Building Support for Web Information Gathering Tasks

Anwar Alhenshiri
Dalhousie University
alhenshiri@dal.ca

Carolyn Watters
Dalhousie University
watters@cs.dal.ca

Michael Shepherd
Dalhousie University
shepherd@cs.dal.ca

Jack Duffy
Dalhousie University
jack.duffy@dal.ca

Abstract

Web information gathering includes tasks in which users attempt to locate, organize, and use information from different sources on the web to satisfy an information need, often over multiple sessions. An example is the case of writing a report or beginning an investigation. This paper presents the results of a research study to explore the difficulties users experience during such tasks. Twenty users participated in the study in which the frequency of specific kinds of activities involved was recorded along with qualitative measures. The results of the study provide insight into the area’s most needing support in order to provide guidance for the building of new tools to support users in these tasks. The results indicate that tools are needed to bring information into a final report, facilitate editing, and to support re-finding information in its original context over multiple sessions.

1. Introduction

Due to the enormous amounts of data posted and updated on the web on a daily basis, the web has become a prime source of information for a wide variety of tasks. Even with the continuous evolution of the web and its use, the browsing and search models have not changed much for users in the context of specific tasks. In this paper we concentrate on information gathering tasks. An information gathering task typically requires that the user perform several activities to achieve his or her goal to collect information from different sources to satisfy an information need. The support provided in current browsing and search models is largely focused on browsing and searching within single sessions.

Recent research has been focused on identifying the types of tasks users perform and formalizing taxonomies and frameworks to understand the nature of each task. For example, user activities on the web have been categorized into high-level tasks (Broder, 2002; Rose and Levinson, 2004; Kellar et al., 2007). Depending on the classification, the most frequent tasks performed on the web include information gathering, navigation, fact finding, and performing a transaction. The generic task of web information gathering can be divided into subtasks commonly performed during the task (Alhenshiri et al., 2010). Major subtasks include searching for information and organization of information for subsequent use. The search subtask, of course, can be further refined to include locating sources of information, locating specific information within sources, linking to related information, and re-finding information over one or more sessions. Information gathering tasks account for a large proportion of web tasks (61% (Rose and Levinson, 2004) to 65% (Broder, 2002)).

While search engines and search tools continue to evolve, managing and organizing information during information gathering tasks on the web is a critical subtask of information gathering that has been given less consideration. Users gathering information on the web find and compare information from different sources and different genres selecting some of that information for further analysis and reporting purposes. Not surprisingly, users adopt different approaches for coping with the management and organization aspects of information gathering (Alhenshiri et al., 2011). Dealing with different sources of information, some current and some archived, requires effort on the part of the user to keep track of the task progress. Two questions then arise. First, how effective are the current web tools for managing and organizing information during information gathering tasks on the web? Second, what can be provided by way of guidelines to lead the development of better tools?

The study presented in this paper examines how browser features and other tools are used in web information gathering tasks so that guidelines for the development of future web gathering tools can be established. The study examines the tools used during web information gathering tasks, the organization and management strategies of the user, and the difficulties the user encounters during the execution of the tasks on the web. From these data, we provide recommendations for improving the support provided
by browsers for the management and organization of information found during gathering tasks.

The remainder of this paper is organized as follows. Section 2 discusses the research rationale. Section 3 provides the research study. Section 4 discusses the research findings. The paper is concluded in Section 5.

2. Research rationale

Information seeking models have focused on identifying activities users perform while they attempt to locate information of interest. The web has been treated as a special case in some of the older models such as Ellis’s (1989) model. Marchionini (1995) stated that the process of information seeking consists of several activities (sub-processes) that start with the recognition and acceptance of the problem and continues until the problem is either resolved or abandoned. Wilson and Walsh’s (1996) model of information behaviour differs from earlier models by suggesting more high-level information seeking search processes: passive attention, passive search, active search, and ongoing search. Although those models characterize users information seeking activities, several activities that users perform on the web are not included in these models.

Several other frameworks have been suggested to understand and model the different activities users perform on the web while seeking information. Rose and Levinson (2004) identified a framework for user search goals using ontologies in order to understand how users interact with the web. Their findings indicated that user goals can be informational, navigational, or transactional. Similarly, Sellen et al. (2002) found that user activities can be categorized into finding, information gathering, browsing, performing a transaction, communicating, and housekeeping. Moreover, Broder (2002) studied different user interactions during web search and identified three types of tasks based on the queries submitted by users which are: navigational, informational, and transactional. In addition, Kellar et al. (2007) investigated user activities on the web to develop a task framework. The results of their study indicated that the four types of web tasks are: fact finding, information gathering, browsing, and performing transactions.

Based on the different classifications of web tasks, research has shown that information gathering tasks represent a significant proportion of the tasks performed on the web (65% according to Broader, 2002). Amin (2009) identified different characteristics in web information gathering tasks. Information gathering was shown to be a more complex task than keyword search tasks. Information gathering includes different kinds of search involving comparison, relationship, exploratory, and topic searches in addition to combinations of more than one type of search (Amin, 2009). Information gathering tasks have high level goals and require using multiple information resources.

Alhenshiri et al. (2010) defined information gathering as a composite of subtasks that include: finding information sources, finding information on the located sources, finding further related information, and preserving and re-finding information. In all of these subtasks, managing and organizing the task information is a subtask that may take place at different times during the task of information gathering. Consequently, we are specifically interested in these subtasks.

Managing web information is concerned with how people keep, organize, and re-find web information (Elsweiler and Ruthven, 2007). Research has considered personal information management with less focus on the web. The web provides a context in which more information can be located, and subsequently requires storing, and relocating at later times. Other research has focused on how users manage their information for re-finding (Jones et al., 2003; Mackay et al., 2005; Elsweiler and Ruthven, 2007).

Jones et al. (2003) studied the strategies users follow to manage web information to relocate and reuse information previously found. Their work showed that users, while gathering web information, follow different preserving strategies to re-find and compare information later. For information gathering tasks, most users gather information over multiple sessions (Spink et al., 1996; Mackay and Watters, 2008), indicating a growing need for strategies and tools for keeping and re-finding such information for reuse. The variety of finding, re-finding, organizing, and management strategies and approaches users follow is an indication that current web tools lack usable reminding, integration, and organization features (Cutrell et al., 2006).

Jones et al. (2008) found that users abandon the use of an information management tool for one or more of the following reasons: visibility, integration, co-adoption, scalability, and return to investment. These reasons need to be further investigated in the context of web information gathering. Such an investigation may reveal further reasons why users use certain tools over others, why they do not use the same tools, what tools most users actually use to keep track of their gathered information, and how they maintain the consistency of information located for the task. Other questions include the identification of tools that are supportive of information organization and management during
information gathering. Practical recommendations for supporting the design and implementation of web information gathering tools with respect to information organization and management are needed.

Schraefel et al. (2002) described a model to allow users to have more focus on the task of information gathering by reducing the effort needed for managing information during the gathering process. During information gathering, the user is likely to collect pieces of information from web pages, i.e., the subtask of finding information on located sources as discussed above. This procedure requires the user to find the sources (pages) first and then copy or summarize information from those pages to be edited and formatted using other tools such as text editors (information management). The ‘Hunter-Gatherer’ (Schraefel et al., 2002) allowed the user to do both subtasks in one tool. That is, to locate pages that contain the information of interest and copy and edit the information required by the task in the same tool.

In a field study, Elsweiler and Ruthavan (2007) asked participants to describe their re-finding tasks. Tasks were either related to email or web re-finding. Three types of tasks were recorded: lookup tasks, item tasks, and multiple item tasks. Lookup re-finding was meant for locating a piece of information about which the user may or may not know. Item re-finding tasks were concerned with a single piece of information that the user knows and has encountered before. Multi-item re-finding was meant for locating multiple pieces of information that are known to the user. In addition to creating a model for re-finding tasks, the experiment lead to developing a common approach that research can utilize to create simulated tasks based on realistic ones. Moreover, Elsweiler and Ruthavan (2007) found that users had increased difficulty in performing a re-finding task as the time increased since the information was originally found.

Mankowsky and Watters (2011) developed Webscraprs, a technique intended to help users gather and later re-find parts of pages during web information gathering tasks. WebScraps focused on the use of specific parts of web pages in which information of interest to the task are found, which is different from thumbnails that capture the entire page with a less detailed view of the page. The study showed that WebScraps increased the user’s ability to remember information they already gathered and kept, were more effective for comparing information than thumbnails, and they were more helpful and enjoyable.

While research has focused on managing and organizing information for re-finding purposes, less consideration has been given to how users organize and manage information during web information gathering tasks. Understanding the behaviour of users in information gathering tasks will provide insight into the design characteristics of tools to support this activity. The research study described in Section 3 had this goal.

3. Research study

A user study was conducted to examine how users manage and organize information during web information gathering tasks. The study was intended to reveal difficulties the users encountered as well as the tools and strategies used specifically for the organization and management phase of the task. Understanding the effectiveness of current tools and strategies provides a base for recommendations for improving these important subtasks of information gathering on the web. The study was designed to answer the following three questions: What activities do users perform during information gathering tasks and at what frequency? What tools and strategies do users use to manage information for gathering tasks? What difficulties do they encounter while managing and organizing the task information? Answering these questions will be used to provide design recommendations for the development of new tools for managing and organizing information during web gathering tasks.

3.1. Study design and population

The design of the study was complete factorial and counterbalanced. There were 20 participants in the study of which ten were graduate students while the remaining ten were undergraduate students. All participants were from the Faculty of Computer Science at Anonymous University. The study used a special version of the Mozilla Firefox browser called ‘DeerParkLogger’, designed at Anonymous University. DeerParkLogger allows the researcher to log all browser interactions during the task, as well as gather other user input during the study period. Each user was asked to complete two information gathering tasks.

3.2. Study tasks

The study used four different information gathering tasks, each with two components. Each task was created following the guidelines described in Kules and Capra (2008) summarized by the following characteristics:

http://www.mozilla.com
The task should indicate uncertainty or ambiguity in information need, or need for discovery.

The task should require knowledge acquisition, comparison, or discovery.

The task should provide a low level of specificity about the information required in the task and how to find such information.

The task should provide enough imaginative contexts for the study participants to be able to relate and apply the situation.

A focus group of sixteen graduate students was used to ensure that the tasks were of the same level of complexity. Over two meetings, the focus group evaluated the equivalence of the tasks with respect to the following criteria.

- The time needed to complete the task.
- The amount of information that is required to be gathered in the task.
- The clarity of the task description.
- The possible difficulties that may be encountered by the user during the task.

Four tasks, each with two parts, were then chosen for the study. An example of one task is shown in Figure 1.

### Information gathering task example

**Task 1a**
You have a friend who asked you to provide her with valuable information about Canadian universities that she may consider for a graduate degree in business. What kind of information would you like to send to your friend providing her with a comparison of two universities? Provide your choice of the universities. Provide links to at most five web pages you find helpful in making your choices. Also, provide a copy of the information you would send to your friend which shows the comparison. You will need to come back to reuse the information you found in this task.

**Task 1b**
Last time, you selected two Canadian universities for your friend to pursue a graduate degree in business. Now, your friend asked you to provide her with two choices of American universities to consider for a graduate degree in business. Choose two American universities with graduate degrees in business and then find up to five websites that would allow you to make a comparison between the Canadian universities you already selected and these two American universities. Provide the results of your comparison (information you used in the comparison) in addition to links to at most five web pages you find useful in making the comparison.

### 3.3. Study methodology

Each participant was randomly assigned two of the four tasks, referred to as Task1 and Task2 each of which was divided into two related tasks, the second following up on the first, referred to as Task1a and Task1b, for example. The reason for splitting each task into a sequence of two related tasks was to provide a context in which participants would find some advantage in reusing or re-finding information for Task1a that was found during Task1b. The participants performed the study in the sequence shown in Figure 2. A pre-study and two post-task questionnaires were used in the study.

### 3.4. Study results

The focus of the analysis was on behavior related to information management and organization during the execution of the study tasks. The following data were recorded: the use of browser tabs, the use of the search feature on web pages, the use of bookmarks, the use of copying and pasting information from web pages, typing information into documents, and the time on task.

#### 3.4.1. Task effect

To ensure the equivalency of the tasks with respect to the level of complexity, the differences in the overall task performance among tasks were measured. This analysis used the time spent on the task, the user confidence in the task results, and the degree of difficulty of the task as perceived by the participants and stated in the questionnaires. The ANOVA results for these three categories indicate that there was no significant differences between the tasks used in the study (task time, $f=0.98$, $p=0.33$; user confidence, $f=1.97$, $p<0.14$; task difficulty, $f=0.97$, $p=0.33$).
This confirmed the results of the focus group prior to the study.

3.4.2. User activities. To identify the different kinds of activities the users performed while managing and organizing information for the task, each instance of each activity was recorded in the log file. Participants performed an average of 125.2 activities (σ=41.9) over both Task1 and Task2, ranging from 42 to 198 activities per user. These activities included: opening URLs, using URL auto-completion, submitting search queries, following links on pages, bookmarking, using the find-on-page browser feature, using browser tabs, using browser windows, copying and pasting information from web pages into files and emails, typing information for the task, and using the browser history. In the second session or part of each task, the users were asked to use information from the first session.

A subset of those activities used for managing and organizing of information during the tasks was selected for further examination, including, the use of browser tabs (and windows), the use of text editors and copy-paste, re-finding links or information, the use of the browser history, the use of bookmarks, and the use of emails. Table 1 shows the frequency of each of these activities during the tasks. The results indicate that among the activities related to managing and organizing information during web information gathering tasks, using browser tabs and copying and pasting information represented a large portion (45% of the overall activities) of the total number of activities.

Table 1. Activities performed by users during the study

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using a browser tab</td>
<td>501</td>
</tr>
<tr>
<td>Following a link</td>
<td>416</td>
</tr>
<tr>
<td>Copying and pasting</td>
<td>241</td>
</tr>
<tr>
<td>Submitting a query</td>
<td>166</td>
</tr>
<tr>
<td>Typing in a URL</td>
<td>80</td>
</tr>
<tr>
<td>Bookmarking</td>
<td>71</td>
</tr>
<tr>
<td>Finding information on a page</td>
<td>44</td>
</tr>
<tr>
<td>Auto-completing a URL</td>
<td>41</td>
</tr>
<tr>
<td>Using a browser window</td>
<td>37</td>
</tr>
<tr>
<td>Typing information for the task</td>
<td>20</td>
</tr>
<tr>
<td>Searching on a site</td>
<td>10</td>
</tr>
</tbody>
</table>

3.4.3. Organization tools. The results showed that over all the tasks performed in the study, participants used text editors 26 times (MSWord 13 times, Notepad 13 times) and emailing (5 times) to edit and manage task information. There was a significant different between the use of text editors and the use of emails in the study (z =1.95, p<0.03). The results indicate that users rely heavily on editing to manage information between the parts of these tasks. Interestingly although users used these tools, 95% (19 users) indicated in the interview that none of the tools used for organizing the task information was sufficient for the purpose of the tasks. The reason they gave for using these tools was that such tools were the only ones available.

3.4.4. Organization strategies. To organize and manage information for the task, users employed different strategies including: the use of bookmarking, editing information in text editors, use of email features, keeping information in email drafts and local files, re-finding information (documents and parts of documents) in emails and files, temporary storage of pages on open browser tabs and windows, copying and pasting information, typing information located in web pages and search results summaries, and temporary storage of URLs for comparisons and decision making.

3.4.4.1. Use of Bookmarks. Even though bookmarking is one of the most well-known features in the web browser for keeping links in order to re-find web pages, the users in the study did not use this feature much for the tasks of information gathering in the study. As Figure 3 shows, 65% (13/20) of the users did not use bookmarks whatsoever. After removing the outlier, the average use of bookmarking in the study was 1.1 (σ=1.8). When questioned in the interview, the users indicated that it was hard to re-find pages kept in a previous session and compare them to pages actively open in the browser during the current session. Users also indicated that completing the task required more than just keeping links to pages but also required access to parts of pages and other information such as annotations created by the user.
3.4.4.2. Editing information in text editors. Users in the study relied heavily on the use of text editors including MSWord and Notepad to organize the task information they had found. Of all uses of tools used during the tasks, 55% were text editors. Participants made use of text editors while browsing, searching, comparing, and selecting information from web pages.

All of the users in the study used text editors to edit and format information they found in web pages. Most users (12/20) extracted information from result hit summaries for editing and further formatting in text editors. The remaining users (8/20) copied information from web pages that were actively open in the browser and pasted the information into text editors for further organization. In all cases, users reported that the process of accumulating, organizing, editing, and formatting task information in text editors and email drafts was demanding. However, they indicated that using text files made it easier to manage and remember information than using the browser history or bookmarks for keeping track of web pages.

3.4.4.3. Use of Email Features. The use of email drafts was mainly for re-finding and making notes on the task requirements in the study. Only three participants in the study (shown in Figure 4) used emails for either saving information for later or to explicitly satisfy the task requirements. In this case, the email composing utility was basically used as a text editor. Users indicated that they used emails because it would allow them to find the information faster for later reuse in multi-session tasks due to the easiness in searching content in the email account. Those users stated that using text editors and creating text files would result in difficulties in finding the information later since this process implies that some folder and file structure organization would be needed.

3.4.4.4. Using browser tabs and windows. To manage tasks during a session, the users relied heavily on the use of browser tabs. On average, users opened 21 tabs over both tasks (σ =22.5). Figure 5 shows the results of the browser tab usage. Of the participants, 95% (19 users) used at least two browser tabs at least once to keep track of open pages, to compare information, and to keep information for later use within the same session. They indicated that they used tabs rather than windows or bookmarks because they could see the information faster by switching among tabs, they were able to keep all pages open, and they were able to compare the information for decision making. Nonetheless, eight participants (40%) indicated that they lost track of pages opened and they used only a few of the tabs they opened during the task. When asked, users indicated that this was the best available strategy offered by the current browsing model for managing pages during a session.

3.4.4.5. Copying, pasting, and typing task information. While most users used copy-paste features, the use of this feature had two variations.
40% of the participants (8/20) used copying and pasting information (as shown in Figure 6) from web pages into text files, emails, and so on. These users also frequently opened pages by clicking links to find information. The interviews revealed that it was convenient for those users to copy and paste the information directly from web pages. The remaining users (12/20) rarely used this feature. Instead, they opened more browser tabs to submit many more queries and extract or type into their word processor or email information from hit summaries provided by the search engines rather than from the web pages themselves.

The use of text editors for keeping and gathering information argues for adding editing features into the current browsing model to improve how user gather and organized web information. Of the participants, 13 users (65%) expressed interest in having editing coupled with gathering capabilities embedded in the web browser. One user indicated that she wanted ‘a browser with Notepad’ as a form of editing and formatting while gathering.

3.4.4.6. Keeping links in emails, files, and bookmarks for re-finding.

Users kept links along with information of interest in text files and email drafts. They reported a variety of reasons for saving links, shown in Figure 7. The most common reasons for keeping links were to reference content already found (50%), to revisit a page (42%), and for comparisons and decision making (8%).

![Figure 7. Why users kept links](image)

Of the participants, 95% (19/20) stated in the interview that they preferred to keep links in documents rather than in bookmarks so that they could have the links related to a task in one place. Users also reported using those links as references to information, to remind themselves where to find more information for the task, and as sources for comparing information. One user indicated that she would like to have ‘automated referencing’ to embed links in the gathered information.

3.4.5. Information organization and management related difficulties.

In the post study questionnaire, users reported that they had encountered problems related to the organization of information (as shown in Figure 8). The problems reported by the participants were categorized by a group of eight graduate students. The most frequent categories identified were organizing, storing, prioritizing, and filtering information. These categories included problems such as, storing and organizing information, formatting information, filtering through information, copying and pasting information, editing information required in the task, switching among pages opened on browser tabs during the task, and prioritizing sources of information for selecting the task requirements.

![Figure 8. Difficulties encountered during the study](image)

The results of the study showed that for some tools, availability rather than effectiveness was the reason why participants used such tools. The main concern expressed by the participants, 50% of the comments, was the inability to edit and browse information simultaneously. Furthermore, eight participants indicated that even though they used tabs to view, compare, and select information on the web, they lost track of pages opened on those tabs. In the study, participants opened an average of 21 browser tabs each during the study.

Participants reported other difficulties including: switching among open browser tabs and windows, and having to read much text. Only 15% of the users indicated that they had difficulties dealing with more than one session both in terms of giving priorities to the information kept for later reuse and in terms of re-finding such information. Users need tools that
support the simultaneous display of multiple sources of information (documents) for comparison and decision making, that allow for editing information as part of the gathering activities, that permit the user to save information and links for later reuse and re-finding purposes, and that support effective management of the task results.

Although the participants used browser tabs frequently, they still reported that they lost track of the open tabs. They also indicated that multiple browser windows were even less effective since it becomes harder to manage multiple open windows simultaneously. In addition, switching among different applications during the tasks made it harder for the users to organize information. For example, one user had the browser open with multiple tabs in an active state in addition to a calculator and a text editor. This user was also attempting to use her email to keep links to pages she expected to reuse later. When asked about the reasons for not using bookmarks, she said that it would be easier to remember where the information could be found if she used her email. Those strategies highlight the need to reexamine the tools currently available for web information gathering tasks.

3.5. Study limitations

There are three main limitations to the user study conducted in this research. First, the study used simulated tasks. Those tasks do not reflect the range of information gathering task on the web even though the researchers took care to make the tasks realistic and comparable. Second, all of the participants were computer science students, clearly a special population of web users, and the results will not be applicable to the broader demographic of web users. Finally, there are now features and browser add-ons, such as the session manager provided by Google Chrome\(^2\), that might have helped with the tasks and that were not used in the version of the web browser used in the study.

4. Discussion

The goal of the study was to propose guidelines for the development of new tools to support information gathering on the web with respect to information organization and management during such tasks. The results provide the following insights:

- Users want capacity to edit and format web information during sessions.
- Browser tabs used for keeping track of information during a session resulted in losing information and were ineffective.
- Users had trouble managing information with the use of several tools for different activities simultaneously.

Users adopted several strategies to organize their information using varied tools during gathering tasks. Those strategies involved:

- Users stored links into text files instead of bookmarks or favorites. Most users (65%) indicated that they rarely, if ever, used bookmarks.
- Users stored complete pages on their computer for later reuse.
- Users created folders and text files to organize the information both for intermediary use and for the final results.
- Users also had to switch between editing and browsing for organizing and gathering information due to the need for the two activities for this type of task.
- Users used browser tabs for organizing information for comparisons because this was the best available feature on the browser for this activity.
- Users used their email ‘compose’ feature to store links and information for later re-finding.

From the findings of the study, several guidelines that would benefit the design of new tools intended for web information gathering are proposed.

- Web gathering tools should allow users to edit and format information while performing gathering tasks. The design of such tools should allow users to browse information while being able to extract and manage information for further editing and formatting.
- Web gathering tools should allow users to compare multiple sources of information for decision making. This can be done by providing, perhaps, excerpts or thumbnails to capture areas of interest on pages being viewed.
- Web gathering tools should allow keeping and re-finding information through features that preserve the context of the task.

\(^2\) https://chrome.google.com/webstore
In future research, several factors will be examined to evaluate their impact in improving user effectiveness in organizing and managing information during gathering tasks on the web: editing information in terms of copying, pasting, formatting, and typing while searching. Furthermore, an alternative to bookmarks will be examined to study the effectiveness of storing information for re-finding. Users should be able to keep information in different forms depending on the task and the interest of the user. For example, some users may prefer to store not only links to web pages but also parts of the pages and the information from those pages. Others may prefer to store groups of links representing sessions or complete tasks and be able to edit such groups. We need to understand more about the impact of variations on these factors to organize and manage information in the context of web information gathering tasks.

5. Conclusion

This paper presented a user study in which the components of organizing and managing web information during gathering tasks were investigated. The results of the study showed that the current browser design requires additional functionalities related to the organization and management of information during web information gathering. New functionalities would include the ability to edit information during the gathering process, to view and compare multiple pages and/or parts of pages simultaneously, and more effective keeping and re-finding techniques. Users should be able to keep documents, parts of documents, search results, and earlier sessions. Improving the process of organizing and managing web information is a first step in developing web tools that will improve the effectiveness of users gathering web information on the web.

6. References


