness and the usability revealed that the combined display was more effective and useful than the linear display when a subject conducted more tasks, or when he experienced a familiar non-mediated system prior to utilizing a new mediated one. First, in the final task (task 4), the subjects scored the best aspectual recall from the saved aspects in the combined display; the difference between the combined display and the linear display was also the biggest in their final task (Table 9). In objective efficiency, it revealed that the linear display group was always faster than the combined display group throughout any order of the tasks (Table 10). In objective usability, the linear display helped its subjects to find and save more documents that contain aspects compared to the combined display in the first task; however, it became equal in the last task (Table 11). Similar to what we found in the research question 1 (the mediation), this result implies that a subject needs an adaptation process to use the new display, such as the combined display of the hierarchical structure and the linear ranked list.

TABLE 9. Effectiveness- Objective and Subjective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
Display condition	L	C	L	C	L	C	L	C
Aspectual recall for saved aspects (percentage)	.15 (.18)	.15 (.17)	.17 (.24)	.18 (.16)	.17 (.21)	.16 (.17)	.22 (.20)	.26 (.26)
Satisfaction with the search results (7 Likert scales)	3.44 (1.79)	2.69 (1.66)	3.75 (2.05)	3.31 (1.45)	4.50 (2.28)	3.63 (1.75)	4.62 (1.78)	4.37 (1.86)
Perception of the effective search time (7 Likert scales)	3.13 (1.63)	2.81 (1.80)	3.44 (1.75)	2.81 (1.60)	5.19 (1.97)	3.38 (1.59)	4.50 (1.86)	4.13 (2.12)

TABLE 10. Efficiency- Objective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
Display condition	L	C	L	C	L	C	L	C
Task completion time (seconds)	1189.06 (36.95)	1195.75 (15.43)	1142.44 (169.29)	1189.69 (41.25)	1151.56 (87.45)	1190.50 (30.53)	1088.88 (210.40)	1111.63 (181.15)

TABLE 11. Usability - Objective (per Task Order).

Task order	Task 1		Task 2		Task 3		Task 4	
Display condition	L	C	L	C	L	C	L	C
Number of saved documents (frequency)	3.63 (2.03)	2.81 (2.43)	3.31 (1.49)	3.06 (2.17)	3.75 (2.24)	3.06 (1.73)	3.63 (1.15)	3.88 (1.93)

Second, the result of the aspectual recall from the saved aspects indicates that the combined display was more helpful when the subjects used the non-mediated condition before using the mediated one than when they used the system in the opposite order (MN order) (Table 12). Unlike the objective effectiveness, the results regarding the user’s perception of task performance showed that the linear display condition was better regardless of the order of the system usage (Table 12). The findings imply that the subjects tend to perceive that they had better search results in a more familiar interface, such as the linear display, regardless of their actual retrieval performance in either of the system usage order. In subjective efficiency, in both the NM and the MN system usage orders, the subjects perceived that the time spent compared to the ordinary Web searching was longer than that of the linear display when they were in the combined display group (Table 13)¹². In subjective usability, the subjects in the MN system usage order answered that they perceived both starting on the search tasks and topic easiness in the traditional linear display condition to be easier than in the combined display condition. However, there was smaller or no difference in both measurements in the two display conditions when a subject faced the NM system usage order (Table 14). Also, the subjects in the combined display answered that it was more difficult to learn and use the system in the MN system usage order compared to the ones in the linear display; but there was smaller difference in the NM order (Table 14). In terms of usefulness, when the subjects experienced the MN system usage order, the ones in the linear display answered that they received significantly more support from the source collection (NJEDL) than the ones in the combined display ($F = 6.42, p < .05$); however, there was no difference between the two display conditions when they were in the NM order (Table 14).

¹² The difference between the two displays was larger in the MN system usage order than in the NM order.

Also, the combined display of the search results was slightly more useful in their searching in the NM system usage order than the linear display compared to the opposite result in the MN order (Table 14). Finally, in the objective usability, when the subjects used the non-mediated condition prior to the mediated condition, the subjects in the combined display used slightly more queries than the ones in the linear display. On the other hand, the subjects in the MN order used significantly more queries in the linear display than in the combined display ($F = 9.74, p < .01$) (Table 15). The average query length between the two display conditions was not affected by the order of system usage (Table 15).

TABLE 12. Effectiveness- Objective and Subjective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Display condition	L	C	L	C
Aspectual recall for saved aspects (percentage)	.17 (.19)	.16 (.19)	.18 (.22)	.22 (.19)
Satisfaction with the search results (7 Likert scales)	3.63 (1.93)	3.19 (1.67)	4.53 (2.00)	3.81 (1.80)
Perception of the effective search time (7 Likert scales)	3.66 (1.99)	3.06 (1.83)	4.47 (1.85)	3.50 (1.81)

TABLE 13. Efficiency- Objective and Subjective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Display condition	L	C	L	C
Task completion time (seconds)	1176.50 (68.57)	1171.78 (87.91)	1109.47 (187.81)	1172.00 (86.22)
Spent time on searching task compared to the ordinary Web searching (7 Likert scales)	4.06 (1.12)	5.00 (1.21)	3.50 (1.41)	4.06 (1.18)

TABLE 14. Usability- Subjective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Display condition	L	C	L	C
<i>Degree of ease per topic</i>				
Ease of starting on a topic (7 Likert scales)	4.56 (1.70)	3.63 (1.38)	5.56 (1.34)	5.19 (1.42)
Topic easiness (7 Likert scales)	4.09 (1.87)	3.31 (1.49)	4.53 (1.66)	4.56 (1.52)
<i>Degree of ease per system</i>				
Difficulty of learning a system (7 Likert scales)	1.87 (1.50)	2.75 (1.06)	1.81 (1.28)	2.13 (1.02)
Difficulty of using a system (7 Likert scales)	2.19 (1.56)	2.94 (1.06)	1.88 (1.50)	2.31 (1.14)

Understanding the way to use the system (7 Likert scales)	5.12 (1.31)	4.19 (1.38)	4.81 (2.10)	5.44 (1.09)
<i>Usefulness</i>				
Support from Source Collection (mediated condition group, 7 Likert scales)	3.94 (1.29)*	2.69 (1.49)*	4.25 (1.29)	4.13 (1.59)
Usefulness of the system for tasks (7 Likert scales)	4.00 (1.26)	3.63 (1.09)	4.63 (1.41)	4.94 (1.06)

* p > 0.05

TABLE 15. Usability- Objective (per Order of System Usage).

Order of system usage	Mediated & Non-mediated		Non-mediated & Mediated	
Display condition	L	C	L	C
Number of queries (frequency)	8.78 (4.28)*	6.00 (3.24)*	5.59 (3.93)	6.19 (2.93)
Number of query terms (frequency)	47.94 (26.33)	36.16 (23.64)	37.47 (54.36)	34.03 (18.71)
Average query length (term length)	5.47 (1.40)	5.76(1.09)	6.06 (2.61)	5.76 (2.27)

* p > 0.05

Overall, a variety of measurements show that as a subject conducts more multifaceted information searching tasks, the combined display revealed at least similar results with the linear display or better. It was the same case when a subject experiences a familiar system prior to facing a new system, but the linear display revealed better results in the reverse system usage order.

8. Conclusion and Future study

The major findings of this study contribute to the research on the development of Web IR systems that have a mediation function and a structured display of the search results, which may support complex Web search tasks such as exploring a multifaceted topic. First, the mediation function for the Web searching is notably influenced and improved by the adaptation process¹³ for a complicated multifaceted topic task and a novel system. Second, the combined display of the document cluster and the linear ranked list is also clearly influenced by the adaptation process for a multifaceted task and a novel system. In the future study, we should investigate the system order

¹³ As explained, the adaptation process is the stage where a user becomes fully involved in an activity by gaining the experience and the skills.

effect for the two displays as a within-subject experiment. We also found that it is not ideal to offer an un-trained Web user the combination of the mediation system and the structured display, regardless of the system's powerful functionalities; the subjects did not favor the most complicated MC mode in many aspects. The findings also contribute to the research area on the Web IR system with the mediation function in terms of system testing because the mediated Web IR system has been rarely tested by real subjects or a specific task type such as the multifaceted topic. Furthermore, few studies have considered the possibility of comparing the effectiveness of the ranked list of search results to the combination of the ranked list and a clustered document display.

Results of this research suggest additional research agendas involving both the mediation condition and the display condition. We may investigate the level in which a novice user makes the optimum information retrieval results by experiencing how many number of tasks or system usage. Such research should have practical implications for the Web IR system design. Future studies might also explore identifying the effect of different types of tasks on either the mediation condition or the combined display. What should be also remembered is that the subjects of this research had similar levels of domain knowledge on the given topics and that all were of the same educational background. Thus, if a system makes use of the characteristics of the user group or the characteristics of the given task, then the system may suggest which system or display of the search results would fit better for a specific user group or a specific kind of tasks. More research needs to be conducted on understanding how a subject uses the mediated IR system or the combined display in real life situations as well. This study was conducted in an experimental setting and several subjects of this study pointed out their pressure of the limited time and their realization of the fact that they were conducting an experiment. Are subjects of the mediation system or the combined display going to demonstrate a similar Web searching performance in their everyday searching? A longitudinal study would be required to answer such a question.

Future study might also investigate the effect of different clustering methods, clustering algorithms, or different types of cluster labels. Despite an effort to adjust an ideal hierarchical structure and cluster labels for the source collection, several subjects in this study indicated that they had difficulty understanding how the structure was composed and that the structure was also confusing due to the large number of clusters and terms per cluster representation. Application of the recent findings such as Geraci et al.'s study (2007) would be an interesting agenda. Further investigations may be necessary to find out whether a particular clustering method or algorithm groups the documents better for a particular document collection, and what types of cluster labels facilitate Web searching performances.

Even without further experiments more conclusions can be drawn from more analysis of the interaction logs. We could verify the mediation assumptions by verifying if the users were able to find the best clusters and if the quality of their queries improved via mediation. Simulations can also be conducted to compare the mediated queries formulated by the real users with queries generated by the mediation system: it may be that the algorithms are better than humans at formulating queries that better capture, in a statistical sense, the content of a document cluster, and the user's own information need.

In conclusion, the findings of this study contribute to a better understanding of how the mediation system and the combined display support a Web information user having a multifaceted search topic. The research findings have practical implications for the development of the Web IR system in terms of the possibilities of the mediation system or the document structuring for user-centered information retrieval. In addition, the findings of this research suggest future research agendas that can be further investigated.

Acknowledgments

I would like to thank Nick Belkin, my dissertation advisor, as well as my committee members, Gheorghe Muresan, Dan O'Connor and David Harper, for their contributions and

continuous support to this work. I would like to give special thanks to Stewart Mohr who kindly supported the recruitment of subjects. Finally, I would like to give special thanks to the SILS department at Texas Women's University for their generous support in allowing me to complete the dissertation and to work at the same time.

References

Allan, J., Leuski, A., Swan, R., & Byrd, D. (2000). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management*, 37, 435-458.

Anick, P. G., & Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 314-323). New York: ACM.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982a). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2), 61-71.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982b). ASK for information retrieval: Part II. Results of a design study. *Journal of Documentation*, 38(3), 145-164.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information by retrieval. *Information Processing & Management*, 31(3), 431-448.

Belkin, N. J., Kelly, D., Lee, H. J., Li, Y., Muresan, G., Tang, M. C., Yuan, X. J., & Zhang, X. M. (2003) Rutgers' HARD and Web interactive track experiments at TREC 2003. *Proceedings of the 12th Text REtrieval Conference (TREC-12)* (pp. 532-543). Washington, D.C.: GOP.

Carpineto, C. & Romano, G. (2004). Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science* 10(8), 985-1013.

- Cole, C., Mandelblatt, B., & Stevens, J. (2002). Visualizing a high recall search strategy output for undergraduates in an exploration stage of researching a term paper. *Information Processing & Management*, 38(1), 37-54.
- Croft, W. B. (1995). What do people want from information retrieval?, *D-Lib Magazine*, November. Retrieved July 22, 2004, from <http://www.dlib.org/dlib/november95/11croft.html>
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.
- Efthimiadis, E. N. (1996). Query expansion. In Williams, & E. Martha (Eds.), *Annual Review of Information Systems and Technology (ARIST)* (Vol. 31, pp. 121-187).
- Ellis, D., Furner-Hines, J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in information management*, 3(2), 129-149.
- Geraci, F., Pellegrini, M., Maggini, M. & Sebastiani, F. (2006). Cluster generation and cluster labelling for Web Snippets: A fast and accurate hierarchical solution. *Proceedings of String Processing and Information Retrieval 13th International Conference (SPIRE 2006): Lecture Notes in Computer Science*, 4209, (pp.25-36). Berlin/Heidelberg: Springer.
- Haynes R. B., Johnston M.E., McKibbin K.A., Walker C.J., & Willan A.R. (1992). A randomized controlled trial of a program to enhance clinical use of MEDLINE. *Online Journal of Controlled Clin Trials*.
- Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96* (pp.76-84). New York: ACM.
- Huang, C-K., Chien, L-F., & Oyang, Y-J. (2003). Relevant term suggestion in interactive Web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, 54(7), 638-649.

Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht, Netherlands: Springer.

Jansen, B. J., Spink, A., et al. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum: A Publication of the Special Interest Group on Information Retrieval*, 32, 5-18.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42, 248-263.

Jardine, N., & Van Rijsbergen (1971). The user of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7, 217-240.

Koshman, S., Spink, A., & Jansen, B. J. (2006). Web searching on the Vivisimo search engine. *Journal of the American Society for Information Science and Technology*, 57(14), 1875-1887.

Kural, Y., Robertson, S. E., & Jones, S. (2001). Deciphering cluster representations. *Information Processing and Management*, 37(4), 593-601.

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. *Proceedings of 10th International Conference on Information and Knowledge Management (CIKM'01)* (pp. 41-48). New York: ACM Press.

Leuski, A., & Allan, J. (1999). The best of both worlds: Combining ranked list and clustering. *CIIR Technical Report IR-172*, University of Massachusetts. Retrieved June 12, 2003, from <http://people.ict.usc.edu/~leuski/publications/papers/ir-172.pdf>

Leuski, A., & Allan, J. (2000). Improving interactive retrieval by combining ranked lists and clustering. *Proceeding of 6th Conference on Content-Based Multimedia Information Access (RIAO 2000)*, (pp. 665-681).

Luhn, H. P. (1958). A business intelligent system. *IBM Journal of Research and Development*, 2, 314-319.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed Digital Libraries: treatment of term vagueness and document re-ranking. *Library Review* 57(3), 213-224.

McCune, B. P., Tong, R. M., Dean, J. S., & Shapiro, D. G. (1985). RUBRIC: A system for rule-based information retrieval, *IEEE Transactions on Software Engineering (TSE)*, 11(9), 939-945.

Muresan, G. (2002). Using Documents Clustering and Language Modeling in Mediated Information Retrieval. Doctoral dissertation, The Robert Gordon University, Aberdeen, Scotland.

Muresan, G., & Harper, D. J. (2001). Document clustering and language models for system-mediated information access. *Proceedings Series: Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries 6th European Conference (ECDL 2001)* (pp. 438-449). Berlin / Heidelberg: Springer.

Muresan, G. & Harper, D. J. (2004). Topic modeling for mediated access to very large document collections. *Journal of the American Society for Information Science*, 55(10), 892-910.

Nordlie, R. (1996). Unmediated and mediated information searching in the public library. In S. Hardin (Ed.), *Proceedings of the 59th ASIS Annual Meeting*, (Vol. 33, pp. 41-46). Medford, NJ: Information Today.

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of Documentation*, 33(1), 1-14.

Oh, S. G. (1998). Document representation and retrieval using empirical facts: Evaluation of a pilot system. *Journal of the American Society for Information Science*, 49(10), 920-931.

Osdin, R., Ounis, I., & White, R. W. (2002). Using hierarchical clustering and summarization approaches for Web retrieval. *Proceedings of the 11th Text REtrieval Conference (TREC-11)*. Washington, D.C.: GOP.

Over, P. (1997). TREC-6 Interactive track report. *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* (pp. 73-82). Washington, D.C.: GOP.

Palmer, C. R., Pesenti, J., Valdes-Perez, R. E., Christel, M. G., Hauptmann, A. G., Ng, D., & Wactlar, H. D. (2001). Demonstration of hierarchical document clustering of digital library retrieval results. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2001) (pp. 451). New York: ACM.

Roussinov, D. G., & Chen, H. (2001). Information navigation on the Web by clustering and summarizing query results. *Information Processing & Management*, 37, 789-816.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Co.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (3rd ed.). MA: Addison-Wesley.

Spink, A. (1995). Term relevance feedback and mediated database searching: Implications for information retrieval practice and systems design. *Information Processing & Management*, 31(2), 161-171.

Spink, A & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. Dordrecht, The Netherlands: Kluwer Academic.

Spink, A., Wilson, T. D., Ford, N., Foster, A., & Ellis, D. (2002a). Information-seeking and mediated searching. Part 1. Theoretical framework and research design. *Journal of the American Society for Information Science and Technology*, 53(9), 695-703.

Spink, A., Wilson, T. D., Ford, N., Foster, A., & Ellis, D. (2002b). Information seeking and mediated searching study. Part 3. Successive searching. *Journal of the American Society for Information Science and Technology*, 53(9), 716-727.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29, 178-194.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworth

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5), 577-597.

Wu, M., Fuller, M., & Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. *Information Processing and Management*, 37(3), 459-484.

Zamir, O., & Etzioni, O. (1998). Web document clustering: A feasibility demonstration. *Proceedings of the 21st ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 46-53), New York: ACM.

Appendix A. The topics

Topic 1

Description:

You are an environmental activist from the Green Party. Among various air pollutants, there are major pollutants causing air pollution. Please gather information on which ones mainly comprise air pollution in New Jersey and Pennsylvania since the year 2000. It should be limited to an officially recognized pollutant.

Aspects:

In the time allotted, please find as many different kinds of air pollutants described above as you can. Save at least one document for each pollutant. If one document discusses several kinds, then you need not save other documents that repeat those same pollutants.

Topic 2

Description:

You are a researcher in New Jersey Department of Environmental Protection. You are reviewing education programs related to water or air pollution for elementary schools in New Jersey since the year 2000. Please locate different kinds of attempts that have been implemented.

Aspects:

In the time allotted, please find as many different attempts of the sort described above as you can. The program does not have to be limited to a regular course but any type of program, past (after year 2000) and future, is relevant as well. If the project covers several student levels implicitly including an elementary school, that is relevant. However, project which is not related to an elementary school student is not relevant. Nation-wide programs are not relevant either. Please save at least one document for each different kind of program. If one document discusses several such programs, then you need not save other documents that repeat the similar curriculums, since your goal is to identify as many different kinds of the sort described above as possible.

Topic 3

Description:

You are a member of the Bird Lovers Group in New Jersey. Your report topic for this month's group meeting is about the species of birds that are in danger of disappearing from New Jersey and Pennsylvania; and the steps that have been taken to preserve them. Please gather these information on birds and at least one related plan to preserve each species.

Aspects:

In the time allotted, please find as many different kinds of birds and plans /projects described above as you can. Only species most likely to disappear are relevant. In addition, the most recent information is the priority. A relevant item should specify the involved bird and the steps taken to save it. A document dealing with a bird without mentioning any steps to solve the problem is not relevant. Description of projects

without mentioning any specific kind of bird is not relevant either. Please save at least one document per each different plan/project (one project per a species is enough). If a document discusses several such plans/projects, then you need not save other documents that repeat those. If one plan/project covers several kinds of birds, please regard each as a different aspect. Projects conducted in different states for the same bird should be treated as different aspects as well.

Topic 4

Description:

You are an owner of a company that opened this month. Since the enactment of the Pollution Prevention Act, industries in the U.S. have tried to reduce the use and generation of hazardous substances. You want to review the kinds of actual benefits such industries can acquire from the implementation of pollution prevention since the year 1991. Please identify as many different specific benefits as possible.

Aspects:

In the time allotted, please find as many different benefits of the sort described above as you can. Save at least one document for each different kind. If one document discusses several different kinds, then other documents that repeat those are not needed. Either financial or non-financial benefits are relevant. The relevant information should be limited to the ones in the U.S. after year 1991, so documents without exact year or location are not relevant.

Appendix B. Questionnaires

PRE-SEARCH QUESTIONNAIRE

Here is the information task you will search:

Topic X (Topic description here)

Please answer the following questions, as they relate to this specific information task.

1. How interested are you in this specific topic?

None			Some			A great deal
1	2	3	4	5	6	7

2. Have you done any searching for information on this specific topic before (online & offline)?

None			Some			A great deal
1	2	3	4	5	6	7

3. Please indicate your level of expertise with this specific topic.

Novice						Expert
1	2	3	4	5	6	7

POST-SEARCH QUESTIONNAIRE

Please answer the following questions, as they relate to this specific information task.

	Not at all			Some-what			Extremely
1. Was it easy to get started on this search?	1	2	3	4	5	6	7
2. Was it easy to do the search on this topic?	1	2	3	4	5	6	7
3. Did you have enough time to do an effective search?	1	2	3	4	5	6	7
4. How satisfied are you with the results of this search?	1	2	3	4	5	6	7
5. Did your previous knowledge help you with your search?	1	2	3	4	5	6	7
6. (only for Mediated search) Did the source collection help	1	2	3	4	5	6	7

in formulating your queries?							
------------------------------	--	--	--	--	--	--	--

POST-SYSTEM QUESTIONNAIRE

	Not at all			Somewhat			Extremely
1. How difficult was it to learn to use this information system?	1	2	3	4	5	6	7
2. How difficult was it to use this information system?	1	2	3	4	5	6	7
3. How well did you understand how to use the information system?	1	2	3	4	5	6	7
4. How useful was the information system in helping you accomplish your search tasks?	1	2	3	4	5	6	7

5. Compared with your usual searching on the Web, how much time did you spend for information searching?

Much less	Less	Somewhat less	No difference	Somewhat more	More	Much more
-----------	------	---------------	---------------	---------------	------	-----------

EXIT QUESTIONNAIRE

	Not at all			Somewhat			Completely
1. How different did you find the systems from one another?	1	2	3	4	5	6	7

Why do you say that?

2. Which of the two systems did you find easier to learn to use?

1 st system	No difference	2 nd system
------------------------	---------------	------------------------

Why do you say that?

3. Which of the two systems did you find easier to use?

1 st system	No difference	2 nd system
------------------------	---------------	------------------------

Why do you say that?

4. Which of the two systems did you think support your searching better?

1 st system	No difference	2 nd system
------------------------	---------------	------------------------

Why do you say that?

5. Which of the two systems did you like the best overall?

1 st system	No difference	2 nd system
------------------------	---------------	------------------------

Why do you say that?

6. What did you like about each of the systems?

1st System: 2nd System:

7. What did you dislike about each of the systems?

1st System: 2nd System:

Only for combined group	Not at all			Somewhat			Extremely
8. How useful was the display of search results in helping your searching?	1	2	3	4	5	6	7

9. Did you feel any difficulty in understanding how the display of search results was organized?	1	2	3	4	5	6	7
--	---	---	---	---	---	---	---

(8) Why do you say that?

(9) Why do you say that?

10. Please list any other comments that you have about your overall search experience on the back of this page. THANK YOU.