VISTO for Web Information Gathering and Organization

Anwar Alhenshiri
Dalhousie University
6050 University Avenue
Halifax, NS, Canada
+1 (902) 494-2093
anwar@cs.dal.ca

Michael Shepherd
Dalhousie University
6050 University Avenue
Halifax, NS, Canada
+1 (902) 494-1199
shepherd@cs.dal.ca

Carolyn Watters
Dalhousie University
6050 University Avenue
Halifax, NS, Canada
+1 (902) 494-6723
watters@cs.dal.ca

ABSTRACT
This paper presents a Visual Search Task Organizer (VISTO). VISTO is a visual tool with effective information gathering task capabilities for the Web. In this prototype system, the task of Web information gathering is taken into consideration with respect to how user locate information for the task, organize task information, preserve and re-find task information, and compare information for effective reasoning and decision making. VISTO was designed and built based on recommendations from previous studies in a large research. The prototype is ready to be evaluated in the next step of the research.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Clustering, Search Process.
H.1.2 [User/Machine Systems]: Human Factors
H.5.2 [User Interfaces]: User Centered Design.

General Terms
Human Factors, Theory, Measurement.

Keywords
Information retrieval, Web search, user tasks, Web information gathering, Web information organization and management

1. INTRODUCTION
Web information retrieval has been studied in the light of request-response for a relatively significant period of time. The user submits a query trying to convey their information need to the Web and in return, they receive a response from the search engine in the form of document hits. In many occasions, a search activity may require that the user continues interacting with the search engine to achieve a higher-level Web task [15]. Research has studied user tasks in order to identify a task framework that would help with understanding user interactions with the Web. Web tasks have been classified into fact finding, navigation, performing a transaction, and information gathering [7, 14]. The latter type accounts for a large portion of the overall tasks on the Web, representing between 51.7% [7] and 61.5% [19].

A Web information gathering task is a composite of subtasks/activities users perform while interacting with the Web for accomplishing a goal described in the task. User activities during Web information gathering may involve finding sources of Web information (Web documents), searching for information on the sources located for the task, finding related information to the already located sources and information, comparing information for reasoning and decision making, organizing task information, and preserving and re-finding information [2]. Figure 1 shows these subtasks in the Web information gathering task. For example, when planning for a trip to a foreign country for the first time, one has to gather different kinds of information to accomplish the goal of the trip. The plan may include looking for sources of information regarding currency exchange rates, flight rates, hotel rates, and so on. While trying to find information regarding each criterion, locating further information and comparing information are possible activities. Re-finding information during the process of information accumulation is an activity that may occur at any time. Moreover, information organization in addition to reasoning and decision making are required for ensuring the validity of the information located for the task and the degree to which the task goal is satisfied.

Figure 1. The Web information gathering task [2]
Since information mismatching and overloading are two significant problems regarding the way search engines gather and present information [23], it becomes the user’s role to locate, compare, and manage the required information in the task. A Web search engine sees the sequences of a task as separate interaction steps. It also provides no means for re-finding information, which is an activity that represents one third of the user interactions during information gathering tasks according to Kellar and Watters [13]. Moreover, search engines do not usually provide support for representing task results according to the type of information being sought in the task. Consequently, there is a very
limited understanding by the design of current Web search models of the fact that a search activity may not be only a one-time query, but rather a more complete and sophisticated task.

With regard to information gathering tasks, there has been no specific focus in the literature on the effects of visualization, clustering, and the concepts of information re-finding and information organization on the effectiveness of users performing Web tasks for information gathering. Visualization and clustering have been investigated for improving the effectiveness of Web search techniques in general. In addition, re-finding is a factor that has been studied in isolation, and there are several techniques intended for improving re-finding of information on the Web either locally on the Web browser or, sometimes, on the entire Web through the use of a search engine. However, information organization is barely studied in the context of the Web. In this article, a Web information gathering and organization prototype (VISTO) is presented. VISTO exploits visualization, visual clustering, and several Web information preserving, re-finding, and organization strategies for Web information gathering tasks.

The paper is presented as follows. Section 2 discusses related research work. Section 3 presents principles and motivations for the design of VISTO. Section 4 describes VISTO in details. Section 5 discusses further research directions. Section 6 concludes the paper.

2. RESEARCH RATIONALE

2.1 Visualizing Web Information

Visualization is a concept that has been in focus for research in information retrieval [15, 20]. Information visualization is suggested to improve users' performance by harnessing their innate abilities for perceiving, identifying, exploring, and understanding large volumes of data. There are several visualization-based prototypes that have been investigated for improving the effectiveness of Web search [5, 6, 20]. For example, Teevan et al. [23] investigated the use of visual snippets in Web search results presentation and compared the effectiveness of this approach to the conventional text snippets provided by search engines such as Google. In addition, Bonnel et al. [6] showed that users favored visual clustering in a 3D City Metaphor. Visual thumbnails of Web search results that accompany textual presentations were also shown to be effective in searching the Web for revisiting [22].

There are several search tools on the Web that use visualization such as the search engines Gceel (www.Gceel.com), Nexplore (www.nexplore.com), and Viewz (www.viewzi.com). Visualization of Web search results has also been investigated in several layouts including the use of hyperbolic trees, Scatterplots, Self-Organizing Maps, and thematic maps such as in the visual search engine Kartoo (www.kartoo.com). Most of these approaches were intended for improving how users find sources of Web information. Exploring multiple features of Web documents such as their content similarities, page thumbnails, URLs, and document summaries in a visualized approach should be exploited in Web information gathering tasks. These features—when visualized properly—can help users find sources of information on the Web, find information in such sources, compare information, and make more effective and efficient decisions during the task accomplishment process.

2.2 Clustering Web Information

Clustering is intended for grouping together items that share similar characteristics and attributes. In Web information retrieval, clustering is meant for grouping similar documents [18]. The use of clustering has been widely investigated in Web information retrieval [18]. Clustering is usually intended to provide overviews of information categories (topics) in the result set. Hence, efficient subtopic retrieval is anticipated with the use of clustering in Web search results presentations [8]. When more than one topic is desired while gathering information on the Web, clustering may provide effective topic exploration in the high-level views of the result hits. Clustering can also decrease the need for scrolling over multiple pages of results and motivate users to look beyond the first few hits, resulting in more effective and efficient user performance.

In Web information retrieval, clustering has been investigated in several prototypes such as in the work of Alhenshiri et al. [1]. Clustering has also been implemented in conventional search engines such as Clusty (www.clsuty.com), Gceel (www.Gceel.com), and Google (in their “see similar” feature and Google Wonder Wheel). Although the performance of users with row presentations of Web documents is comparable to their performance with clustering-based presentations, user preference usually comes in favor of clustering-based methods [8]. In addition, there are indications that clustering can even be more effective [24]. With the variety of information that is gathered on the Web, clustering can play a significant factor in Web information gathering tasks. Clustering should be investigated with regard to finding related information to the task during the gathering process.

2.3 Preserving and Re-finding Web Information

Research has focused on enhancing re-finding Web information locally on the Web browser. However, re-finding strategies such as the back button, favorites, and bookmarks can maintain limited numbers of information sources. In addition, the use of those strategies is limited to pages and sites of interest during particular Web sessions. Therefore, searching the Web for re-finding, also known as re-searching [22], has been studied for assisting users in locating results of interest that were found interesting in previous sessions. Research shows that a great deal of Web search visitations is for revisiting [22]. Consequently, Re-finding is a common activity in Web information gathering tasks accounting for (53.27%) according to Mackay and Watters [17]. For information gathering tasks of multi-session nature, which may require multi-topic search, re-finding can play a significant role in the effectiveness of tools designed for this type of task. Re-finding should be focused not only on preserving active Web pages in the browser but also on Web search results. In addition, research should further explore the role of re-finding in the context of a complete information gathering task.

2.4 Organizing and Managing Web Information

Research has focused on investigating how users manage their information for re-finding [9, 10, 16]. Important reasons behind giving up on certain personal information management tools were
discussed in the work of Jones et al. [11]. Strategies users follow to manage Web information in order to be able to relocate and reuse previously found information are discussed in the work of Jones et al. [10]. Their work shows that users—while gathering Web information—follow different preserving strategies to re-find and compare information. Most users gather information over multiple sessions [17], which indicates the need for management strategies for preserving and re-finding such information for reuse. The variety of finding, re-finding, organizing, and management strategies and approaches users follow while seeking and gathering Web information can be related to the fact that current Web tools lack important reminding, integration, and organization schemes.

Research has had little consideration to factors that would improve how Web users collect, manage, compare, and organize information for gathering tasks. On the Web, research has only considered the case of managing and organizing information for re-finding [10]. How users organize and manage information during Web information gathering has had little consideration. Since Web information gathering tasks may take several sessions, involve looking at information from different sources, and require comparing information that may belong to varied topics, investigating organizational and management strategies users follow on the Web is necessary.

3. VISTO Principles and Motivations

To further exploit the concepts of information visualization, visual clustering, re-finding, and organization, VISTO was designed. The prototype offers the following features for supporting information gathering tasks:

a. Effective visualized search.
b. Intuitive visualized clustering.
c. Effective Web information organization.
d. Effective preserving and re-finding strategies.

VISTO was designed based on the recommendations drawn from the studies in [1, 2, 3, 10, 23]. Research has studied visualization and clustering for improving the effectiveness of general Web search tasks. In addition, re-finding has been investigated for providing effective techniques that allow users to relocate previously preserved Web documents. What is lacking in research regarding Web information gathering tasks is threefold. First, visualization and clustering should be investigated for improving the process of accomplishing the whole task. VISTO attempts to utilize visualization and clustering to allow users to find, compare, and relate information to the already located sources of information more effectively.

Second, re-finding has been studied only for permitting users to effectively and efficiently re-find Web documents that were preserved in previous sessions. VISTO attempts to create a more effective storing and re-finding environment. This is done by allowing users to store individual documents as well as complete sessions. In addition, re-finding is done not only by searching a list of documents, but also by using keyword search to re-find individual documents, sessions, and whole tasks previously preserved by the user. Third, Information organization has been studied in the desktop environment for personal information management. The Web has had little consideration except in the case of information management for re-finding [10]. VISTO takes a step further and attempts to provide effective organizational schemes for information during information gathering tasks.

4. VISTO Design

The VISTO interface was designed using Java swing components and the prefuse visualization toolkit (http://prefuse.org/). A snapshot of the VISTO interface is shown in Figure 2. There are four different models, shown in Figure 3, employed in the design of the VISTO interface as follows:

4.1 Search Model

VISTO provides search services to users gathering Web information. It combines the powers of Google and Yahoo Web search engines. When a user submits a query, VISTO uses Google’s spell correction service and then submits the query to both search engines. VISTO eliminates repeated hits and prepares the results for display. Three main features are provided to the user in results displayed by VISTO. First, context is provided by
presenting document titles on visual glyphs. Document glyphs are clustered so that relations among search results are conveyed to the user. Second, look-ahead is provided by VISTO. The user can see the document thumbnail and summary to predict the page content. Third, focus is provided to the user by hovering over the visual glyphs. The hover-over feature narrows the focus of the user to the specific cluster/document so that the user can see the document summary and URL. Moreover, the user can eliminate clusters and individual glyphs from the display for reducing clutter and achieving more focus.

4.2 Clustering Model
VISTO uses intuitive visual clustering to render its search results and its preserved task and session documents. Clustering is performed based on one of four criteria. First, network domain clustering permits the user to see visual clusters of Web documents categorized based on network domains such as commercial, organizational, and educational domains. Second, the user can select clustering by location in which documents are categorized based on the country of origin. Third, VISTO permits clustering by content similarity (topical clustering) in which similar documents are grouped together. Clusters are labeled using cluster-internal labeling [18]. The title of the document closest to the centroid of the cluster is used as the label of the cluster. Last, VISTO provides clustering by genre. Documents that belong to the same genre are grouped together. Currently, VISTO uses 17 of the 25 genre types described in the work of Santini et al. [21]. For computing genre-based similarity, the structure and the content of the documents are taken into account.

4.3 Organizing Model
To assist users with organizing task information, VISTO uses different strategies. Based on the study described in [3], users find it hard to keep track of task information especially in multi-session information gathering tasks. Therefore, VISTO allows the user to store partial information during a task by preserving current session information. This is done by either preserving active visual views of the current display or by selectively preserving particular documents among the search results. Preserved documents are grouped under a task title (name) and sorted by date for later retrieval. The user can continue working on the same task over multiple sessions while adding and eliminating documents. The user can add annotations to the preserved task information along the way towards completing the task. To further assist users with organizing the task information and making effective decisions regarding the task, two different views of the task gathered information are available in VISTO. Search results and accumulated task information (documents) can be viewed either visually or in HTML format. The study in [1] showed that users prefer to have both views during information gathering. VISTO provides continuous task information accumulation strategy so that users do not lose track of their information.

4.4 Re-finding Model
Re-finding is a heavily studied subject in Web information retrieval. In VISTO, re-finding for information gathering and organization is emphasized best. VISTO allows users to store complete sessions and individual documents for re-finding. It also allows search within sessions and within tasks by either selecting from a list of tasks/sessions or by keyword search to further assist the user. The keyword search matches the task name given previously by the user and the annotations preserved alongside the task. Moreover, VISTO allows users to email task information including accumulated documents, task subject and date, and task annotations. The emailing strategy was recommended in the work of Jones et al. [10]. However, VISTO adds the organization of a task to the subject matter by submitting all the aforementioned information items. With VISTO, the user can follow the preserving and re-finding strategy that suits their needs and accommodates the task requirements.

5. Discussion
VISTO is a prototype system ready to be tested on the Web for information gathering tasks. A complete factorial study will be conducted for evaluating the efficiency, effectiveness, and engagement of VISTO in performing Web information gathering tasks. In this study, participants will be asked to perform the tasks using VISTO and other Web tools. The study will evaluate effectiveness with respect to the relevancy and accuracy of task information, the number of queries required for the task, the number of sequences/steps needed during the task, and the completeness of the task requirements. Efficiency will also be measured with regard to the time on task. User engagement will be measured using the user confidence in the task results, their interest in the tool, and other self-reported comments. The results of the study will provide practical research recommendations for Web tools intended for Web information gathering tasks.

6. Conclusion
This paper presented VISTO, a prototype system for improving the effectiveness of gathering and organizing Web information. The current state in Web information gathering necessitates studying challenges users encounter during this type of tasks. Based on three previous studies [1, 2, 3], VISTO was designed. Our previous studies revealed several questions regarding which visualization, clustering, re-finding, and organizing factors would improve the process of Web information gathering. The research will continue with a user study to evaluate VISTO.

7. REFERENCES


