

Supporting Webpage Revisiting with History Data and Visualization

Trien Van Do

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy



The University of Leeds
School of Computing

March, 2013

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Trien Van Do to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Declarations

Chapter 5 of this thesis has been based on work from the jointly-authored publication below. All the material in this publication is my own work under the supervision of Dr Roy Ruddle.

Do, T.V., & Ruddle, R. A. (2012). The design of a visual history tool to help users refind information within a website. In: Proceedings of the 34th European Conference on Information Retrieval (ECIR 2012). Springer-Verlag, 459-462.

To my family, teachers, and friends.

Acknowledgements

My PhD has been a three and a half year journey and I have enjoyed all of its ups and downs. I would like to take this opportunity to extend my thanks to the following people who have supported me along the way.

Firstly, I am very grateful to my supervisor Dr Roy Ruddle, my mentor Dr Vania Dimitrova, and Professor Ken Brodlie. Roy, I cannot thank you enough for being such a great supervisor, for having enough faith in my abilities to give me the chance to do this PhD with you; from the moment you asked what I did in my free time in the application interview, I knew it would be enjoyable to work with you. Vania, thanks for your sharp questions in my assessments which helped me think through my research, for social events within the School of Computing and with your research group, and for your support. Ken, thank you, you are an inspiration; it has been a privilege to have the benefit of your counsel.

I would like to thank Dr Lydia Lau and Dr Max Wilson, my examiners, for their useful suggestions for improving the final version of my thesis.

I would also like to thank the members of the Visualization and Virtual Reality Group, past and present, for making the trip a pleasant one. Thanks Chris Rooney for helping me find my feet in my first days in Leeds. Special thanks to Jeremy Swann for being such a great companion on the journey.

Thanks also go to the administration and support staff in the School of Computing and the University of Leeds for keeping everything running smoothly. Especially, this research could have not been conducted without a Fully Funded International Research Scholarship by the University of Leeds for my first three years and a scholarship by the School of Computing for my last six months.

I am indebted to all participants who took part in my two user studies. They are my heroes. Without them, this research could have never been completed.

To all my friends back home in Vietnam, here in Leeds, and somewhere else now, thank you for your care and understanding. You all have made my life more interesting.

Last but not least, I would like to thank my family who have always trusted and supported me in whatever I do.

Abstract

This research addresses the general topic of “keeping found things found” by investigating difficulties people encounter when revisiting webpages. The overall aim of the research is to design, develop and evaluate a web history tool that addresses these difficulties.

An empirical study has been conducted. Participants recorded their web navigation for three months using a Firefox add-on. Each participant then took part in a controlled laboratory experiment, to revisit webpages they had visited neither frequently (on only one day) nor recently (1 week or 1 month ago). Ten underlying causes of failure were discovered. Overall, 61% of the failures occurred when the target page: 1) had originally been accessed via search results; 2) was on a topic a participant often looked at; or 3) was on a known but large website.

Based on the findings of the empirical study, a new visualization history tool which supports people in revisiting webpages has been designed and developed as an add-on for Firefox. The tool has two main novel aspects. Firstly, by providing different navigation techniques, it enables users to revisit webpages within their long-term web history. Secondly, the visualization presentation is created based on the user’s navigational paths (even crossing different tabs) rather than the chronology which webpages were visited.

Evidence about the benefits of the visualization history tool has been provided through a three month field study. The results showed that such a history tool solved the identified causes of failure and helped participants succeed on 96% of revisiting occasions. They particularly used the tool to revisit webpages which had been visited neither frequently and nor recently. Participants often took only 3 steps to revisit a webpage. Overall, they were satisfied with the tool and rated it 4.1/5.0, and 84% of them wanted to keep using the tool after the evaluation.

Table of Contents

| | |
|---|-------------|
| Acknowledgements | v |
| Abstract | vi |
| Table of Contents | vii |
| List of Tables | xi |
| List of Figures | xiii |
| List of Pseudo Code | xv |
| Chapter 1. Introduction | 1 |
| 1.1 Research questions | 2 |
| 1.2 Approach overview..... | 2 |
| 1.3 Contributions | 3 |
| 1.4 Thesis outline | 4 |
| Chapter 2. Background | 7 |
| 2.1 Definitions | 7 |
| 2.2 How people navigate on the WWW..... | 8 |
| 2.3 How people revisit information | 9 |
| 2.3.1 Using explicit web history | 13 |
| 2.3.2 Using automatically recorded web history | 15 |
| 2.3.3 Search again from scratch using search engines..... | 19 |
| 2.4 Analysis of tools for revisiting | 23 |
| 2.4.1 List-based..... | 26 |
| 2.4.2 Visualization | 27 |
| 2.5 Technologies used in history tools | 30 |
| 2.6 User study methodologies..... | 32 |
| 2.6.1 Approaches | 32 |
| 2.6.2 Participants | 33 |
| 2.6.3 Tasks | 34 |
| 2.6.4 Measures | 35 |
| 2.7 Discussion..... | 36 |
| Chapter 3. The logging tool | 38 |
| 3.1 Requirements..... | 38 |
| 3.2 Design and implementation..... | 39 |
| 3.2.1 The tool bar | 41 |

| | | |
|-------------------|--|-----------|
| 3.2.2 | The history editor form | 48 |
| 3.2.3 | The privacy form | 48 |
| 3.3 | Summary..... | 49 |
| Chapter 4. | The underlying causes of revisit failure..... | 51 |
| 4.1 | Method..... | 52 |
| 4.1.1 | Participants | 52 |
| 4.1.2 | Revisiting experiment..... | 52 |
| 4.1.2.1 | Target page criteria..... | 52 |
| 4.1.2.2 | Target description and cue generation..... | 53 |
| 4.1.2.3 | Experiment procedure..... | 54 |
| 4.2 | Results | 55 |
| 4.2.1 | Logfile data..... | 55 |
| 4.2.2 | Revisiting experiment results | 57 |
| 4.2.2.1 | Success and failure..... | 57 |
| 4.2.2.2 | Revisiting strategies..... | 58 |
| 4.2.3 | The underlying causes of failure | 59 |
| 4.2.4 | Pattern of causes of failure for revisiting strategies | 64 |
| 4.3 | Discussion..... | 65 |
| 4.4 | Summary..... | 67 |
| Chapter 5. | The new visualization history tool..... | 68 |
| 5.1 | Requirements..... | 68 |
| 5.1.1 | Functional requirements..... | 68 |
| 5.1.2 | Data requirements..... | 69 |
| 5.1.3 | Environmental requirements | 69 |
| 5.1.4 | Usability requirements..... | 69 |
| 5.2 | The first iteration: paper-based prototype | 69 |
| 5.2.1 | Design | 70 |
| 5.2.1.1 | Global navigation | 71 |
| 5.2.1.2 | Result view..... | 71 |
| 5.2.1.3 | The toolbar..... | 73 |
| 5.2.2 | Initial feedback from users | 73 |
| 5.3 | The second iteration: the visualization design details | 73 |
| 5.3.1 | Addressing the feedback from the previous iteration | 73 |
| 5.3.2 | Detailed design of the <i>Result View</i> | 74 |
| 5.3.2.1 | The list view | 74 |

| | | |
|-------------------|--|-----------|
| 5.3.2.2 | The tree view | 75 |
| 5.3.3 | The visibility of the tool as an add-on in Firefox | 80 |
| 5.3.4 | Feedback from users..... | 81 |
| 5.4 | The third iteration: the refined visualization history tool..... | 81 |
| 5.5 | Technical implementation..... | 83 |
| 5.5.1 | Technologies chosen | 84 |
| 5.5.2 | Tool architecture | 84 |
| 5.5.3 | Implementation details | 85 |
| 5.5.3.1 | The heat map calendar | 85 |
| 5.5.3.2 | The Google searches tab..... | 86 |
| 5.5.3.3 | Search session reconstruction | 87 |
| 5.5.3.4 | Tree construction | 88 |
| 5.5.3.5 | Tree layout algorithm | 90 |
| 5.5.3.6 | Filters | 90 |
| 5.5.3.7 | Back and forward | 91 |
| 5.6 | How the new tool addresses the causes of failure | 91 |
| 5.6.1 | Known website | 92 |
| 5.6.2 | Search results | 92 |
| 5.6.3 | Deleted links..... | 93 |
| 5.6.4 | Links from email & social networks | 95 |
| 5.6.5 | Topic | 95 |
| 5.6.6 | Knowing the visited date | 96 |
| 5.7 | Summary..... | 96 |
| Chapter 6. | The user evaluation of the history tool | 99 |
| 6.1 | Method | 99 |
| 6.1.1 | Participants | 99 |
| 6.1.2 | Procedure..... | 100 |
| 6.1.3 | Logging participants' activities..... | 100 |
| 6.1.4 | The diary form | 101 |
| 6.1.5 | The follow-up interview..... | 102 |
| 6.2 | Results | 103 |
| 6.2.1 | Logfile data..... | 103 |
| 6.2.2 | How participants used the history tool..... | 104 |
| 6.2.3 | How the tool solved the underlying causes of failure | 107 |
| 6.2.4 | What participants thought about the tool | 112 |

| | |
|---|------------|
| 6.2.5 Diversity of participants | 114 |
| 6.2.6 Other comments and reflections from participants | 116 |
| 6.3 Discussion..... | 119 |
| 6.4 Summary..... | 121 |
| Chapter 7. Conclusions and future work..... | 123 |
| 7.1 Conclusions..... | 123 |
| 7.2 Future work | 126 |
| 7.2.1 Enhancing the visualization history tool..... | 126 |
| 7.2.2 Other directions for future work | 127 |
| Appendix A: Research ethics approval..... | 129 |
| Appendix B: Participant consent form..... | 130 |
| Appendix C: Participant information sheet for the user study described in Chapter 4..... | 131 |
| Appendix D: Information sheet for the user study described in Chapter 4..... | 132 |
| Appendix E: Participant information sheet for the user study described in Chapter 6..... | 134 |
| Appendix F: Information sheet for the user study described in Chapter 6..... | 137 |
| Appendix G: User manual for the logging tool of the user study described in Chapter 4..... | 139 |
| Appendix H: User manual for the logging tool of the user study described in Chapter 6..... | 142 |
| Bibliography | 150 |

List of Tables

| | |
|--|-----|
| Table 2.1 Presentation, recency, frequency, and scale supported by history tools for revisiting..... | 24 |
| Table 4.1 Comparison of web navigation studies. | 56 |
| Table 4.2 Percentage of revisited pages that were informational vs. navigational, subdivided according to recency and frequency..... | 56 |
| Table 4.3 Comparison of performance indicators across different revisiting strategies. | 58 |
| Table 4.4 Comparison of failure rates due to different causes. | 60 |
| Table 4.5 Summarisation of failures due to the <i>Topic</i> cause. | 61 |
| Table 4.6 Summarisation of failures due to the <i>Search results</i> cause..... | 62 |
| Table 4.7 Number of pages visited before and/or during the revisiting experiment and recency for each <i>Known website</i> failure. | 62 |
| Table 4.8 The number of failures of each cause for revisiting strategies..... | 64 |
| Table 6.1 Comparison of web activity with the study in Chapter 4..... | 103 |
| Table 6.2 The percentage of informational vs. navigational pages that were revisited for each combination of recency and frequency..... | 104 |
| Table 6.3 The percentage of exploring, revisiting and reviewing sessions..... | 104 |
| Table 6.4 Percentage of sessions that employed for each reviewing pattern. | 105 |
| Table 6.5 Percentage of sessions that employed each revisiting pattern..... | 106 |
| Table 6.6 Classification of diary entries into the underlying causes of failure..... | 109 |
| Table 6.7 Percentage of revisit diary entries that fell into each combination of recency and frequency. | 111 |
| Table 6.8 Percentage of participants who noticed each type of information encoded by the tool..... | 113 |
| Table 6.9 Tool loading time for different numbers of pages. | 114 |
| Table 6.10 Number of participants in each group and average number of diary entries, and usage sessions. | 115 |
| Table 6.11 The revisiting patterns used by each group of participants..... | 115 |

| | |
|---|------------|
| Table 6.12 Causes of failure distribution of each group of participants..... | 116 |
|---|------------|

List of Figures

| | |
|--|-----------|
| Figure 2.1 Thumbstrips creates a filmstrip of visited webpages' thumbnails at the bottom of Firefox browser. | 17 |
| Figure 2.2 History Tree visualizes webpages visited in each tab in a separate branch. | 18 |
| Figure 3.1 Three icons of the logging tool placed in the status bar of the Firefox browser: the first to start/pause the capturing, the second to go to the data folder, and the last to open the history editor form. | 41 |
| Figure 3.2 Browsing history editor form: Users can select a date to review their history on that date. Clicking on a list entry displays detailed information about that webpage. They can also delete unwanted entries. | 48 |
| Figure 3.3 User privacy form: users can block a specific webpage or the whole website by putting a URL in the "Never record this website" button. The domain of a URL is extracted by the tool. | 49 |
| Figure 4.1 Target page dialog, showing one with a thumbnail cue. | 55 |
| Figure 4.2 Percentage of unsuccessful revisits for each recency/cue combination. Error bars show standard error of the mean. | 57 |
| Figure 5.1 The paper-based prototype of the visualization history tool includes: the Global Navigation at the left with a heat map calendar and a tab view; the Result View at the right with a list view and a tree view; and the Toolbar at the top containing buttons. | 70 |
| Figure 5.2 The visualization design details of the history tool. | 74 |
| Figure 5.3 An example of the list view entries, with a bold maroon title for a webpage of interest (dwell time ≥ 30 seconds) and normal title for another webpage. | 75 |
| Figure 5.4 A horizontal orthogonal tree view with 63 nodes. Each node represents a webpage by its thumbnail. Frequency of visits to a page is encoded by node size, and edges are connected based on a user's navigational path. | 76 |
| Figure 5.5 An example of the overall network for the filtered set of webpages: nodes connected to the nominal node R are visited by direct entry, nodes connected to other nodes are visited by hyperlinks; the number on an edge presents the number of times users use that edge to visit a webpage. | 78 |
| Figure 5.6 The final spanning tree for the browsing session in Figure 5.5 after applying the Dijkstra algorithm. | 79 |

| | |
|--|------------|
| Figure 5.7 The dialog opened by right clicking on a node in the tree view displays a webpage's details and navigation options. | 80 |
| Figure 5.8 Icons of the visualization history tool, situated in the status bar of the Firefox web browser. | 81 |
| Figure 5.9 Sliders are placed in the Toolbar. Webpages that do not satisfy the filter criterion are removed from the list view but are shown in the tree view with reduced size. | 82 |
| Figure 5.10 An example of the list view entries, with a bold maroon title for a webpage of interest (dwell time \geq 30 seconds) and normal title for another webpage. | 82 |
| Figure 5.11 The tree's ROOT node is used to provide general information. | 83 |
| Figure 5.12 The final visualization history tool with: the Global Navigation at the left with a heat map calendar, a search box, and a tab view; the Result View at the right with a list view and a tree view; and the Toolbar at the top containing buttons and sliders. | 83 |
| Figure 5.13 The architecture of the visualization history tool: The logging module tracks and saves data about user web history in a SQLite database then the Visualization module retrieves and presents them to users. | 85 |
| Figure 5.14 An example of a month view calendar which spans six different weeks. | 85 |
| Figure 5.15 A user selects the domain of the target page from the domain list, to see all webpages within that domain. | 92 |
| Figure 5.16 If a user knows the target page belonged to a search session, they can select the relevant search query from the list of searches. | 93 |
| Figure 5.17 Right clicking on a node to select the "View all webpages visited from this webpage" option. | 94 |
| Figure 5.18 Example of a tree view that looks similar to the list view. | 94 |
| Figure 5.19 The "List view only" mode displays the full thumbnail of a webpage. | 95 |
| Figure 5.20 To display all visited webpages that are about the topic "bike", a user types "bike" in the search box. | 96 |
| Figure 6.1 The quick user guide. | 100 |
| Figure 6.2 An excerpt of the logfile of a participant's activity, recorded by the history tool. | 101 |
| Figure 6.3 The diary form. | 101 |
| Figure 6.4 The follow-up interview sheet | 102 |

List of Pseudo Code

| | |
|--|-----------|
| Pseudo code 3.1 The load event..... | 42 |
| Pseudo code 3.2 Creating and saving a webpage thumbnail. | 43 |
| Pseudo code 3.3 Calculating MD5 Hash for a webpage. | 44 |
| Pseudo code 3.4 LocationChange event..... | 45 |
| Pseudo code 3.5 Blur event. | 45 |
| Pseudo code 3.6 Focus event. | 45 |
| Pseudo code 3.7 Tracking referrer of a webpage..... | 46 |
| Pseudo code 3.8 Extracting anchor text of a clicked link..... | 47 |
| Pseudo code 5.1 Updating the calendar. | 86 |
| Pseudo code 5.2 Reconstructing the search session of a search query. | 88 |
| Pseudo code 5.3 Creating a tree for a filtered set. | 89 |
| Pseudo code 5.4 Filtering the tree view by resizing nodes..... | 91 |

Chapter 1. Introduction

A basic human principle is referring to what we have known in the past to accomplish our present missions. However, we cannot remember all that we have read and seen in our life. To assist our memory, external aids are utilised. For example, we often spend a great amount of our time to manage our personal information, highlight important text in a book, classify files into different folders, or direct emails to appropriate categories.

Today, millions of people all over the world navigate the World Wide Web (WWW) to get information (Fox, 2002; Cole et al., 2003). The whole WWW is huge but each person visits only a certain number of webpages which form an individual's web history. It is predicted that an "average" person will look at approximately one million webpages during their lifetime (Weinreich et al., 2006). People may bookmark some webpages or save them as files on hard disk but no one would manage all the webpages that they have been to because of the required time and effort. Therefore, revisiting webpages is often more challenging than other personal information such as files and emails.

At the same time, revisiting is common. Studies (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006) have shown that between one third and one half of visits are return visits to pages that had been previously seen. The plethora of techniques that people use to assist revisiting ("keeping found things found") are well documented (Jones et al., 2001). A number of tools supporting revisiting have been developed. The tools vary from a browser's built-in functionality (e.g., back and forward buttons, bookmark, and history list) to browser extensions, and independent commercial and research applications (see Chapter 2). However, occasionally people still have frustration at not knowing where to "go" in order find a webpage again (Bruce et al., 2004; Teevan, 2007b). The true cost of revisiting is hard to calculate, but it has been estimated that knowledge workers wasted 15% of their time as a result of difficulties experienced while trying (and often failing) to find information that they knew already exists (Feldman, 2004). This doctoral thesis explores how to help people navigate effectively within their own long-term web history to find webpages again.

1.1 Research questions

To explore how to help people revisit webpages more effectively, the research described in this thesis focuses on answering the following questions:

What are the difficulties that people encounter when revisiting webpage? Revisiting is often not too difficult but occasionally people still show frustration at not knowing where to “go” in order re-access a webpage (Bruce et al., 2004; Teevan, 2007b). So when do people find it challenging to revisit a specific webpage? Answering this question may yield solutions to support revisiting.

What tools will support realistic revisiting? How should a future web history tool be built? Which techniques are needed for people to effectively navigate within their long-term web history to find webpages again? How should a web history be presented to users?

Once built, how effective is the history tool? How should the tool be evaluated? Does it help people eliminate the difficulties of revisiting as designed?

1.2 Approach overview

This thesis aims at helping people navigate effectively within their own long-term web history to find webpages again. The ultimate goal is to design, develop and evaluate a web history tool that supports revisiting. The research adopts the user-centred design methodology which “starts with the users, and to work from there” (Norman and Draper, 1986). The approach follows the ISO standard Human-centred design for interactive systems (ISO 9241-210, 2010)¹. In short, the user-centred design is a multi-stage problem solving process that puts users at the heart of the design process. It starts with an explicit understanding of users, tasks and environment. Then users are involved throughout the design and development stages in several iterations.

Two factors that clearly affect the ease of revisiting are the frequency and recency with which a webpage has been visited. The lack of frequency and recency means that it is difficult for users to remember a page’s domain or

¹ See http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075

URL, and the vagueness with which a page is recalled (“I remember seeing a page about something like X”) makes it difficult to search/browse for it from scratch. Previous research has divided frequency and recency into multiple categories (e.g., weekly vs. monthly vs. less often (Capra, 2006); < 1 day vs. < 1 week vs. ≥ 1 week (Mayer, 2009)), but the present research simplifies this to four categories, according to whether a webpage was previously visited:

1. Both frequently (on more than 1 day) and recently (less than 1 week).
2. Frequently but not recently.
3. Recently but not frequently.
4. Neither frequently nor recently.

The most challenging category (neither recently nor frequently - the 4th category listed above) is the primary focus of this thesis. To investigate how people find webpages again and what difficulties they encounter, an empirical study has been carried out (see Chapter 4). Based on the findings of the empirical study, a new history tool has been iteratively designed (see Chapter 5). Finally, it has been evaluated with 19 participants in a three-month field study (see Chapter 6).

1.3 Contributions

This thesis makes three main contributions to the areas of information retrieval, personal information management, and human-computer interaction.

First, the thesis investigates the underlying causes of failure when people try to revisit webpages (see Chapter 4). Ten causes are identified by analysing the unsuccessful revisiting trials of a controlled laboratory experiment, data about participants’ navigational actions during the experiment, video/audio of participants’ thinking aloud and related data from participants’ logfiles. The three main causes (accounting for 61% of the failures) are: (1) participants visiting a large number of pages on a particular topic, (2) webpages that have originally been accessed via search results, (3) participants knowing which website contains a page but that site itself being large.

Second, the thesis proposes a novel visualization web history tool (see Chapter 5) which supports people in revisiting a complete history using visualization rather than a list-based approach. The tool has two main novel aspects. Firstly, by providing different navigation techniques, it enables users to revisit webpages within their long-term web history. Secondly, the

visualization presentation is created based on the user's navigational paths (even crossing different tabs) rather than the chronology which webpages were visited.

Third, the thesis provides evidence about the benefits of the visualization history tool (see Chapter 6). The results show that such a visualization history tool enables users to navigate effectively within their long-term history to find webpages again. The tool helped participants succeed on 96% of revisiting occasions. Overall they are satisfied with the tool, rated it 4.1/5.0, and 84% of them want to keep using the tool after the evaluation.

1.4 Thesis outline

This thesis is divided into seven chapters.

Chapter 2 provides background and has seven sections. Section 2.1 explains definitions such as revisit, re-finding, search, search session, search trails, query, search query, and keyword. Then how people navigate on the WWW is described in Section 2.2. In general, this navigation involves three mechanisms: browsing, searching, and direct entry. Section 2.3 categorises how people revisit webpages into three mechanisms that involve using explicit web history, automatically recorded web history, and search engines. Tools that support revisiting are examined in Section 2.4. Based on presentation, they are divided into two approaches: list-based and visualization. Then visualization history tools are analysed deeper in terms of representation, presentation, and interaction. Section 2.5 briefly summarises technologies used to develop web history tools. Some experimental research methodologies are presented in Section 2.6. The final Section 2.7 differentiates this thesis from previous research.

Chapter 3 describes the web history logging tool with the requirements and challenges for implementing it. The logging tool has been developed as an add-on for Mozilla Firefox to automatically capture an individual's web history, and has been tested and used on Windows, Mac, and Linux computers. The add-on stores webpage information (thumbnail, URL, title, description) and navigational information (visited time, URL and ID of referrer, anchor texts used to access the page, dwell time) in a database in the individual's personal file space. Section 3.1 of this chapter states the requirements of the logging tool. Section 3.2 describes the design of the tool and discusses key challenges in the development (e.g., capturing user's navigational paths and tracking dwell times in a multi-tabbed browser, and

removing unwanted entry such as ads, frames, and private pages). Section 3.3 summarises and discusses the limitations of the logging tool.

Chapter 4 presents an empirical user study to investigate how people revisit webpages and the difficulties they encounter. Chapter 4 has four sections. Section 4.1 describes the methodology of the study. In this study, participants recorded their web navigation for three months using a Firefox add-on. Each participant then took part in a controlled laboratory experiment, to revisit webpages they had visited neither frequently (on only one day) nor recently (1 week or more ago). Section 4.2 reports the results of the user study. First participants' logfiles are analysed to compare with web navigation activities reported by previous studies. Then the results of the revisiting experiment sessions are reported. Finally, the experiment and logfile data are combined to determine the underlying causes when participants failed in their attempts to revisit webpages. These results are discussed in Section 4.3 and summarised in Section 4.4.

Chapter 5 describes the requirements, design, and technical implementation of a new visualization history tool that addresses difficulties that people encounter when revisiting webpages. Consistent with established processes for interaction design, Section 5.1 of this chapter presents the requirements of the tool in four aspects: functional, data, environmental, and usability. Then, three design iterations of the visualization history tool are described in Sections 5.2, 5.3, and 5.4. To meet the data and environmental requirements, the history tool has been developed as another module of the logging tool. Section 5.5 emphasises some important aspects of the implementation of the tool. Then how the history tool would address difficulties of revisiting is explained in Section 5.6. Finally, Section 5.7 summarises the chapter.

Chapter 6 presents a three month field study of how participants used the visualization history tool. In this study, an electronic diary methodology was employed. At the end of the study, a follow-up semi-structured interview was conducted to clarify aspects of the diary entries and to learn what people thought about the tool. The method of the study is described in Section 6.1. Then Section 6.2 reports the results of the study in four aspects: the logfile data, the usage of the tool, the diary entries, and the follow-up interview. Some other comments and reflections from participants are also included. Finally, Section 6.3 summarises the chapter.

Chapter 7 concludes the thesis and discusses some implications for future work.

Eight appendices of documents related two user studies described in Chapter 4 and Chapter 6 are also added at the end of the thesis.

Chapter 2. Background

This chapter provides the background for the research described in this thesis and the context for the work. The chapter starts with some definitions to eliminate any ambiguity in Section 2.1. As the title says, this doctoral research aims at supporting webpage revisiting with history data and visualization. To do that, how people revisit webpages must be reviewed. However, revisiting webpages depends on how people have previously reached to the webpages, which is also related to human memory. To cover these topics, how people navigate the WWW is summarised in Section 2.2, and Section 2.3 discusses how people revisit webpages. Section 2.3 also highlights the findings of human memory and personal information management research.

To design a new history tool to support revisiting, the pros and cons of existing tools need to be examined. Section 2.4 carries out this task. It first classifies main history tools by presentation, recency, frequency, and scale that they support revisiting. After that, it analyses in detail two approaches of presenting a history to users: list-based and visualization presentation. The technologies used to develop history tools are reviewed in Section 2.5 to justify technical solutions for the development of the new history tool.

In order to support better webpage revisiting, this thesis needs to investigate difficulties of people when they attempt to revisit webpages. An empirical user study is required. However, there are different methodologies to do such a user study. Section 2.6 surveys popular experimental research methodologies employed by previous research.

Finally, Section 2.7 of this chapter provides an overall view of the problem tackled in this research and differentiates this study from previous ones.

2.1 Definitions

Previous studies have used two terminologies “search” and “query” inconsistently. In this thesis, a query means a formal database query such as an SQL query. A search, on the other hand, is the use of search engines to find information by typing a string into a search box. The string typed into a search box is called a search query. But a search query is not a keyword. A keyword can be considered the ideal of a search query. It is an abstraction from multiple search queries. A search query may be misspelled, out of

order or have other words tacked on to it, or conversely it might be identical to the keyword.

A search session is defined as “*a process that involves one or many rounds of searches dealing with the same information need*” (Jiang et al., 2012). This definition is aligned with previous studies (Jansen et al., 1998; Silverstein et al., 1999; He and Göker, 2000; Cacheda and Vinã, 2001; Spink et al., 2001). Search trails “*originate with a directed search (i.e., a query issued to a search engine), and proceed until a point of termination where it is assumed that the user has completed their information-seeking activity. Trails can contain multiple query iterations, and must contain pages that are either: search result pages, visits to search engine homepages, or connected to a search result page via a hyperlink trail.*” (White and Drucker, 2007)

Mayer (2009) defines “*Revisit (or revisitation) is the repeated visit to a webpage as identified by its location, i.e., its address*”. With this definition, a revisit might refer to the same or modified content but tab/window switching does not count. In her PhD thesis, Teevan (2007b) has a definition for re-finding “*Re-finding is the process of finding information that has been seen before*”. This definition makes re-finding similar to revisiting. However, re-finding is based on the content rather than the address of a webpage. Later, Tyler and Teevan (2010) say “*When an individual clicks a URL following a search and then later clicks on the same URL via another search, we call it re-finding*”. So in their definition, re-finding is only a subset of revisiting. In fact, people would like to go back to webpages that they have visited not only by search but also by other mechanisms such as direct entry or browsing (see Section 2.2). Therefore this thesis uses Mayer’s definition, which is consistent with earlier studies (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006).

2.2 How people navigate on the WWW

Navigation on the WWW involves in three basic mechanisms: browsing, searching, and direct entry (Liebscher and Marchionini, 1988; Marchionini and Shneiderman, 1988; Marchionini, 1997).

- Browsing: Users travel from one page to another by clicking on hyperlinks.

- Searching: Users use search engines to search the WWW as a whole (global search) or within a website (local search) and subsequently select webpages by scanning the results list.
- Direct entry: Users directly enter the URL of a page, for example, by selecting it from bookmarks or a history list, typing it into the browser address bar, or copying and pasting a URL from somewhere.

These navigation mechanisms are often combined during the same activity (Belkin et al., 1993), for example, searching or direct entry of URL are often followed by browsing (Bates, 1989).

Since the early days of the WWW, studies (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006; Adar et al., 2008) have been made of people's navigation patterns so that browsers and search engines can be improved. Although the number of webpages that people visit has increased over time, people's navigation patterns have remained broadly the same. For example, URLs are most often accessed via hyperlinks (43-45%), and direct access to URLs accounts for 9-13% of visits. The use of the Back button has declined with the introduction of tabbed browsers, but still accounts for 14%. These studies have also shown that between one third and one half of visits to pages are revisits.

The length of time that people dwell on webpages varies considerably, with around 50% of webpages looked at for 12 seconds or less, 70% for 30 seconds or less, and only 10% for more than two minutes (Weinreich et al., 2006). During search sessions, a dwell time of 30 seconds or more on a webpage can be indicative of webpage utility (Fox et al., 2005), and this threshold was used to analyse search trails in web logs (White and Huang, 2010).

Because users often leave browsers running for extended periods of time without interacting with it, one of the first studies on WWW (Catledge and Pitkow, 1995) determined session boundaries using a 25.5 minute period of user inactivity. A similar threshold of 30 minutes has been adopted in later studies (Kelly and Belkin, 2004; Liu et al., 2010; Tyler and Teevan, 2010).

2.3 How people revisit information

Revisiting is often motivated when people remember they have seen information somewhere and need it again. To do so, the address of the information must be relocated. Revisiting can be in different forms, from

relocating a paper document in an office, or a file on a PC to an email in a mail box or a webpage on the WWW. This is part of the personal information management (PIM) which investigates how people create, organise, manage, archive, relocate, and reuse a variety of types of information including paper documents, email messages, webpages, electronic documents, files, contacts and calendar information (Jones, 2008). As human memory plays an important role in PIM (Elsweiler, 2007), this section first briefly reviews important studies on human memory and revisiting personal information (e.g., files and emails) then analyses in detail how people revisit webpages.

The way people navigate and reach to information the first time can vary. Sometimes, the purpose and scope of such navigation are well-defined (e.g., finding the Visualization and Virtual Reality group's homepage). In other cases, they may be more vague (e.g., doing some research to buy a suitable DSLR camera), or even involve chance (e.g., reading interesting news, links suggested by friends). Typically, revisiting is more purposeful and well defined: people often try to re-find specific items. By definition, a revisit involves looking for something that has been seen before. Therefore, revisiting can make use of the knowledge that is remembered from the previous visits such as starting points and waypoints, date and time, and other contextual information. Clearly, human memory somehow affects this process.

In general, it is agreed that there are three types of memory: sensory memory, short-term or working memory, and long-term memory (Atkinson and Shiffrin, 1968; Baddeley and Hitch, 1974; Dix et al., 2003). Sensory memory acts as a buffer for stimuli received through any of the five human's senses. Sensory is either quickly passed into short-term memory by attention or overwritten by new information. Short-term memory acts as a "scratch-pad" for the temporary recall of information. It can be accessed rapidly, but it also decays quickly and has a limited capacity. Information may be transferred from short-term memory into long-term memory by rehearsal (Eysenck, 2001). Long-term memory stores factual information, experiential knowledge, procedures - in fact, everything we "know". It has essentially an unlimited capacity which can hold information over long periods of time, perhaps indefinitely. Revisiting could involve both short-term and long-term memories and has been classified into short-term and long-term revisit (Mayer, 2009).

Evidence has shown that information can be encoded in human memory by different ways such as visual encoding (Bahrick et al., 1967; Haber, 1969; Frost, 1972; Kosslyn, 1973, 1975, 1976; Kosslyn et al., 1978; Farah, 1993), spatial encoding (Thorndyke and Stasz, 1980; Brewer and Treyns, 1981; Kosslyn, 1981; Kerr, 1983; Tversky, 1991; Cohen, 2004), semantic encoding (Anisfeld and Knapp, 1968; Grossman and Eagle, 1970; Bruder and Silverman, 1972; Cramer and Eagle, 1972), acoustic encoding (Baddeley, 1966; Nelson and Rothbart, 1972), and temporal encoding (Smith et al., 1978; Rubin, 1982; Brown et al., 1985; Larsen et al., 1996; Huttenlocher and Prohaska, 1997; Friedman, 2004). There is also an important link between context and memory (Fleeson and Kihlstrom, 1988). Evidence revealed that context has a strong influence on what people can remember. If the context of encoding and storage information is returned at the retrieval time it can improve retrieval performance (Godden and Baddeley, 1975; Smith et al., 1978). Research on the effects of context on recognition has also shown that the abilities of recognition are superior to recall (Gillund and Schiffrin, 1984). It is easier for people to recognise that they have seen objects before than to list all of the objects that they have seen. The theory of cue-dependent forgetting mentions that information is available in memory but cannot be accessed without the appropriate “cue” (Tulving, 1974). Evidence for this theory was supported by several other studies (Underwood and Schulz, 1960; Tulving and Psotka, 1971; Czerwinski and Horvitz, 2002). Further evidence suggested that people are likely to remember the context in which objects are used more than their specific properties (Jaimes et al., 2004). For example, people may not remember all the details about a document, but they may remember why they read it or who gave it to them to read. Visual, spatial, and semantic encodings were exploited in many previous studies to support PIM in general and webpage revisiting in particular (see Section 2.3.1 and 2.3.2).

People often consider the value of personal information to make judgments about whether to keep or delete such information (Bergman et al., 2009). Personal information items (e.g., files, emails, bookmarks and contacts) vary in their subjective importance and even this may change over time. If information is not kept it is unavailable to re-access when needed later (Jones, 2004). On the other hand, if kept, irrelevant information may create clutter and obscure important information. A lot of information is kept for anticipated future use but in fact never needed (Whittaker and Sidner, 1996; Abrams et al., 1998; Jones, 2004; Bruce, 2005).

People tend to organise papers into “piles” and files (Malone, 1983). Piles are placed spatially around the office and support short-term memory. However, people have difficulty keeping track of piles over time as their number increases (e.g., on average each person in an office had 18 boxes of paper (Whittaker and Hirschberg, 2001)). A solution for this situation is to arrange papers into named files (or folders) to support longer-term storage. However, people still have difficulty in retrieving information by location when the number of folders is more than ten (Jones and Dumais, 1986). People also keep a large number of personal electronic files (Boardman and Sasse, 2004) and personal pictures (Whittaker et al., 2010). Similar to papers, computer files are organised in folders and subfolders (Boardman and Sasse, 2004). To access a piece of information again, people’s preferred strategy is first to recall which folder a desired file is in. Then they look at the list of files in that folder and attempt to recognise the desired file.

Sometimes, users sort files by name, date, file type or some other characteristic. They tend not to search files or folders by name and only employ a full-text search as the last method (Barreau and Nardi, 1995).

Personal email archives are growing larger (Fisher et al., 2006; Whittaker, 2013). The patterns of managing and retrieving emails have also been investigated. Emails are increasingly organised for both task management and personal archiving in the same ways as electronic files (Whittaker and Sidner, 1996; Bellotti et al., 2005; Fisher et al., 2006). Filing decisions (e.g., which folders to create, what to name them, how to organise them) partly depend on whether an item is an email message or a personal file (Whittaker and Sidner, 1996). Therefore, filing takes time and the folders that are created today may be ineffective in the future. Although, folder names are assigned by users, they are often not descriptive of folder contents and purpose. Besides, items placed in a folder are sometimes forgotten until after the period of their usefulness has passed (Whittaker et al., 2006). Users apply three main strategies to revisit email: (1) identifying folders (containing manually classified messages), (2) searching, and (3) sorting (Whittaker et al., 2006). Previous studies have also investigated the latent that users need regarding handling emails (Szóstek, 2011), topic detection and tracking (Cselle et al., 2007), and difficulties of re-finding email (Elsweiler et al., 2011).

Similar to revisiting emails and files, revisiting webpages is part of personal information management. People also manage visited webpages by organising bookmarks into folders in the same ways as files and emails

(Boardman and Sasse, 2004). Sometimes people save webpages as files on hard disk, print them out, or email to themselves for later re-access (Jones et al., 2001). This makes revisiting webpages become relocating papers, files, or emails. However, the number of webpages that people visit is much larger than that of files and emails whereas they only explicitly manage a few webpages. This makes revisiting webpages more challenging. The rest of this section provides a more comprehensive review of revisiting webpages.

Research into revisiting webpages indicates that 43% of revisits occur within one hour (often shopping and reference webpages or as a result of hub and spoke navigation), another 17% within one day, and the remainder being longer-term revisits (Adar et al., 2008). Jones et al. (2001) conducted a user study with 11 participants (four researchers, three information specialists and four managers) to explore how people keep found things found on the WWW. This study focused on the methods people used to manage webpages for revisit. Results showed that besides features supported by web browsers (e.g., bookmarks and history), people also emailed webpages' addresses (sometimes with comments) to themselves or others, printed them out, saved them as files on a hard disk, pasted URLs in documents or personal websites, wrote down notes on paper containing URLs, directly typed URLs into the address bar and searched again from scratch. The following sections review three distinct approaches to revisiting where people: (1) explicitly record the location of a webpage, (2) use automatically recorded web history, or (3) search again from scratch.

2.3.1 Using explicit web history

Explicit web history is a collection of webpages that people think that they would need to visit again in the future so they manually record the location of webpages or manually organise them. This allows users to revisit webpages in the long term. The most popular way of doing this is to use web browsers' Bookmarks (in Chrome, Firefox, and Safari) or Favourites (in Internet Explorer). Current web browsers allow users to organise bookmark entries in hierarchical folders (Abrams et al., 1998). Each webpage is usually presented by its title and its favicon (a small picture assigned for each website by web developers). Each bookmark entry references only a single page, which loses contextual information (e.g., other webpages visited in the same session, navigational path of the page) (Jones et al., 2001).

Bookmarks free users from remembering and typing URLs explicitly (Abrams et al., 1998). Although many people add webpages to their bookmarks (Pitkow and Kehoe, 1996), they rarely re-access them (Catledge and Pitkow,

1995; Jones et al., 2001). Another problem is that the size of a user's bookmark collection grows steadily and roughly linearly over time (e.g., more than 40 webpages after a year and more than 200 after two years (Abrams et al., 1998)). Therefore, similar to personal files and emails, bookmarks require a lot of effort to maintain.

Methods like emailing webpages' addresses, printing webpages out, saving them as files on a hard disk, pasting URLs in documents or personal websites, writing down notes on paper containing URLs (Jones et al., 2001) are essentially other forms of bookmarks either in tangible (printing out, writing notes) or electronic formats. People still need to anticipate which webpages they might need to revisit. Over time, people will end up with a list of URLs, emails or piles of papers. To retrieve information from those archives easily, considerable additional maintenance is required.

Several studies have attempted to reduce the maintenance of bookmarks. HiBo (Kokosis et al., 2005) automatically organises bookmarks into topical categories using a built-in subject hierarchy. Different to HiBo, HyperBK (Staff and Bugeja, 2007) automatically classifies a webpage into an existing bookmark category. These systems partly address the problem however, to revisit a webpage users need to figure out which category the webpage belongs to. Besides that, the accuracy of categorising algorithms need improving, for example, only 61% of bookmarks were classified correctly with HyperBK.

A number of research projects have developed tools to visualize explicit web history and evaluated the tools with a small number of participants. Some of the tools use network visualizations. For example, SessionGraphs (Mayer and Bederson, 2001) requires users to name a task that defines what they are browsing for. Then visited webpages are added to one of the sessions attached to the task. A session can be created either manually by a user or automatically by an underlying heuristic algorithm based on pausing time and new windows. This approach is aimed at a group of users that is both able to define their tasks and willing to spend some effort organising their tasks hierarchically.

Spatial metaphors such as book, bookcase and desktop are used in WebBook and Web Forager (Card et al., 1996). Important webpages are bound into different books by users. Using 3D, the workspace contains three different areas: (1) the "Focus Place" to view a full webpage of a WebBook, (2) the "Immediate Memory" space behind the "Focus Place" to hold the WebBook temporarily in use, and (3) the "Tertiary Place" including a

bookcase and a desk to store other WebBooks. The metaphors are intuitive and conventional to users, however users have to interact with the Webbook by flipping sequentially page by page. Besides that, it is not easy to remember which book a page belongs to. In another project, the Data mountain concept exploits the use of spatial memory for data management (Robertson et al., 1998). Users may arrange thumbnails of a collection of webpages arbitrarily on an inclined 2D plane like a mountain in a 3D environment. This method was motivated by the fact that humans can often remember where they have placed objects. Different textures of the 2D plane are used as landmarks, which are important for readability and recognisability (Darken and Sibert, 1996).

The explicit web history approach supports revisiting a limited collection of webpages at any recency and frequency. However it only works well when: (1) the number of recorded webpages is relatively small, (2) users can anticipate which webpages they need to visit again in the future, and (3) they are willing to maintain webpages in categories. Managing and revisiting webpages with this approach is rather similar to dealing with personal files and emails.

2.3.2 Using automatically recorded web history

There are a number of entirely automatic methods for recording history. One that is familiar to all WWW users is links changing colour when a page has been visited, which lets users recognise which links they have visited before to make navigational decisions. However, the colour change will expire after a certain period of time, depending on users' browser settings.

A browser's Back and Forward buttons are also familiar to all WWW users, and are frequently used to return to pages in the current navigational session. Backtracking was the second most used navigation method after hyperlinks. It accounted for 32% to 36% of all navigation actions (Catledge and Pitkow, 1995; Tauscher, 1996; Weinreich et al., 2006). In the old days, this activity was often motivated by the need of exploring different navigational branches from each webpage (Nielsen, 1995). Nowadays, with the introduction of tabbed browsers, the use of the back button has reduced significantly to 14.3% (Weinreich et al., 2006). Complementing the Back and Forward buttons, the History menu of Firefox lists the 10 most recently visited webpages across all tabs. Similarly, the Recently Closed Tabs/Windows list displays 10 pages have been most recently closed, for example, to revisit a page that was accidentally closed.

Long-term revisiting is supported by a web browser's history list, which automatically adds addresses of visited pages into the archive. Although people rarely access the history list directly (Tauscher and Greenberg, 1997; Jones et al., 2001; Weinreich et al., 2006) the information that it contains underpins URL auto-completion, and displays of a person's most frequently visited pages. To exploit URL auto-completion, users need to type some characters of either a webpage's URL or title in the address bar. If the title and URL are poorly assigned, maybe no webpages contain them be offered users. By contrast, sometimes many webpages on the same topic containing these characters make users difficult to decide which webpage they are looking for. Today some browsers provide a search capability within the history list (e.g., Firefox integrates this feature into the address bar) however its existence is not widely known. Users can either open the history dialog to browse by date/site/most visited/last visited or type a search query in a textbox to search within their history. All webpages whose titles contain the search query will be displayed in a linear list of titles with favicons.

Google History improves on a web browser's history list by adding three new features. First, it provides a heat map calendar so users can easily navigate to different points of time by date, month and year. The heap map uses four different shades of blue to encode the number of webpages visited on each day (e.g., 0-5, 6-10, 11-20, 21+). Second, webpages can be filtered by categories such as web, images, news, products, sponsored links, videos, maps, blogs, and books. Third, Google History captures all search queries and the webpages clicked on results pages of each search. Of course users can also search within their history. Recently, Google History list also lets users explore more webpages from the same website by clicking on the small arrow at the end of each list item.

xMem (Ceri et al., 2006) operates in a similar way to a browser history list, and also categorises visited pages (titles and URLs) into topics by exploiting semantic information. By contrast, CWH (Won et al., 2009) improves "Search Within History" of a web browser by (1) letting users search with meta data such as date and time, and (2) adding the webpage's thumbnail for each result item. Both approaches share the shortcoming that information about users' navigational paths and sessions is lost.

Today, the functionality of most web browsers can be expanded with extensions. With Firefox browser, an extension is called an add-on. Some interesting add-ons have been developed to support revisiting. Flipora is a popular history add-on that is available for Firefox, Chrome and Internet

Explorer, and has more than 10 million users² (as in November 2012). It works like the history list of web browsers but stores the list on a server so users can access it anywhere, and then browse or search within that list. WebMynd³ is another add-on which detects users' Google searches and automatically displays only visited webpages in a widget at the right side of the Firefox browser. Rather than using a conventional list of text, Thumbstrips⁴, as its name suggests, creates a filmstrip of visited webpages' thumbnails at the bottom of the Firefox browser (see Figure 2.1). Looking at the filmstrip, users can recognise webpages they wish to find again. However it is not easy for users to relate a given page to their navigational path and the history is deleted when the browser is closed.

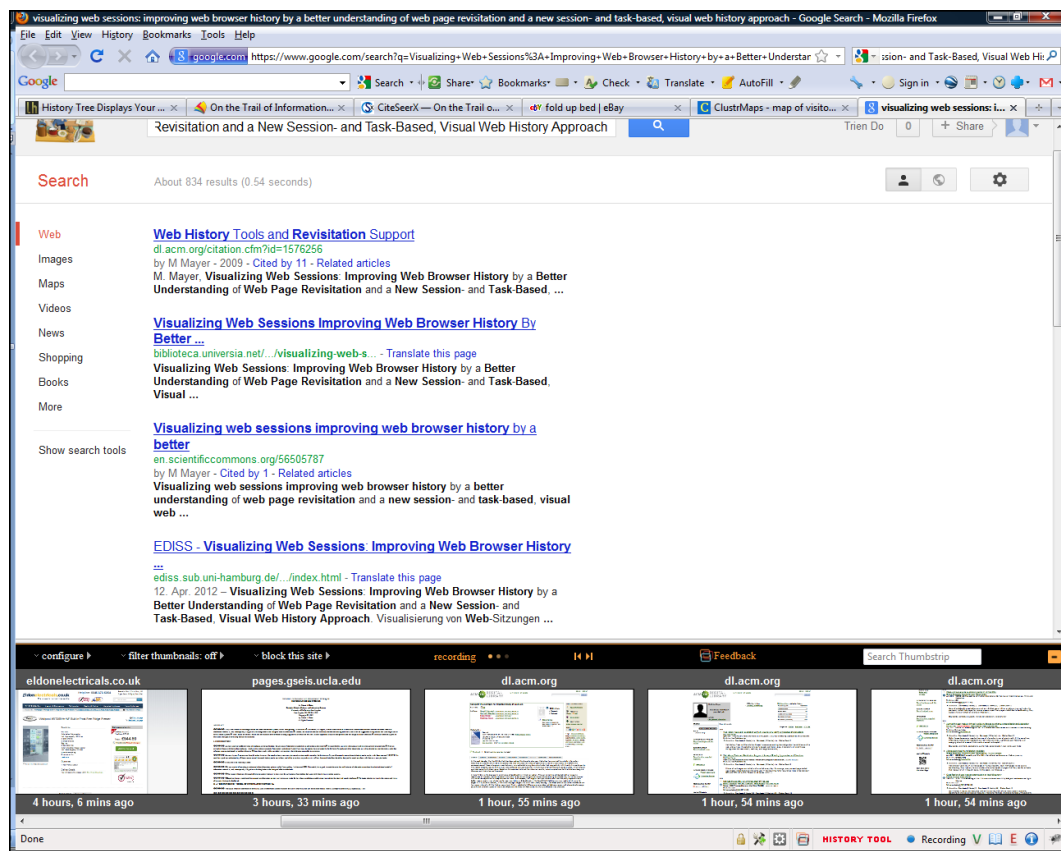


Figure 2.1 Thumbstrips creates a filmstrip of visited webpages' thumbnails at the bottom of Firefox browser.

Extracting information directly from the Firefox browser's history data file (except webpage thumbnails), the add-on BrowseLine (Hoeber and Gorner,

² See <http://www.flipora.com/>

³ See <http://www.webmynd.com/html/nytimes.html>

⁴ See <http://rockyourfirefox.com/2010/05/thumbstrips>

2009) visualises an individual's web history in two orthogonal timelines: macro-time and micro-time. Running from bottom to top vertically, the macro-time is divided into one hour slots. The micro-time is in the horizontal direction, from left to right, presenting webpages visited within each hour slot of the macro-time. This presentation wastes much of real estate because users often do not browse the WWW in continuous hours. Again, this approach arranges visited webpages by the time they are open. Multiple tabs and navigational paths are not taken into account. Participants complained that the tool took too much time to load the whole Firefox history data file for each revisiting.

The History Tree (Panasiti, 2009) add-on visualizes history of open tabs as a tree. Each branch of the tree represents a tab of the browser (see Figure 2.2). Nodes in each branch are the sequence of webpages opened in that tab, and each node contains a webpage's title and visited time. A disadvantage of the tool is that it might lose the user's navigational path because, if a page is opened in a new tab, it will start a new branch from the root of the tree losing the connection between the newly open webpage with the previous one. Also, when the browser is closed, all the history is deleted.

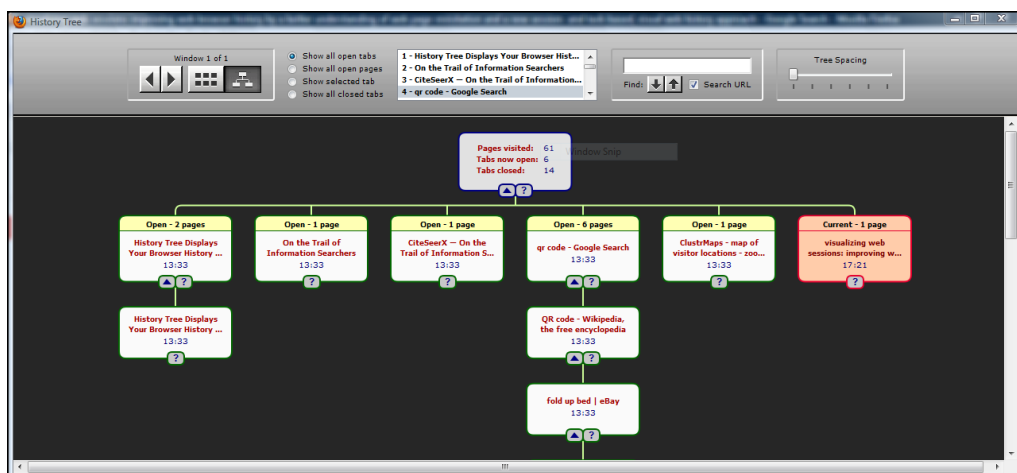


Figure 2.2 History Tree visualizes webpages visited in each tab in a separate branch.

Browser Extensions are a new technology. However, visualization was first adopted to address the revisiting problem more than 15 years ago. Webmap (Dömel, 1995), MosaicG (Ayers and Stasko, 1995) and PadPrints (Hightower et al., 1998) are three of the first tools to present a web history by spanning trees. One of the main problems of these two tools is the tree sizes grow rapidly when more nodes are added. Domain Tree Browser (Gandhi et al., 2000) improves on this limitation by dividing the presentation into three panes: domain pane, tree pane, and webpage pane. With this approach,

each domain (website) is represented by a separate tree to address the scale problem. However the way a new node is added to a tree might confuse the users because that node is always attached to the last visited node of its domain regardless of a user's navigational path. Nestor Navigation (Eklund et al., 1999) and WebNet (Cockburn and Jones, 1996) create 2D directed graphs of a user's navigational paths. The graph is drawn as a straight line of connected nodes if users do not backtrack. Otherwise, if the users backtrack and follow new directions, the graph branches. Loops could also happen. Using the same approach, WebQuilt (Hong and Landay, 2001) was designed to analyse browsing patterns of multiple users on a website. WebPath (Frécon and Smith, 1998) generates a three-dimensional representation of a web browsing history. A webpage is represented as a cube which is labelled with the webpage's title and appears as any one of its background image, first image in-lined, or background colour. Navigational paths are represented by arrows between cubes. The colour of an arrow indicates if the two connected webpages are in the same website or not. The vertical position of a new webpage is incremented by time so that the most recently visited cube is at the highest point. The horizontal position of a new webpage can be determined by mapping the horizontal axes to two of a possible eight metrics (e.g. loading time, number of images, and sever name). The drawback of these four tools is that their presentations become cluttered as more and more webpages are visited.

Compared to explicit web history, theoretically, the automatically recorded web history approach can support revisiting for any recency and frequency of a complete web history. Users are not required to anticipate about future revisits and manage their web history. This enables user to concentrate on their tasks instead of being worried about keeping track of what webpages they have been to. However, as the information space of a complete history is large over time, the two key aspects need to be taken into account when designing a new history tool are (1) how to help users navigate more effective their web history and (2) how to present a web history to them.

2.3.3 Search again from scratch using search engines

Web searching differs from traditional information retrieval in three main aspects: (1) dynamic vs. static documents, (2) heterogeneous vs. homogeneous in terms of content, media, form, and producers, and (3) heterogeneous vs. homogeneous in demographic characteristics such as position held, education, and searching experience (Park et al., 2005). According to PewInternet (2012), more than 74% of WWW users use search

engines to reach other web sites and the use of search engines has overcome email to become the most popular internet task.

During a search session, users may perform several actions including submitting a search query, viewing results pages, clicking on URLs, viewing documents, and returning to the search engine for query reformulation (Jansen et al., 2007). Using a version of the Dempster-Shafer theory (Voorbraak, 1991), He et al. (2002) determined the average duration of a search session was about 12 minutes. Later studies reported that this duration was about 15 minutes (Jansen and Spink, 2003) or less than 30 minutes (Jansen et al., 2007).

In each search session, users often submit more than one search query, for example 1.7 queries in (Cacheda and Vina, 2001), 1.8 queries in (Park et al., 2005), 2.0 queries in (Silverstein et al., 1999), 2.8 queries in (Jansen et al., 1998), and from 2.3 to 2.9 in (Jansen et al., 2007). Users need three or more search queries in 17% to 29% search sessions (Cacheda and Vina, 2001; Spink et al., 2002; Park et al., 2005). Users often view about two to three documents per search query and about eight documents in a given session (Cacheda and Vina, 2001; Jansen and Spink, 2003). Most often, users view only the first page of results (Jansen and Spink, 2006), which is typically the top ten search results (Silverstein et al., 1999). Reformulated search queries constitutes 40-52% of all search queries (Spink et al., 2000). Research on query modifications examines transitions between consecutive user search queries based on the overlap in term such as term addition, term removal, and term substitution (Bruza and Dennis, 1997; He et al., 2002; Jones and Fain, 2003; Costa and Seco, 2008; Huang and Efthimiadis, 2009; Jansen et al., 2009). Jansen et al. (2007) reported that users often modified their queries by changing query terms (nearly 23% of all query modifications) rather than adding or deleting terms. Previous studies (Beitzel et al., 2004; Park et al., 2005) also revealed that users rarely used advanced features of search engines: only about 2%-8% of the queries contain query operators (Hoelscher, 1998; Jansen and Pooch, 2001; Beitzel et al., 2004).

The average length of search queries has also been measured. Examining approximately 730,000 queries, Baeza-Yates and Castillo (2001) found that search queries had an average length of 2.43 terms. Spink et al. (2001) analysed 1,025,910 queries and revealed this average length was 2.6 terms. Other studies (Silverstein et al., 1999; Cacheda and Vina, 2001; Beitzel et al., 2004; Park et al., 2005) reported similar numbers.

Anchor text in webpages is known to be useful in improving the quality of web searching (McBryan, 1994). In fact, many commercial search engines rely heavily on anchor text because, like real user queries and titles, anchor text generally consists of a few terms that summarise a webpage (Jin et al., 2002; Eiron and McCurley, 2003). However, anchor text is often more carefully assigned than a webpage title because many webpage titles are repeated or meaningless (Won et al., 2009). Eiron and McCurley (2003) reported that among 2,395,766 webpages which had anchor text, content, and title information, 60.6% of the terms in users' search queries were contained in the webpages' content and their anchor text, but not in their title. They also concluded that "The advantage of anchor text over titles grows with the number of terms in the query". Craswell et al. (2001) and Westerveld et al. (2002) pointed out that anchor text provided a significant boost to the quality of results for site finding or homepage finding tasks. Other research attempted to model anchor text and classify queries to enhance webpage retrieval (Fujii, 2008) or to mine anchor text trends for retrieval (Dai and Davison, 2010). Due the findings of these studies, the research in this thesis selected anchor text as one of the cues for the revisiting user study in Chapter 4.

People often adopt a strategy of search from scratch (Jones et al., 2001) using search engines. However even a perfect search engine is not always enough for revisiting because instead of trying to "teleport" or jump directly to target pages using search queries, people often prefer an "orienting" strategy (Barreau and Nardi, 1995; Ravasio et al., 2004; Teevan et al., 2004). The first large step is made to the local area containing the information and then, based on cues and contextual knowledge, other local small steps are taken to reach to target pages. A subsequent experiment (Ruddle, 2009) investigated how a group of students revisit webpages in a familiar website (the department website). The result supported findings about the "orienting" strategy and highlighted the fact that participants had many more difficulties finding the local area than the target page. It also reported that despite having browsed frequently a website for 8-20 months, participants could recall only a small amount of the content and structure of the department website.

An analysis of a one-year web search query log of 114 anonymous users revealed that as many as 40% of all queries were re-finding queries (Teevan et al., 2007), even though the search queries used to re-find were different from the original search queries in ways such as word order, stop words,

non-alphanumerics, word merge, stemming and pluralisation, words swaps, add/remove word, abbreviations, synonyms, etc. So the first challenge of re-finding is that it is difficult for users to remember the exact search queries used to find information in the first place (Aula et al., 2005). Then, even when recalling the correct search query, recognising the pages clicked on the results pages and effectively browsing further from those pages are other challenges (Obendorf et al., 2007). The problem becomes more severe when the search results themselves change due to new ranking algorithms or updated databases (Aula et al., 2005; Teevan et al., 2007). In another study (Tyler and Teevan, 2010), 22% of all search queries were re-finding and the authors explained that this lower percentage was due to the study being only for one month. They also noticed with multiple click search queries, the result which was clicked first following the search query was more likely to be useful later, and the results found at the end of a search session were more likely to be re-found.

To support the search again strategy, some tools have been developed. Re:Search Engine (Teevan, 2007a) customises search results of search engines by fetching relevant previously viewed results from its cache. Revisit Rack (Morgan and Wilson, 2010) uses the Yahoo Boss API to return results for each search query. The results are then paginated with 8 per page and their thumbnails are displayed together at the top of the page rather than beside each result so users can more effectively utilise visual recognition without scrolling. SearchBar (Morris et al., 2008) and Google Search History list all search queries and webpages clicked from them. All these tools have been proved to be more useful than search engines for the search again strategy, however with Re:Search Engine and Revisit Rack users might still have to repeat the search process (e.g., recalling search queries, recognising clicked results) and all these tools have not dealt with the cases when desired pages were browsed further from pages clicked on results pages.

Similar to automatically recorded web history, the search again from scratch approach can support revisiting for any recency and frequency. However it relies on search engines to search the WWW as a whole. Searching a webpage again on the WWW is much difficult than searching personal files and emails because the information of space of the WWW is huge and dynamic. It requires users either to form new search queries to find again webpages they visited by browsing or recall old search queries to re-find

webpages. Then they need to recognise a target webpage from results lists or maybe have to browse further.

2.4 Analysis of tools for revisiting

This section reviews the suitability of main history tools. Table 2.1 summarises history tools by presentation, recency, frequency, and scale that they support revisiting. The ways in which each history tool supports revisiting for different frequencies, recencies, and scales has been discussed in Section 2.3, so this section first classifies history tools in terms recency, frequency and scale then analyses in detail two approaches of presenting a history to users: list-based and visualization presentation.

Table 2.1 Presentation, recency, frequency, and scale supported by history tools for revisiting.

| Tool | Presentation | Recency | Frequency | Scale |
|-------------------------------|---------------------|----------------|------------------|----------------------|
| Back/Forward Button | List-Based | Recently | | Session |
| ThumbStrips | | Recently | | Session |
| Google New Tab | | | Most visited | 9 Pages |
| Most Frequently Visited Pages | | | Most visited | 12 Pages |
| Bookmarks | | | | Collection |
| URL auto-completion | | | | Complete History |
| History List | | | | Complete History |
| Google History | | | | Complete History |
| Flipora | | | | Complete History |
| WebMynd | | | | Complete History |
| xMem | | | | Complete History |
| CWH | | | | Complete History |
| Re:Search Engine | | | | Complete History |
| Revisit Rack | | | | Complete History |
| SearchBar | | | | Complete History |
| WebMap | Visualization | Recently | | Session |
| MosaicG | | Recently | | Session |
| PadPrints | | Recently | | Session |
| Nestor Navigation | | Recently | | Session |
| WebNet | | Recently | | Session |
| History Tree | | Recently | | Session |
| Domain Tree Browser | | Recently | | Session |
| WebPath | | Recently | | Session |
| SessionGraphs | | | | Collection |
| WebBook and Web Forager | | | | Collection |
| Data Mountain | | | | Collection |
| BrowseLine | | | | Firefox history file |

Note:

- Blank cells in the Recency and Frequency columns indicate that history tools support revisiting to webpages of any recency and/or frequency.

- Collection: individual webpage is manually added to an archive by users.
- Session: webpages of a browsing session are deleted when the browser is closed.

It is understandable that some tools focus on recently visited webpages because that accounts for 70% of all revisits (Mayer, 2007; Adar et al., 2008). To support revisiting recently visited webpages, web browsers provide Back and Forward buttons. These buttons allow users to go back immediately to the previous page in the navigational path of a current webpage. Other history tools such as Thumbstrips, WebMap (Dömel, 1995), MosaicG (Ayers and Stasko, 1995), PadPrints (Hightower et al., 1998), Domain Tree Browser (Gandhi et al., 2000), Nestor Navigator (Eklund et al., 1999), WebNet (Cockburn and Jones, 1996), and History Tree display all webpages that have been accessed since web browsers opened. That is why these tools only deal with the scale of a session with a small number of webpages.

The most frequently visited pages are shown in a dropdown list of web browsers when users click on the small arrow at the end of address bars. So people can quickly revisit them. Google new tab provides the same utility by displaying the thumbnails of most frequently visited webpages when users open a new tab of web browsers. However, to maintain the effectiveness, these functions offer a scale of less than 15 webpages. Note that these webpages are not necessarily those that were visited recently as long as they have been visited most in the past.

URL auto-completion, History List, Google History, Flipora, WebMynd, xMem (Ceri et al., 2006), CWH (Won et al., 2009), Re:Search Engine (Teevan, 2007a), Revisit Rack (Morgan and Wilson, 2010), and SearchBar (Morris et al., 2008) support revisiting for any recency, frequency and scale of a complete history. However the way they present a web history is rather simple. All webpages are displayed in a list.

Bookmarks, WebBook and Web Forager (Card et al., 1996), Data Mountain (Robertson et al., 1998), and Session Graphs (Mayer and Bederson, 2001) also let users revisit for any recency and frequency. However the scale of these tools is limited because the collection is manually created by users.

There are two approaches of presenting a web history to the users: list-based and visualization. While the list-based tools support different scales of revisiting (e.g., from collection, session to complete history), visualization

history tools have not attempted to address the scale of a user's complete history.

2.4.1 List-based

In the list-based approach, webpages are presented as items in a list and users need to scan through the list to identify the target page. Webpages are often sorted by recency. Most recently visited pages are at the top. Lists can be divided into different types based on how each item is constituted.

Typically, each item contains a favicon and some text about a webpage (e.g., title, description, URL, domain, frequency of visits, and date of the last visit). This type of list was employed in URL auto-completion, bookmarks/favourites, history lists, Google History, Flipora, WebMynd, xMem (Ceri et al., 2006), Re:Search Engine (Teevan, 2007a), SearchBar (Morris et al., 2008), and Back/Forward Button (right clicking on them to view the list). CWH (Won et al., 2009) enriches the typical list by adding a small thumbnail of each webpage and Revisit Rack (Morgan and Wilson, 2010) displays the thumbnails of webpages together at the top of each results page rather than beside each result for more effective recognition.

Thumbstrips and Google New Tab⁵ even use thumbnails as the main element of list items. A study on how people recognise previously seen webpages from titles, URL, and thumbnails revealed that thumbnails were the most important cue for user recognition of visited webpages (Kaasten et al., 2002). Since then thumbnails have been exploited to support searching and revisiting in many other studies (Woodruff et al., 2001; Dziadosz and Chandrasekar, 2002; Woodruff et al., 2002; Teevan et al., 2009; Aula et al., 2010; Jiao et al., 2010; Loumakis et al., 2011; Badesh and Blustein, 2012). This is why the present research chose webpage thumbnails as another cue for the revisiting user study in Chapter 4.

Interaction in the list-based approach is typically simple. Scrolling up and down a list is provided by most list-based history tools and going to next or previous results pages is popular for tools which support search functionalities (e.g., Flipora, Google History). Clicking on a hyperlink is a common way to open a visited page in the web browser.

The advantage of this approach is that it is simple, conventional, and scalable to support long-term revisits. The main drawback is it loses contextual information such as user's navigational paths.

⁵ See <http://support.google.com/toolbar/?hl=en>

2.4.2 Visualization

In the visualization approach, webpages are presented either as elements of spatial metaphors or nodes in connected networks. Representing a webpage by its thumbnail, some tools utilise familiar metaphors to present a user's web history. The metaphors of book, bookcase, and desk are used in WebBook and Web Forager (Card et al., 1996). Data Mountain (Robertson et al., 1998) utilises the mountain. Relationships between webpages are shown by the way they are arranged in the display space. Users can left click to flip through webpages in a book in WebBook and Web Forager (Card et al., 1996) or bring a webpage to front view in Data Mountain.

The path taken to find information is particularly important for re-finding (Capra and Pérez-Quiñones, 2005). As people are able to remember only important "waypoints" (Maglio and Barrett, 1997), visualization history tools attempt to visualize the complete paths. They are particularly helpful when people revisit a webpage by making a series of small "orienteeing" steps (Teevan et al., 2004). Besides freeing user's cognitive capacities, visualizing webpages in their navigational paths enables users to navigate more easily in an information space by jumping to any webpage they can see rather than following hyperlinks page by page. This is why users' navigational paths were chosen as a cue for the revisiting user study in Chapter 4.

The History Tree (Panasiti, 2009) and many research tools such as WebMap (Dömel, 1995), PadPrints (Hightower et al., 1998), Domain Tree Browser (Gandhi et al., 2000), Nestor Navigator (Eklund et al., 1999), WebNet (Cockburn and Jones, 1996), SessionGraphs (Mayer and Bederson, 2001) adopted a graph-based approach to draw navigational paths. The user interface of this kind of tool is often more complicated using information visualization techniques. The human perceptual system is highly attuned to images, and visual representations can communicate some kinds of information more rapidly and effectively than text (Hearst, 2009). The rest of this section analyses those tools further in terms of key aspects of information visualization (e.g., representation, presentation, and interaction).

Representation

A simple representation of a webpage is a small circular node in a spanning tree (e.g., as in WebMap (Dömel, 1995)). Each node is assigned a number based on the order it is visited. The benefit of this approach is that the size of a node is small which leads to a compact tree. However, the only way to determine which webpage a node represents is to move the mouse over and

look at a complementary list view. WebNet (Cockburn and Jones, 1996) and SessionGraphs (Mayer and Bederson, 2001) improve on this by adding a short label for each node. The label is a short version of a webpage's title. Nestor Navigator (Eklund et al., 1999) allows users to use different icons for nodes and let them annotate nodes. But this labelling makes the tree cluttered when text crosses edges and nodes. Later tools like PadPrints (Hightower et al., 1998) and Domain Tree Browser (Gandhi et al., 2000) represent a webpage by its thumbnail. Webpages' thumbnails were proved to be useful for user recognition, however a node often takes more real estate.

Relationships between webpages are represented by edges in a network. In History Tree, a webpage can be displayed more than once in a branch and in different branches depending how many times it has been visited and in which tabs. This approach might confuse users because they might need to locate the same webpage at different locations to find another webpage visited from that webpage. Other tools like WebMap (Dömel, 1995), PadPrints (Hightower et al., 1998), and Domain Tree Browser (Gandhi et al., 2000) utilise spanning trees. However, the way of building the tree is rather simple. If a page is visited the second time it is simply ignored and the "current node" is set back to its first drawn. No weight function was considered. The reason might be most tools were designed for short-term revisit. The parent and child relationship in each branch partly represents recency of webpages. Other tools such as Nestor Navigator (Eklund et al., 1999), WebNet (Cockburn and Jones, 1996), and SessionGraphs (Mayer and Bederson, 2001) use a connected network. This means if a user navigates in a loop, the graph also shows this circle. This makes the network become busy and complicated over time.

Information coding is used in several tools. Webmap (Dömel, 1995) uses colour-coding to distinguish between normal nodes (blue), the current node (red), and mouse over nodes (pink). Links between nodes are also colour-coded: black for intra-site and green for inter-site. Domain Tree Browser (Gandhi et al., 2000) uses colours to highlight current domain and current page. To convey more information of a web history, Domain Tree Browser, and SessionGraphs (Mayer and Bederson, 2001) encode frequency by node size. The more the number of visits to a node, the bigger the node becomes. Depending on a user's preferences, node sizes in Nestor Navigation (Eklund et al., 1999) and WebNet (Cockburn and Jones, 1996) represent frequency or recency.

Presentation

Space is always not enough for a visualization system no matter how large the screen is (Spence, 2007). That's why tools use different techniques to present a web history. Scrolling is provided by History Tree, ThumbStrips, WebNet (Cockburn and Jones, 1996), and WebBook & Web Forager. Focus+Context is employed in Webbook and Web Forager (Card et al., 1996) and SessionGraphs (Mayer and Bederson, 2001) in two different techniques. Webbook and Web Forager use distortion in the form of "document lens" to help users inspect portions of interest while keeping the context. In SessionGraphs, a suppression technique is utilised. There are three different panes: task pane, session pane, and session view pane. Selecting each task in the first makes the second pane view only the session belonging to that task and clicking on each session in the session pane commands the third pane to view only webpages of that session. Finally, Zoom and Pan are used in PadPrints (Hightower et al., 1998) and Domain Tree Browser (Gandhi et al., 2000).

Interaction

The interaction mode is one of the first things to be considered in visualization tools. Passive interaction with a static display is used in History Tree, WebMap (Dömel, 1995), PadPrints (Hightower et al., 1998), Nestor Navigator (Eklund et al., 1999), WebNet (Cockburn and Jones, 1996), WebBook & Web Forager, and Data Mountain (Robertson et al., 1998). ThumbStrips provides both passive interaction with a static display and a moving display. With these tools, all users can do is to look at the visualization and select a desired page. More complicated tools such as Domain Tree Browser (Gandhi et al., 2000) and SessionGraphs (Mayer and Bederson, 2001) utilise composite interaction. First, stepped interaction is required to select a subset of data and then passive interaction is followed for further exploration. In Domain Tree Browser, users first select a domain in the domain panel to view the corresponding history tree in the tree panel while with the SessionGraphs (Mayer and Bederson, 2001), users need two steps to view a sub history tree: selecting task then selecting session.

Left clicking is the most common interaction to open a webpage in a browser in visualization history tools (History Tree, ThumbStrips, PadPrints (Hightower et al., 1998), Domain Tree Browser (Gandhi et al., 2000), and WebNet (Cockburn and Jones, 1996)), however double clicking is used in WebMap (Dömel, 1995). Mouse over is used to view a webpage's details in Domain Tree Browser, SessionGraphs (Mayer and Bederson, 2001).

WebMap provides a play back feature which answers the navigation question “where am I”. When a node is selected in a list, a path to that node is highlighted in the map. Filters have not been supported much so far. History Tree lets users filter webpages by tab ID and Webnet (Cockburn and Jones, 1996) does this with frequency and recency.

Some time ago, a history tree (Brodlić et al., 1993) was demonstrated to be useful in supporting scientists in solving time-dependent problems. Similarly, evaluations have shown that visualization history tools support revisiting better than conventional list-based history tools (e.g., PadPrints (Hightower et al., 1998) vs. Netscape Navigator 3.0, Data Mountain (Robertson et al., 1998) vs. Internet Explorer, SessionGraphs (Mayer and Bederson, 2001) vs. Netscape 4.7). Users found desired webpages significantly faster with fewer steps and fewer visited webpages. They were also more satisfied with visualization history tools. But all these tools share two main drawbacks. First, as shown in Table 2.1, they do not support revisiting a complete history. Second, they have not dealt well with tabbed browsers where navigational paths could cross each other. Existing visualization history tools often add a new node to the current active node. This makes navigational paths misleading. If these two limitations can be solved, the visualization approach could be a better solution than the list-based.

2.5 Technologies used in history tools

History tools are different from standalone systems because they need to keep track activities of web browsers. This section summarises technologies used to develop history tools.

Domain Tree Browser (Gandhi et al., 2000) and Nestor Navigator (Eklund et al., 1999) embed a light weight version of web browsers (Java Web Browser and Internet Explorer) inside them. Thanks to that they can directly communicate with browsers to extract information and keep track of browsers’ events. Domain Tree Browser uses Jazz ⁶ library in Java to do visualization while Nestor Navigator uses Visual Basic 5. With this approach, the tools have full control of the browsers however users do not have their familiar and fully functioned browsers. Updating to a new version of the browser is another problem.

⁶ See <http://www.cs.umd.edu/hcil/jazz/>

Exploiting remote procedure calls, WebMap (Dömel, 1995) and WebNet (Cockburn and Jones, 1996) communicate with browsers (Mosaic and tkWWW) by commands. This type of communication allows processes to exchange data without using kernel communication support like sockets or pipes. They both use Tcl taking advantage of the Motif-like Tk widget library, especially the widget canvas for visualization. The main problem of this approach is that users have to remember to explicitly run both the history tool and the web browsers.

Another way of extracting information loaded in a web browser is using a proxy that acts as an intermediary for requests from web browsers and the WWW. The proxy helps extract information of webpages. PadPrints (Hightower et al., 1998) and SessionGraphs (Mayer and Bederson, 2001) use a modification of WBI⁷ proxy developed by IBM for this purpose. While PadPrints (Hightower et al., 1998) uses Pad++ (Bederson and Hollan, 1994) for zooming interface using Tcl scripting language, SessionGraphs uses Jazz. The main advantage of this approach is that the history tools can be used with any web browsers. However, similar to using the remote procedure calls approach, users have to remember to explicitly run both the history tool and the web browsers.

Recently, the browser extension concept was introduced and has become popular to users thanks to Firefox. It is a computer programme that extends the functionality of a web browser in some way and can be integrated seamlessly into a web browser. When users start their browsers, an extension is automatically started so users do not need to remember to run another programme. The browser extension can both control and keep track activities of a web browser. This approach is adopted in Flipora, History Tree, WebMynd and ThumbStrips. With the new version of HTML 5, visualization can also be done easily within these extensions. The disadvantage of this approach is that an extension developed for a specific browser cannot be added to another browser. However, today the Crossrider⁸ development platform lets developers make extensions that can work across different web browsers such as Chrome, Internet Explorer, Safari and Firefox.

⁷ See <http://www.almaden.ibm.com/cs/wbi/>

⁸ See <http://crossrider.com/>

From the above analysis, developing a web history tool as a browser extension would be the best choice at the moment.

2.6 User study methodologies

Two user studies have been carried out to complete this thesis. It is important to consider which methodologies should be employed because they will influence the findings in different ways. All studies have their own advantages and disadvantages. This section surveys methods that have been used in revisiting research.

2.6.1 Approaches

Log analysis, observational, controlled laboratory, interview, questionnaire, survey, and diary methods have been employed in many studies on revisiting electronic information. However all research methods have strengths and weaknesses, and sometimes two or more methods are combined to provide a much better understanding of phenomena (Lazar et al., 2010). Log analysis and observational studies show real life behaviour but have limited ability to understand the motivation behind the behaviour. Controlled laboratory studies are difficult to recruit participants for and may change the actions of users (Carter and Mankoff, 2005). Interview or survey studies give insights into participants' motivation, but self-reported data can lead to bias. Diaries fill the gaps between observation and interview/survey (Hyldegård, 2006).

Many interesting findings on personal information management have been discovered through observational method in studies on paper documents (Malone, 1983; Lansdale, 1998), emails (Whittaker and Sidner, 1996), files (Barreau and Nardi, 1995), webpages (Jones et al., 2001; Sellen et al., 2002). Using a modified diary method, a more recent study (Teevan et al., 2004) focused on directed search and looked at user behaviour across a broad class of electronic types (email, files and the WWW).

Some other studies employ log analysis to investigate revisit patterns in the real world. This method was used in previous studies to explore revisiting patterns (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006; Adar et al., 2008).

Sometimes, studies are only interested in investigating in depth for special circumstances. In those cases, the controlled laboratory method is required. This approach allows researchers to conduct controlled experiments and examine users' thought processes during sessions by asking them to think

aloud. Thanks to that more information can be obtained. For example, several controlled studies were conducted to test whether an algorithm developed to support re-finding keeps participants from noticing change and allows participants to conduct re-finding tasks as quickly as a static result list while still enabling the finding of new information (Teevan, 2007a).

The empirical study described in Chapter 4 of this thesis combined different methodologies. The controlled laboratory method was required because the study needed to compare the effect of different cues and recencies to revisiting. In addition, the observational method was essential to investigate how participants revisit webpages. Finally, the log analysis method was important to examine revisiting patterns and the difficulties of revisiting.

Surveys were used in previous studies to explore the major problems with using the WWW⁹ (e.g., 17% of participants were not able to return to some pages they had visited before), and to identify the intent behind observed revisitation (Adar et al., 2008). A questionnaire was the medium to explore previous experience of participants with history mechanisms, knowledge of them about page contents, and their satisfaction with the tool (Ceri et al., 2006). Interviews shed light on the “orienting” behaviour (Teevan et al., 2004), and “keep found things found” activities (Jones et al., 2001). Modified diary study was employed to investigate how people performed personally motivated searches in their email, in their files, and on the WWW (Teevan et al., 2004).

To evaluate the history tool proposed in this thesis (see Chapter 6), an electronic diary methodology was utilised because revisits were not predictable. Diary entries helped participants capture situations where revisits occurred and they needed the history tool. Then a follow-up interview was also conducted to clarify aspects of the diary entries and to learn what people thought about the tool. Log analysis was also employed to examine revisiting patterns with the tool.

2.6.2 Participants

One of the biggest challenges in conducting this type of research was recruiting participants. The first reason is that it relates to privacy of participants. Accessing the WWW is a personal activity so not many people are willing to expose that kind of information to others. Although researchers

⁹ See GVU's Tenth WWW User Survey:
http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/

are often aware of that and provide some ways to protect privacy and anonymity, users might be still unsure about what data are captured. Studies of revisiting patterns (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006) recorded all webpages that participants visited. Second, these studies often last a long period from one month (Catledge and Pitkow, 1995) up to 3 months (Weinreich et al., 2006). During that period, a lot of issues might happen such as user illness, data crash, and machine crash. Because of those reasons, the number of participants in controlled laboratory studies is often rather small (e.g., 15 in the “orienting” study (Teevan et al., 2004), 12 in CWH (Won et al., 2009), 11 in “keep found things found” study (Jones et al., 2001), 10 in SessionGraphs (Mayer and Bederson, 2001), and 4 in Domain Tree Browser (Gandhi et al., 2000)).

2.6.3 Tasks

In the studies which need face to face interaction between researchers and participants, identifying appropriate tasks for participants to perform during user experiments is another challenge. In the “keep found things found” study (Jones et al., 2001), participants were asked to list at least three work-related and web-intensive “free-time” tasks they might like to work on over the week after that should they have an half hour or more of unscheduled time and other web tasks they might expect to perform in a typical work week whether or not work related. During the observation sessions, one of those “free-time” tasks was selected for the participant to perform. They were also asked to think aloud and video recorded. Thanks to that, researchers could observe how participants organise found things and access them again.

In a study (Capra and Pérez-Quiñones, 2005), two sessions were conducted to investigate how people use Web search engines to find and re-find information. In the first session, each participant was given 18 tasks to look for certain information on the WWW. Then in the second one, about a week later, they were asked to re-find the same or similar data. In a later study (Ruddle, 2009), to explore how people find information on a similar website, participants were asked to navigate within their school’s website to look up answers for 12 questions.

When evaluating SessionGraphs (Mayer and Bederson, 2001), participants were first asked to answer a set of questions within a website using Netscape and the tool respectively. One to six days after that, they were asked to do the same things to compare SessionGraphs vs. Netscape

history functionalities. The same method was used to compare SearchBar (Morris et al., 2008) with Internet Explorer 7.

In short, so far there are two main methods which webpages can be chosen for revisiting: either by participants or researchers. If participants choose target pages, they know that they will need to re-access those webpages later. On the other hand, when participants are asked to visit some webpages or do some tasks by researchers in the first time, they can guess they will have to do the same things next time. In both cases, participants might remember target webpages or how to get to them and have certain preparation for revisiting later. This may make revisiting a little easier. The study described in Chapter 4 of this thesis did not adopt any of the above methods. As it focuses on the 4th group of webpages, a computer programme was developed to select webpages which had been visited neither frequently nor recently (see Section 4.1.2.1). A dwell time of 30 seconds or more on a webpage was also used as a condition of selecting target pages. This increased the likelihood that target pages were of interest to a participant. So any qualified webpages in a web history could be selected then described for participants to revisit.

2.6.4 Measures

To evaluate the visualization history tool described in Chapter 5, some measures (e.g., effectiveness, efficiency, satisfaction and ease of use) of interaction design (Preece et al., 2002) are employed.

Effectiveness is about whether the tool helps users accomplish particular tasks. One common way to measure effectiveness is to count the number of tasks a user is able to accomplish successfully (Robertson et al., 1998; Mayer and Bederson, 2001). A tool is efficient if it helps users complete their tasks with minimum waste, expense or unnecessary effort. Efficiency is often measured by recording the time it takes to complete a task (Robertson et al., 1998; Mayer and Bederson, 2001; Ceri et al., 2006) or the number of actions or steps taken to complete a task (Hightower et al., 1998). Satisfaction can be understood as the fulfilment of a specified desire or goal. Ease of use is related to the amount of effort which users expend executing and/or accomplishing particular tasks. It is common for satisfaction and ease of use measures to be gathered via the Likert scale (Hightower et al., 1998; Robertson et al., 1998; Ceri et al., 2006).

2.7 Discussion

Revisiting is so common that between one third and one half of visits are return visits to pages that has been previously seen (Catledge and Pitkow, 1995; Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Weinreich et al., 2006). Despite the plethora of techniques that people use to assist revisiting (Jones et al., 2001), occasionally they still have frustration at not knowing where to “go” in order re-access a webpage (Bruce et al., 2004; Teevan, 2007b).

Two factors that clearly affect the ease of revisiting are the frequency and recency with which a webpage has been visited. This research considers four combinations of them, according to whether a webpage was previously visited:

1. Both frequently (on more than 1 day) and recently (less than 1 week).
2. Frequently but not recently.
3. Recently but not frequently.
4. Neither frequently nor recently.

Pages in the 1st category can often be recalled by a person, either in terms of the URL or the “orientteering” (searching and browsing) steps that are required to reach such a page (Teevan et al., 2004). Web browser functionality such as Back/Forward buttons, URL auto-completion, and lists of recent and most visited pages, complement a person’s memory and simplify the task of revisiting pages that have been visited frequently and/or recently (i.e., the 1st, 2nd and 3rd categories above). History tools such as History Tree, Thumbstrips, WebMap (Dömel, 1995), PadPrints (Hightower et al., 1998), WebNet (Cockburn and Jones, 1996) also assist people for these kinds of revisitation. The greatest problems occur when a page is in the 4th category. With webpages which had low frequency of visits, the failure rate was much higher than ones with medium and high frequency (Bruce et al., 2004). Evidence also proved that recency of visits influenced the difficult of revisiting (Elsweiler and Ruthven, 2007). The lack of frequency and recency makes it difficult to search/browse for it from scratch.

Several tools have been proposed and developed to support revisiting. So far history tool designs were based on users’ revisiting patterns (Mayer and Bederson, 2001; Teevan, 2007a; Morris et al., 2008), classification and management of webpages (Mayer and Bederson, 2001; Ceri et al., 2006), potentially useful cues (Won et al., 2009; Morgan and Wilson, 2010), and enhancing current support of web browsers such as bookmarks (Card et al.,

1996; Robertson et al., 1998), back/forward or recent visits (Dömel, 1995; Cockburn and Jones, 1996; Hightower et al., 1998; Eklund et al., 1999; Gandhi et al., 2000). However, each tool has its own disadvantages as discussed in Section 2.3 and 2.4. In short, list-based tools can support revisiting for any frequency, recency, and scale but they do not provide contextual information. By contrast, visualization tools can provide more contextual information (e.g., users' navigational paths) which supports user recognition but have not been able to deal with a completed web history (see Section 2.4) and tabbed browsers.

This research adopts a new approach. First, the information would be useful for revisiting and the difficulties people encounter are investigated through an empirical study on the 4th group. Based on the findings of this study and the analysis in Section 2.3 and 2.4, a new history tool is designed. The new tool exploits the automatically recorded web history (see Section 2.3) and visualization approach (see Section 2.4). Different navigation techniques are employed to deal with a complete web history.

To conduct the empirical study, a web history logging tool is required. The next chapter describes the requirements, design and implementation of such a tool.

Chapter 3. The logging tool

To be able to run the empirical study to investigate the underlying causes of failure when revisiting webpages, a user's web history needs to be captured. This chapter describes the requirements, design and implementation of a logging tool that captures such history data.

3.1 Requirements

The requirements of the logging tool were as follows. As one of the goals of the forthcoming laboratory experiment (see Chapter 4) was to test the usefulness of different cues (e.g. thumbnail, anchor text, and navigational path) when revisiting webpages at different recency and frequency, the tool needed to extract the following information: webpage thumbnail, visited time (to calculate recency), webpage referrer (to reconstruct navigational paths), and anchor text of the link leading to the page. A thumbnail of the whole webpage is required rather than only the visible part. If the same webpage is visited several times, to save space its thumbnail should not be captured again if its contents do not change. The URL of a webpage is of course important to be captured. The frequency of a webpage's visits can be determined by counting the number of occurrences of a URL, and the number of different days on which a webpage has been visited can be derived from date stamps.

As described in Section 4.1.2.1, dwell time on a page is one of the criteria that was used to choose target pages for participants to revisit, so the tool needs to keep track of this information carefully. Dwell time should be calculated for a webpage only when (1) it is open in a tab, (2) that tab is the active tab of the browser, and (3) the browser is the active application of the computer. When a webpage is opened in hidden tabs and never looked at its dwell time is zero, and reopening a webpage in a hidden tab is not treated as a revisit. This overcomes weaknesses in previous studies (Catledge and Pitkow, 1995; Cockburn and McKenzie, 2001; Weinreich et al., 2006; Liu et al., 2010).

A webpage's title is used to describe the content of a webpage so it is recorded. The tool also stores other necessary information such as all hyperlinks within a webpage to examine clicked webpages in search results pages (see Section 4.2.3) and to identify anchor texts (see Section 3.2.1), a

search query of a Google search page and a webpage's description to present the history (see Chapter 5).

Whenever a webpage is loaded into a web browser, other content might be inserted in that page (e.g., ads, frames). To remove this noise and to address the problem of invalid logged entries mentioned in (Weinreich et al., 2006), the logging tool should only capture information of the page whose URL is displayed on the address bar. A webpage might be visited a few times in the history but it should be saved individually because each time its contents can be different and it can be in another navigational path. To protect user privacy, secure webpages (https) should not be captured and a form is required to allow users to specify webpages and websites they don't want to be recorded.

As a whole, the tool needs to track all activities of a web browser. However, it must run in the background without interfering with the user's browsing activity. The users should not feel any difference after installing the logging tool. Although running in the background, the logging tool needs to provide a simple user interface for users to start/pause capturing history, to edit their history for privacy, and to block private websites.

3.2 Design and implementation

The logging tool has been designed and developed as an extension for a web browser. An extension is a piece of software which can be integrated seamlessly into a web browser to modify the behaviour of existing features or to add entirely new features. Extensions are especially popular with Firefox, because it has been designed as a minimalistic application to reduce software bloat and bugs, while retaining a high degree of extensibility, so that individual users can add the features that they prefer themselves. In Firefox, these extensions are called add-ons¹⁰. The logging tool has been developed as an add-on for Firefox because Firefox has been one of the most popular web browsers¹¹. When users start the browser, the tool will automatically run in the background as part of it. They do not need to remember to run another programme with the browser. The web history of a user is stored in the add-on folder which belongs to that user profile. The Firefox browser protects user privacy by creating a private profile for each

¹⁰ See http://en.wikipedia.org/wiki/Add-on_%28Mozilla%29

¹¹ See http://www.w3schools.com/browsers/browsers_stats.asp

user even on the same machine. So the web history of a user is not accessible and visible by others. The following technologies have been used to develop the add-on:

User interface: XUL¹² (XML User Interface Language) and HTML are used to create components of the tool. This is actually an XML grammar that allows graphical user interfaces (such as buttons, menus, toolbars, tree, etc.) to be written in a similar manner to webpages. User actions are bound to functionality using JavaScript directly within XUL tags. XUL is developed by Mozilla and can be used to write cross-platform applications. More important it can overlay the Firefox browser user interface to make an add-on a seamless component of the browser.

Application logic: As XUL and HTML have been selected to make the user interface and the tool is a client-side application, JavaScript is an obvious choice to write application logic. It can be executed in any browser and it is fast to the end user. To implement the application logic, JavaScript also exploits DOM to access XUL elements and imports Mozilla Firefox APIs to interact with the Firefox browser.




Database: At first, XML was investigated to store data but, to deal with complex queries and large quantities of history data, a more complex relational database management system (RDBMS) is needed. Popular RDBMS such as Microsoft SQL and Oracle were considered but they are designed for more complex client-server systems. The problems of installation and connection are too much for users. Microsoft Access could be another choice, but it only works on Windows. Instead, SQLite has been chosen because it is “a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine”¹³. The best part is that SQLite stores the entire database (definitions, tables, indices, and the data itself) as a single cross-platform file on a host machine. Therefore it is easily integrated into the add-on.

To meet the requirements stated in Section 3.1, the logging tool has been designed with three components: (1) a simple tool bar for users to control the logging tool, (2) a web history editor letting users review their history and delete unwanted entries for their own privacy, and (3) a form for them to block private websites.

¹² See <https://developer.mozilla.org/en/docs/XUL>

¹³ See <http://www.sqlite.org/>.

3.2.1 The tool bar

The tool bar has three icons placed in the status bar of the Firefox web browser (see Figure 3.1): button  to start/pause the capturing,  to open the history editor form, and  to go to the data folder to send the logfile back at the end of the study. A menu item is added to the menu “Tool” of the web browser for blocking webpages/websites. By default, to capture their web history, users do not need to do anything after installing the add-on.

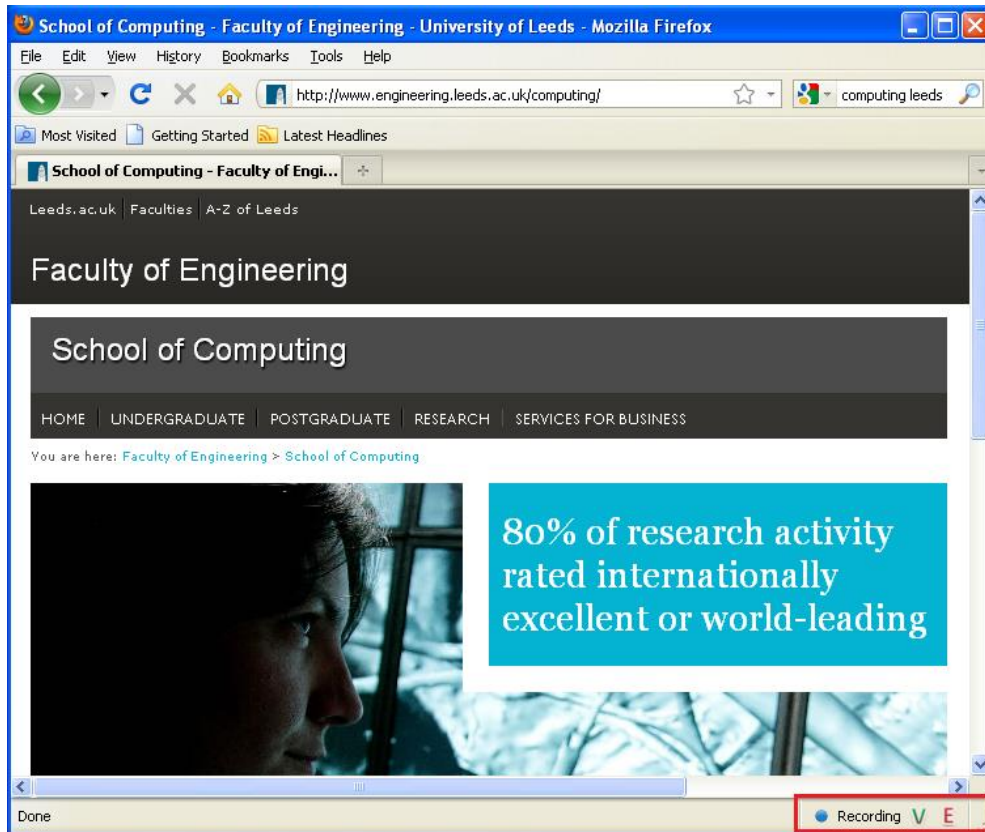


Figure 3.1 Three icons of the logging tool placed in the status bar of the Firefox browser: the first to start/pause the capturing, the second to go to the data folder, and the last to open the history editor form.

To insert the tool bar in the status bar of the Firefox browser, a file named `overlay.xul`¹⁴ needs to be created to describe extra contents for the browser user interface. After the add-on is installed, the Firefox browser automatically reads this file and adds the extra contents. The overlay file also maps user interaction, browser events to application logic scripts.

Load is the first event that the logging tool needs to track the web browser. When a webpage is loaded into a tab, this event occurs and information of

¹⁴ See XUL Overlays: https://developer.mozilla.org/en/docs/XUL_Overlays

the webpage is extracted. However to capture the dwell time on a webpage, extracted information will not be saved to the database until the webpage is closed. To remove noise mentioned in Section 3.1, when the load event is triggered, the logging tool checks each tab of the browser to see if there is any new URL which needs recording. If yes, this new URL is pushed into a tracking list and its information is saved into the database later, otherwise the URL that activates the event is ignored because it is either a noise or a private page. Another problem is that sometimes it takes time to load a webpage into a tab of the web browser, immediately capturing thumbnail of a webpage might result in a blank image. The tool attempts to solve this problem by calling a thumbnail capturing function in another thread after 0.5 second for a webpage and 2 seconds for a PDF file. Pseudo code 3.1 summarises the script for extracting information of a webpage.

Pseudo code 3.1 The load event.

```
function onLoad()
{
    if(Recording == false)
        return;
    //get all tabs of browser
    var numOfTab = gBrowser.browsers.length;
    //Check if a tab is not in the trackingList and needs recording
    //if yes then push it into the trackingList
    for (var i = 0; i < numOfTab; i++)
    {
        var htmlDoc = gBrowser.getBrowserAtIndex(i);
        var checkingURL = htmlDoc.currentURI.spec; //Get URL of the page
        if(checkingURL!="about:blank" && checkingURL not in trackingList
            && checkingURL not in blockedList)
        {
            //Extract webpage's information
            var mTime = new Date().getTime();
            getReferrerID();
            getAnchorText();
            getDomHashForPage();
            getSearchTerm();
            getOtherAttribute: URL,title,referrer,allLinks,description;
            createNewPage();
            curPages.push(newPage);
            trackingList.push(checkingURL);
            if(checkingURL is a PDF file)
            {
                window.setTimeout(function(){saveThumbnail();},2000);
            }
            else
            {
                window.setTimeout(function(){saveThumbnail(); },500);
            }
            break;
        }
    }
}
```

When extracting information of a webpage, some webpage's attributes (e.g., URL, title, description, and all links) are direct. Some are not straightforward (e.g., thumbnail, referrer, anchor text, and dwell time). To capture the full

thumbnail of a webpage, the HTML document needs to be drawn onto a canvas then saved as a picture. To save storage space, the logging tool resizes the thumbnail down to 70% of the original size. Pseudo code 3.2 shows how to capture a webpage thumbnail.

Pseudo code 3.2 Creating and saving a webpage thumbnail.

```
function getAndSaveThumbnail()
{
    var canvas = create a HTML canvas;
    //Scale down the size of document
    canvas.width = DocumentWidth * 0.7;
    canvas.height = DocumentHeight * 0.7;

    //Save the full thumbnail of a webpage
    var ctx = canvas.getContext("2d");
    ctx.clearRect(0, 0, canvas.width, canvas.height);
    ctx.save();
    ctx.scale(0.7, 0.7);

    ctx.drawWindow(content of Document, 0, 0, canvas.width,
                    canvas.height, "rgb(255,255,255)");
    ctx.restore();
    //Use Firefox API to save the canvas as a .png file
    saveCanvasToFile();

    //To save the visible part of a webpage in the browser
    //Only the height of canvas needs to be recalculated
    //canvas.height = canvas.width * screen.height / screen.width;
    //Then do the same as above
}
```

As stated in the requirements, a webpage thumbnail will not be captured again if its contents don't change. As the full text of a webpage is not stored in the database, a simple method is used to identify if the contents of a webpage have changed. First, the html DOM of a webpage is serialised to a string. Then a hash string of 32 hexadecimal characters for this string is calculated using MD5. This hash string is stored in the database to compare with the new hash string of the next revisit.

Pseudo code 3.3 illustrates how to calculate MD5 Hash string for a webpage.

Pseudo code 3.3 Calculating MD5 Hash for a webpage.

```
function getDomHashForPage()
{
    //Convert XML to string
    var objDOM = HtmlDom of Webpage
    var serialisation = new XMLSerializer();
    var logXMLString = serialisation.serializeToString(objDOM);
    //The function below is of Firefox Mozilla
    var converter =
Components.classes["@mozilla.org/intl/scriptableunicodeconverter"]

.createInstance(Components.interfaces.nsIScriptableUnicodeConverter);
    converter.charset = "UTF-8";
    // result is an out parameter,
    // result.value will contain the array length
    var result = {};
    // data is an array of bytes
    var data = converter.convertToByteArray(logXMLString, result);
    //The function below is of Firefox Mozilla
    var ch = Components.classes["@mozilla.org/security/hash;1"]
        .createInstance(Components.interfaces.nsICryptoHash);
    ch.init(ch.MD5);
    ch.update(data, data.length);
    var hash = ch.finish(false);
    // convert the binary hash data to a hex string.
    var s = "";
    for (var i=0; i < hash.length ; i++)
        s = s + convertToHexString(hash.charCodeAt(i));
    return s;
}
```

Tracking dwell time on webpages in a tabbed browser has previously been performed (Weinreich et al., 2006). However, switching between tabs of the browser, and between the browser and other applications, has not been taken into account. To address conditions mentioned in Section 3.1, three events of the browser need tracking. In Firefox, when users switch between opened webpages, the *locationchange* event is triggered. Then the active tab needs to be identified and dwell time is calculated for the webpage in this tab. If users switch between the browser and other applications, the *blur* and *focus* events are triggered. When the *blur* event occurs, the dwell time calculation must pause until the *focus* event is activated. Pseudo code 3.4, 3.5, and 3.6 work together to track dwell time on webpages.

Pseudo code 3.4 LocationChange event.

```
function locationChange()
{
    var newURL = URL of the active tab;
    if (newURL == oldURL)
    {
        return;
    }

    if (oldURL is in trackingList)
    {
        //Update dwell time for the previously active webpage
        var curTime = new Date().getTime();
        var lapseTime = (curTime - timeCounter)/1000 - timeBlurCounter ;

        //As a webpage can could be switched to active several times
        //Accumulative Dwelltime += lapseTime;
        updateDwelltimeForPage(oldURL,lapseTime);
        //Check if the tab displaying the oldURL is closed
        if(oldURL is not in any tabs of browser)
        {
            RemoveFromTrackingList(oldURL);
            SaveWebpageToDatabase(oldURL);
        }
    }
    //Start counter anyway
    timeBlurCounter = 0;
    timeBlur = 0;
    timeCounter = new Date().getTime();
    oldURL = newURL;
}
```

Pseudo code 3.5 Blur event.

```
function firefoxLostFocus()
{
    //Mark the time the browser is inactive
    //Dwell time of a webpage is calculated only when
    //it is in an active tab and the browser is active
    if(timeCounter > 0)
        timeBlur = new Date().getTime();
}
```

Pseudo code 3.6 Focus event.

```
function firefoxGetFocus(source)
{
    //run this function only if the browser lost focus earlier
    if(timeBlur > 0)
    {
        var tmp = (new Date().getTime() - timeBlur) / 1000;

        if(tmp >= 1)
        {
            timeBlurCounter = timeBlurCounter + tmp;
            //For new to stop this event from being done again
            timeBlur = 0;
        }
    }
}
```

With a normal webpage, it is not difficult to identify its referrer by getting the “referrer” attribute. However, if a webpage is clicked from a Google search engine results page (SERP), tracking its referrer is more challenging. For example, the Google SERP for a search query “how to create update Firefox add-on” has the syntax like:

http://www.google.com/search?q=how+to+create+update+firefox+add-on&ie=utf-8&oe=utf-8&aq=t&client=firefox-a&rlz=1R1GGHP_en-GB_GB462

If a webpage is selected from the above SERP, its referrer attribute should be the URL above. However, wanting to keep track of webpages clicked from a SERP itself, sometimes Google SERP directs these pages via other webpages. For example, one of pages clicked in the above SERP has the following referrer:

http://www.google.com/url?sa=t&rct=j&q=create%20update%20firefox%20add-on&source=web&cd=18&ved=0CGEQFjAHOAo&url=http%3A%2F%2Fstackoverflow.com%2Fquestions%2F6484749%2Fxp-create-update-rdf-for-previous-version&ei=ZZ7tT-yvNIO30QXG_oHQDQ&usq=AFQjCNGS2SL8348TltkvoGtyVO5NvAAYNA&cad=rja

This is rather different to the above SERP. Such a webpage can be determined when its referrer attribute contains the two tokens “?sa=” and “&q=”, and needs further analysis. The search query is then extracted from this referrer to compare with the search query of the current SERP to find out the exact referrer. Pseudo code 3.7 implements this procedure.

Pseudo code 3.7 Tracking referrer of a webpage.

```
function getReferrerID(objPage)
{
    var refURL = objPage.referrer;
    if(refURL != "")
    {
        //Get searchQuery from an referrer URL
        var searchQuery = getResultTerm(refURL);
        if(refURL not in trackingList && searchQuery != "")
        {
            //the referrer can be a page clicked from Google search
            for each trackingURL in trackingList
            {
                var sQuery = getResultTerm(trackingURL);
                if (searchQuery == sQuery)
                {
                    refURL = trackingURL;
                    break;
                }
            }
        }
    }
}
```

To identify the anchor text of a clicked link, the link needs to be compared with all the links in the referrer webpage. The problem is the links in the referrer webpages can be either absolute or relative while the clicked link is

always an absolute link as it is displayed in the address bar of the web browser. For example, the URL of a referrer webpage is:

```
http://www.w3schools.com/jsref/jsref_obj_array.asp
```

A relative link inserted somewhere in the page is:

```
<href="/jquery/default.asp" target="_top">JQUERY </a>
```

If users click on that link, the URL displayed in the address bar is:

```
http://www.w3schools.com/jquery/default.asp
```

In this case, the URL in the address bar is rather different to the relative link of the referrer webpage. Pseudo code 3.8 deals with this problem to extract the anchor text of a clicked link.

Pseudo code 3.8 Extracting anchor text of a clicked link.

```
function getAnchorText(curLink, vvrAllLinks)
{
    //Input: current URL and allLink of referrer page
    //Output: anchorText

    var runningLink;
    var snippet;

    for (var i=0; i < vvrAllLinks.length; i++)
    {
        runningLink = vvrAllLinks[i].href;
        if(runningLink == null)
            continue;
        //Normalise each link by deleting the last "/"
        if(runningLink.lastIndexOf("/") == runningLink.length -1)
            runningLink=runningLink.substring(0,runningLink.length-1);

        if(curLink.lastIndexOf("/") == curLink.length - 1)
            curLink=curLink.substring(0,curLink.length - 1);

        //Split each link into an array based one token "/"
        var runParts = runningLink.split("/");
        var curParts = curLink.split("/");
        //If two arrays are the same
        //--> found the clicked link in the referrer page
        if(runParts == curParts)
        {
            //Get the anchor text
            snippet= vvrAllLinks[i].text;
        }
        else if(runParts[runParts.length-1] == curParts[curParts.length-1])
        {
            //Relative link, so compare the last part of the wo links
            snippet= vvrAllLinks[i].text;
        }
    }
    return snippet;
}
```

3.2.2 The history editor form

The history editor form has been designed to allow users to browse in their history by date, view detailed information about each webpage, and delete it if necessary (see Figure 3.2). If users decide to delete a URL, the tool will search throughout their history and delete all entries which have the same URL.

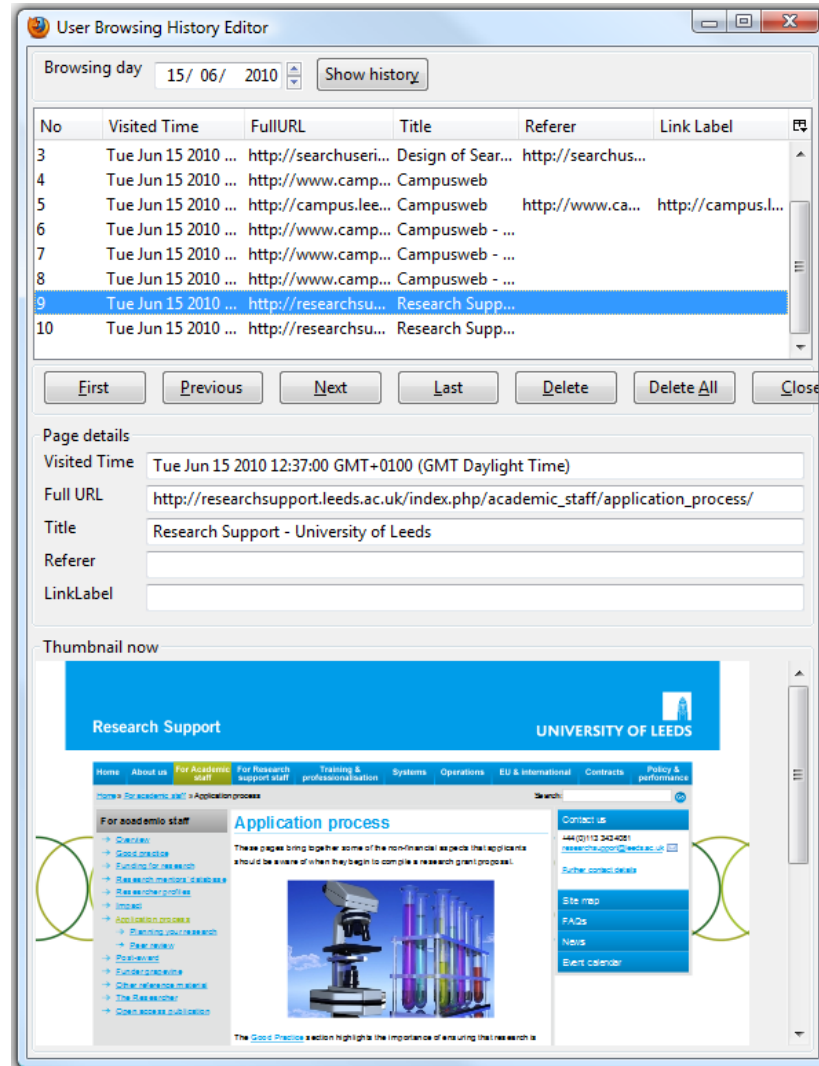


Figure 3.2 Browsing history editor form: Users can select a date to review their history on that date. Clicking on a list entry displays detailed information about that webpage. They can also delete unwanted entries.

3.2.3 The privacy form

A simple form is provided to let users protect certain websites or webpages from being recorded (see Figure 3.3). When viewing a webpage, users can open this dialog from the “Tool” menu, the URL of the webpage is automatically added to the “Never record this website” text box. They can also manually type the URL of a webpage here then decide whether to block

the whole website or only a specific webpage. The tool automatically extracts the domain of a URL if users click on the “Block site” button. These webpages/websites are maintained in a list and ignored by the logging process.

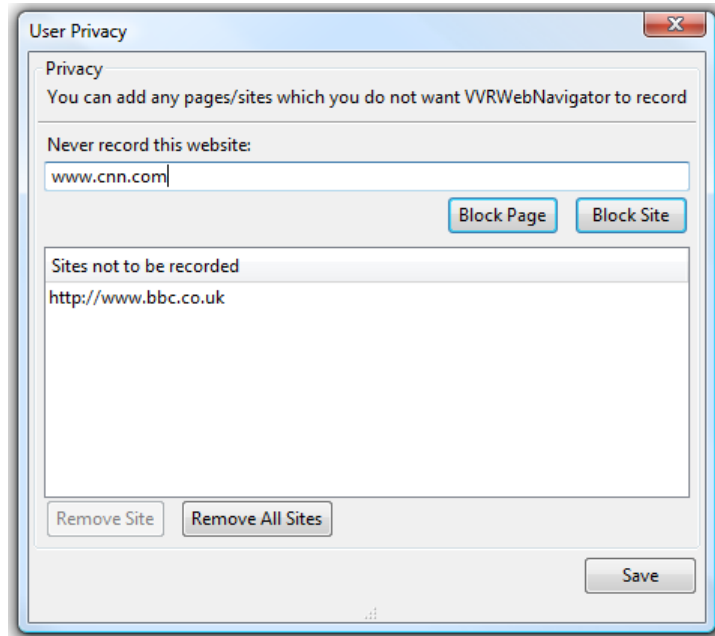


Figure 3.3 User privacy form: users can block a specific webpage or the whole website by putting a URL in the “Never record this website” button. The domain of a URL is extracted by the tool.

3.3 Summary

The logging tool has been developed as an add-on for Mozilla Firefox to automatically capture an individual’s web history, and has been tested and used on Windows, Mac, and Linux computers. The add-on stores webpage information (thumbnail, URL, title, description) and navigational information (visited time, URL and ID of referrer, anchor texts used to access the page, dwell time) in a database in the individual’s personal file space. The referrer ID allows navigational paths to be reconstructed later. When a webpage is loaded in the browser, a check is made to remove unwanted entries (e.g., ads, frames, and private pages). The tool captures the thumbnail of the whole webpage rather than just the visible area. Dwell time on each page is also tracked carefully, because it is counted only when both the browser and the page are active. Switching between tabs and applications is also taken into account. For privacy reasons, the add-on does not capture https webpages. It also allows users to specify that certain websites or pages should not be recorded, and to turn recording on/off by clicking a button that

was added to the Firefox browser window. The add-on also provides a history editor, which allows users to view and delete entries in their web history.

The logging tool also has limitations. If a user uses different computers, the history will be split. In the case where several users use the same account, captured data will be treated as of one user. Another problem is that as webpages are not saved into the database until they are closed, the history is not always up-to-date. A solution for a future version is to save webpages into the database as soon as they are loaded into the web browser and then update their dwell time later. The next chapter describes a user study which uses this logging tool.

Chapter 4. The underlying causes of revisit failure

This chapter describes an empirical study whose goals were to: (1) quantify how often people make occasional revisits to webpages (i.e., revisit neither frequently nor recently), (2) investigate cues that alleviate the difficulties that people encounter during occasional revisiting, and (3) understand the underlying causes of failures that occur when people try to revisit webpages on an occasional basis. The underlying scenario was a person wishing to find again a specific piece of information, either for their own purposes or in response to someone else's question.

The study started by capturing participants' web history for three months, followed by a controlled laboratory experiment during which participants were asked to revisit specific "target" webpages selected from this period. In other words this study, like (Teevan et al., 2004), asked participants to revisit pages they had previously visited during their own day-to-day web usage. By contrast, most previous research used web collections that have been chosen specifically for a given study (Hightower et al., 1998; Robertson et al., 1998; Wexelblat and Maes, 1999; Mayer and Bederson, 2001; Ceri et al., 2006).

Participants captured their browsing by installing and using a Firefox add-on logging tool described in Chapter 3. For the revisiting experiment, each participant was required to revisit 48 webpages but, to avoid fatigue, the experiment was performed in three 1-hour sessions that took place at weekly intervals. In each session a participant was asked to revisit 16 webpages that had previously been visited on only one day, either 7 ± 3 days previously (termed "1 week" in the remainder of this paper) or 28 ± 3 days previously (termed "1 month"). Previous research classified revisits of a week or more as very long-term (Mayer, 2009), this study wanted to investigate whether there were differences between the lower bound of long-term (1 week) and a greater interval (1 month). The ± 3 days was used to increase the number of pages that could be used as targets in the experiment (see Section 4.1.2.1).

Each page was described (see Section 4.1.2.2) and, for two pages out of each set of eight, a participant was provided with: (a) no supplementary cue, (b) the anchor text used to access the page, (c) a thumbnail image of the page, or (d) a page on the browsing path the participant had used when originally visiting the page. The choice of cues was informed by previous

research (Kaasten et al., 2002; Fujii, 2008; Li and Zhao, 2009; Dai and Davison, 2010; Koolen and Kamps, 2010). In summary, this experiment used a within-participants design with factors of recency (1 week vs. 1 month) and cue (none vs. anchor text vs. thumbnail vs. path).

4.1 Method

4.1.1 Participants

The study was approved by the University Research Ethics Committee. All the participants gave their informed consent and were paid an honorarium for their participation (pro rata if they withdrew). Twenty-three individuals (nine females) commenced the study but six of them withdrew and, in line with the University's ethical policies, did not need to give any reason. One participant was excluded from the study because she made little usage of the WWW, and four other participants could not finish all three revisiting sessions due to either an accident or illness.

The data reported in the following sections are from the 12 participants (6 females) who completed the whole study. All of the participants were students, two studying History, one Biology, seven Computer Science, and two Computing & Management. The participants' mean age was 26.2 years ($SD = 3.9$). They all had at least one year of experience with Firefox. Nine of them only used their laptop during the period of the study. The rest used two different computers (one at home and the other at university). In this case, the logging add-on was installed on the computer that participants used most often for accessing the WWW.

4.1.2 Revisiting experiment

All revisiting sessions were done on participants' own computers so they had access to their usual working environment (profile, etc.). However, before each revisiting session, they were asked to return the logfiles so the targets could be determined, and the page descriptions and cues could be generated.

4.1.2.1 Target page criteria

As one of the goals of this study was to identify which cues might be useful for revisiting pages that have been visited neither frequently nor recently, each target page had to meet the following criteria. First, it must have been previously visited on only one day. That meant a page could have been visited several times in that day but not in any other day in the whole period

of the study. Second, a page had to have been visited 1 week (7 ± 3 days) or 1 month (28 ± 3 days) before the revisiting session in which it was used. Third, the participant must have dwelled on that page for at least 30 seconds (webpages utility described in Section 2.2). In other words, the 30 seconds criterion increased the likelihood that target pages were of interest to a participant (this study terms these pages *informational*, and other pages as *navigational*) and would be memorable but, of course, the criterion could have been satisfied for other reasons (e.g., the participant got distracted). Fourth, the page must not be a search results page or a form – this criterion was used because participants were required to retrieve specific information. The first three criteria were implemented in a computer program, which identified possible target pages. The last criterion was checked manually because, once the other criteria had been used as filters, the volume of webpages that remained meant that manual checking was faster than writing software to automate the check.

Once potential target pages had been identified they were reviewed manually to determine which targets could be used in each cue condition. In each of the three experimental sessions there were 16 target pages (two for each combination of recency and cue), and no two selected target pages belonged to the same website. Only pages with meaningful anchor text (not “click here”, “read more”, “next”, etc.) were chosen as targets for the anchor text cue condition. The path cue was the URL of the page that was two clicks before the target page when the participant originally visited the page. Paths often were pages in the same website or a search result, and implicitly provided information about the locality of the target or what the participants were looking for when originally visiting it.

4.1.2.2 Target description and cue generation

The target descriptions were needed to simulate the scenario where the participant had a vague memory of information they wished to find again (“I remember something about X ... but where was it?”). The description was constructed from: (a) two pairs of two consecutive words chosen randomly from a target’s <title> tag (one pair from each half; if a title contained less than four words then all of them were used), and (b) two keywords extracted from the page’s content by the Alchemy web service¹⁵. The four pairs/keywords were then sorted randomly. However, these keywords were

¹⁵ See Transforming text into knowledge:
<http://www.alchemyapi.com/api/keyword/urls.html>.

then reviewed manually because (a) sometimes these keywords might not be appropriately extracted and (b) a participant often visited several webpages on the same topic, these keywords might not distinguish them well. This method attempted to make participants target to an individual page on a topic. Pilot testing showed that this method of describing the pages was sufficiently precise for participants to identify the target (subsequently, they were successful in 80% of the experiment's trials; see Section 4.2.2.1), without trivialising the revisiting task in the experimental setting.

Unlike previous studies (Robertson et al., 1998; Kaasten et al., 2002; Won et al., 2009), the thumbnail cues were generated from the whole of a webpage rather than just the amount that was visible at one time on a screen. The thumbnails were down-sampled to 140 pixels wide, with the height dictated by the page's aspect ratio, so a page's general appearance could be determined but the text could not be read (Kaasten et al., 2002).

4.1.2.3 Experiment procedure

At the start of the first session, participants read an information sheet that described the experiment's procedure and were encouraged to ask questions to clarify that procedure. All three sessions were videoed and logged using the add-on logging tool for subsequent analysis. The videos were used to record what participants said about target pages, including their memory for them and any difficulties that were expressed when revisiting. After that, participants could do anything to revisit the target pages. As previously explained, they used their own browser and computer, which helped the study to be ecologically valid.

In each session, a participant searched/browsed at their own pace until they had attempted to revisit all 16 targets, which were presented one at a time, in a randomly ordered list (see Figure 4.1). For each target page, the procedure was as follows. First, the participant read the description and, for anchor text, path and thumbnail targets, the additional cue information. Next, the participant was asked whether they recalled the page concerned, recalled the topic, how they previously found the page, what they were looking for on that occasion. After that, the participant "thought aloud" while they tried to find the target using any method they wished. Once the participant thought they had found the target they clicked on a button (see Figure 4.1) to open the target page to verify whether the revisit had been successful. If a target page was not found within 3 minutes the trial was terminated as unsuccessful.

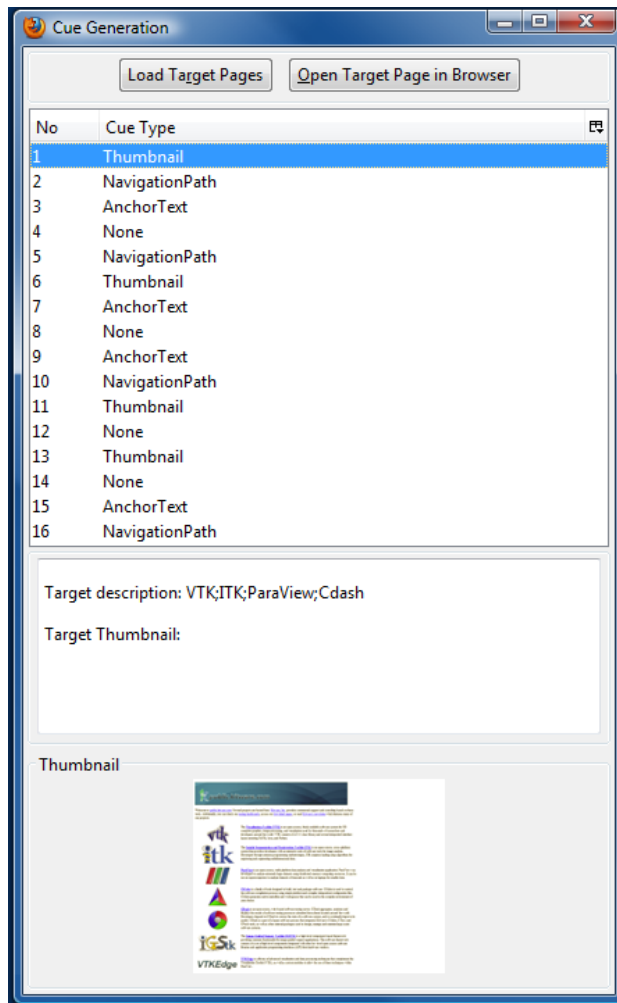


Figure 4.1 Target page dialog, showing one with a thumbnail cue.

4.2 Results

This section is divided into three sections. The first analyses participants' logfiles and makes comparisons with web navigation activities reported by previous studies. The second reports the results of the revisiting experiment sessions. The third combines the experiment and logfile data to determine the underlying causes when participants failed in their attempts to revisit webpages.

4.2.1 Logfile data

Overall, our participants' activity was broadly similar to that reported by another well-known study (Weinreich et al., 2006) (see Table 4.1). The recurrence rate is the percentage of page visits that were revisits, during the logfile recording period.

Table 4.1 Comparison of web navigation studies.

| | Weinreich et al | This study |
|---------------------------|------------------------------|------------------------|
| Year of study | 2004-2005 | 2010-2011 |
| Tools | Intermediary and Firefox 1.0 | Add-on for Firefox 3.x |
| Number of participants | 25 | 12 |
| Duration (days) | 52-195 | 50-97 |
| Recurrence rate | 46% | 36% |
| No. of URL visits per day | 90 | 86 |

Overall, 31% of page visits had dwell time of 30 seconds or more. The percentages that were informational vs. navigational for each combination of recency and frequency are shown in Table 4.2. Only for pages that were visited both frequently (on more than 1 day) and recently (within 3 days) did informational pages substantially outnumber navigational pages, indicative of participants going directly to the page they wished to revisit (e.g., using browser auto-completion functionality). Almost one fifth of the revisited pages were in the neither recent nor frequent category, which is the focus of this research.

Table 4.2 Percentage of revisited pages that were informational vs. navigational, subdivided according to recency and frequency.

| Page Type | Frequent | | Not Frequent | |
|------------------|-----------------|-------------------|---------------------|-------------------|
| | Recent | Not Recent | Recent | Not Recent |
| Informational | 9.5% | 3.4% | 29.7% | 9.9% |
| Navigational | 3.8% | 2.3% | 33.1% | 8.3% |
| Total | 13.3% | 5.7% | 62.8% | 18.2% |

As in another previous study (Cockburn and McKenzie, 2001), each participant had several websites ($M=4.8$, $SD=2.9$) that they visited often (on at least 50% of the days of the study period). Those websites were often online shopping websites, Google search, Facebook, Youtube, online TV, online radio, online music, news, forums, blogs, dictionaries, and organizational websites. Across all of the participants, the most visited websites were Google (14% of all visits), Facebook (13%), BBC (6%), our University website (5%), YouTube (4%), Wikipedia (3%) and eBay (2%). All participants' main search engine was Google.

Participants' activity was divided into sessions using the timeout method (a period of user inactivity) with a criterion of 25.5 minutes (Catledge and Pitkow, 1995). This showed that participants carried out an average of 3.9 sessions per day ($SD = 0.8$) and the average session length was 31.9

minutes ($SD = 11.0$), which are figures comparable with previous research (e.g., 3.2 sessions/day and 24 minutes (Mayer, 2007)).

On average, the length of navigational paths was 2.1 pages. However, 5% of paths had a length of over four pages, and there were a few cases whose lengths were over 20 pages.

4.2.2 Revisiting experiment results

The data in this section were analysed from logfiles and videos of the revisiting sessions. First, an overview of the successful and unsuccessful trials is presented. Then an analysis of variance (ANOVA) is used to investigate the effect of recency and cue, and participants' memory of the target pages is discussed. Finally, participants' revisiting strategies are described.

4.2.2.1 Success and failure

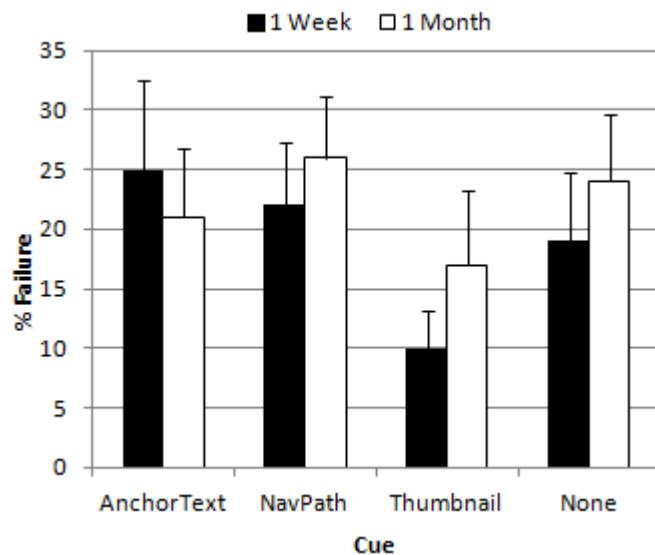


Figure 4.2 Percentage of unsuccessful revisits for each recency/cue combination. Error bars show standard error of the mean.

The percentage of trials in which participants failed to revisit the correct target page was analysed using a repeated measures analysis of variance that had two factors (recency \times cue). Overall participants failed to revisit the target in 20% of trials (see Figure 4.2), but there was not a significant difference in the percentage of failures after one week vs. one month ($F(1, 11) = 0.37, p = .55$) or between the four cues ($F(3, 33) = 2.45, p = .08$), and there was not a significant interaction ($F(3, 33) = 0.38, p = .77$).

Before revisiting each target, participants were asked if they recalled the specific target page, its general topic, or not at all. Participants stated that they specifically recalled the majority (79%) of the targets, but still failed to

revisit 13% of those targets because participants either could not find them again or turned out to have recalled the wrong page. In 16% of trials participants remembered the general topic, but there was a 46% likelihood of failing to revisit it. In 5% of trials did participants not remember anything about a target, but the failure rate was only 57% because sometimes participants were still able to find it by searching again or searching in their history (see Section 4.2.2.2) and recognising the page when they reached it.

4.2.2.2 Revisiting strategies

Each participant tended to adopt a single strategy (search again, or search in history) for most trials, and other strategies on an occasional basis (see Table 4.3). The characteristics of each strategy were:

- *Search again*: search from scratch by typing search queries into search boxes of search engines.
- *Search in history*: some words typed into the address bar or history dialog to see suggestions from the browser history and bookmarks
- *Browse*: direct entry of a URL and then browsing.
- *Bookmarks*: a target page is thought to be in the bookmark and participants select one from there.
- *Mixture*: two or more of above methods are used.

Table 4.3 Comparison of performance indicators across different revisiting strategies.

| Strategy | Percentage of trials | Failure rate for strategy | Proportion of the failures that were | | Mean No. of pages visited per target |
|-------------------|----------------------|---------------------------|--------------------------------------|---------|--------------------------------------|
| | | | 1 week | 1 month | |
| Search again | 50% | 23% | 45% | 55% | 3.7 |
| Search in history | 29% | 7% | 64% | 36% | 1.5 |
| Browse | 15% | 26% | 32% | 68% | 3.8 |
| Bookmark | 1% | 67% | 100% | 0% | 3.7 |
| Mixture | 4% | 38% | 80% | 20% | 5 |
| Did not try | 1% | 100% | 25% | 75% | - |

Using the search again strategy, participants typed search queries into a search engine and used the colour of hyperlinks (the “already visited” link colour) to decide whether or not to follow a given search result. Participants often searched using the keywords provided in the target’s description, and occasionally used the anchor text that was provided. To find the target, participants typically had to browse further webpages after clicking on a search result, try several results and/or refine the search queries (participants performed an average of 1.7 searches/target).

With the search in history strategy, participants just typed words or part of a URL on the address bar of Firefox, even if only the topic was remembered, and the auto-completion function included possible matching webpages from their browsing history and bookmarks in a drop down list. This feature was released in Firefox 3.0, since when the location bar has been called the Awesome Bar, and is also provided by recent versions of Google Chrome 12 and Internet Explorer 9. Despite this, in half of the trials participants had to try two or more sets of search queries before obtaining a suitable set of results. Overall, 51% of searches produced three or fewer suggested pages, making it easy for participants to revisit the target. However, 46% of searches produced more than three suggested pages, making it more difficult for participants to determine which one was correct, and they sometimes tried all the suggested pages and browsed around. No suggested pages were returned in 3% of searches, which caused participants to either look in their history list or search from scratch with the search again strategy (see above).

Participants sometimes adopted the browsing strategy if they recognised that the target page was in a particular website (e.g., shopping, travelling, news or forums), going to that website and then browsing or searching locally. Occasionally, participants initiated a revisit by browsing from the navigational path page that was provided as a cue. Bookmarks were rarely used.

4.2.3 The underlying causes of failure

Each unsuccessful trial was analysed in detail, combining data about participants' navigational actions during the experiment, video/audio of participants' thinking aloud and related data from participants' logfiles, to establish the underlying causes for the failures. Ten causes were identified, and occurrences of these are shown in Table 4.4.

Table 4.4 Comparison of failure rates due to different causes.

| Cause | No. of failures | % of failures | Proportion of the failures that were | |
|---|-----------------|---------------|--------------------------------------|---------|
| | | | 1 week | 1 month |
| Topic | 29 | 25% | 55% | 45% |
| Search results | 28 | 24% | 54% | 46% |
| Known website | 14 | 12% | 21% | 79% |
| Deleted link | 8 | 6% | 0% | 100% |
| Hidden information | 7 | 6% | 14% | 86% |
| Search on specific website | 8 | 6% | 75% | 25% |
| Inappropriate page title | 6 | 5% | 83% | 17% |
| Links from email, forum & social networks | 3 | 3% | 33% | 67% |
| Multi-page thread | 2 | 2% | 100% | 0% |
| Do not remember | 13 | 11% | 46% | 54% |

Topic: At different points of time, each participant might have several interests and be performing a variety of tasks, e.g., following a sport event which might last from several days to several weeks, or doing research for an assignment or dissertation. During those periods, they may visit many pages from a variety of websites. Therefore when asked to revisit one of those pages, participants often could not figure out the correct page. For example, participants typically said something like this: “I’ve visited a few webpages about XXX when I did YYY, I’m not sure which page contains the specific content mentioned in the description. It might be page A”. Next, they tried to revisit page A and could not find that specific information. Then they thought “Oh It might be page B”. After revisiting several candidates they gave up.

To investigate this type of failure, the webpages that were visited in each “topic” failure of the experiment were checked against a participant’s logfile data, to determine how close they came to the target page before failing (see Table 4.5). *On navigational path* was when participants visited a page(s) that were linked directly to the target, but not the target itself. *Same session* was when participants visited a page(s) that had previously been visited during the same session as the target (sessions were defined by the 25.5 minutes timeout; see Chapter 2). *Same day, within a week* and *within a month* were when participants visited a page(s) that had been visited within that time-lapse of the target. *Gave up* was when participants made no serious attempt, and *all pages were new* was when participants did not visit any previously visited pages in the trial.

Table 4.5 Summarisation of failures due to the *Topic* cause.

| Closeness | Number of Failures | Proportion of the failures that were 1 | |
|--------------------|--------------------|--|---------|
| | | 1 week | 1 month |
| On navigation path | 6 | 50% | 50% |
| Same session | 6 | 50% | 50% |
| Same day | 2 | 50% | 50% |
| Within a week | 3 | 33% | 67% |
| Within a month | 3 | 67% | 33% |
| Gave up | 4 | 100% | 0% |
| All pages were new | 5 | 40% | 60% |

Search results: Participants sometimes clicked on several results then browsed deeper before reaching the information they were looking for. Failure could occur at any one of four stages during the revisiting process (see Table 4.6): changes in search results, wrong search query, not recognising the correct result, or not browsing sufficiently from the correct result. Of the 28 failures with this cause, one occurred when a participant remembered the correct search query but the target page was not now returned in the results. For eight of the failures, participants did not correctly recall the same search query they had used when originally visiting the target, so the relevant search result was not listed. In four of those failures the logfile data showed that the participant had refined a search query before visiting the target, and in seven of the failures the participant had clicked on two or more results from the search results pages. Ten of the failures occurred when the correct result was listed, but participants did not recognise it despite links being coloured to indicate that the page had previously been visited. The logfiles revealed that the participants clicked on a few results in several results pages during the original search session. Of those 10, five occurred when participants correctly recalled the search query they had used to originally visit the target, and five occurred when participants used a similar search query. Nine of the failures occurred when participants clicked on the correct search result, but did not manage to browse to the target page from there.

Table 4.6 Summarisation of failures due to the *Search results cause*.

| Stage of Failure | Target was originally visited | | Proportion of the failures that were | |
|----------------------------------|-------------------------------|------------------------|--------------------------------------|---------|
| | As a search result | Browsing from a result | 1 week | 1 month |
| Results changed | 1 | 0 | 0% | 100% |
| Wrong search query | 7 | 1 | 50% | 50% |
| Did not recognise correct result | 5 | 5 | 40% | 60% |
| Not browsing sufficiently | - | 9 | 78% | 22% |

Known website: This cause of failure was when participants correctly recalled that a target page belonged to a particular website, but could not find the page when they browsed/searched that website again. These cases often happened when participants used the browsing strategy on a large website (e.g., a university), and also occurred when they originally visited the regional website (e.g., www.nintendo.co.uk) of a company but tried to revisit the information on the company’s global website (e.g., www.nintendo.com).

Table 4.7 Number of pages visited before and/or during the revisiting experiment and recency for each *Known website* failure.

| Case | Number of pages visited | | | Recency |
|------|-------------------------|----------------------------|------------------------|---------|
| | Only before experiment | Before & during experiment | Only during experiment | |
| 1 | 1 | 1 | 2 | 1 Month |
| 2 | 2 | 1 | 3 | 1 Month |
| 3 | 2 | 1 | 0 | 1 Month |
| 4 | 2 | 2 | 1 | 1 Month |
| 5 | 2 | 2 | 3 | 1 Month |
| 6 | 8 | 1 | 2 | 1 Week |
| 7 | 9 | 2 | 2 | 1 Month |
| 8 | 1 | 5 | 1 | 1 Month |
| 9 | 9 | 3 | 1 | 1 Week |
| 10 | 28 | 3 | 2 | 1 Month |
| 11 | 31 | 3 | 3 | 1 Month |
| 12 | 60 | 1 | 2 | 1 Month |
| 13 | 46 | 10 | 6 | 1 Week |
| 14 | 67 | 4 | 4 | 1 Month |

Table 4.7 summarises a number of webpages visited before and/or during a revisiting experiment for each “known website” that failure occurred. Two characteristics that were typical of the failures were that participants: (a) only

looked at a few pages on the site before thinking that they could not find target pages and giving up, and (b) looked at as many new pages as pages that they had visited before. In 71% of cases, this type of failure occurred when participants used a browsing strategy (see Section 4.2.2.2).

Deleted link: Target pages were previously visited via a link that had subsequently been deleted (e.g., from the homepage of a news website such as the BBC, which is frequently updated). Although participants remembered where they had previously found a target page, to find it again they either had to classify that page to go to the appropriate archive of a website to browse or form new search queries to search for the target. Both approaches have difficulties: classifying is not always correct and searching again might produce a large number of similar results. The most common explanation of the participants was: “*I just opened the homepage of website XXX and clicked on a link there but now I can’t see it anymore*”. Sometimes they just gave up and said: “*I don’t think I can find it again.*”

Search on a specific website: This often happened with websites providing local search function and/or filters, for example, accommodation rental, recruitment, and shopping websites. Repeating the actions that were previously performed to find a specific webpage is not easy. Besides, information on these websites is updated very often.

Hidden information: Some webpages only initially showed some information and users needed to click links to view details. In revisiting trials, participants could go to the correct page but forgot how to reach detailed information.

Inappropriate page title: This was one of the main reasons that made the search in history strategy unsuccessful, because this method relies on words contained in a title or URL. This problem is generally caused by bad title assignment.

Links from email & social networks: Today, links sent and shared by emails, forums and social networks are very popular. If the emails are deleted, or the posts are not available anymore, people will find revisiting difficult.

Multi-page thread: Forums are widely used nowadays to discuss ideas, share knowledge, etc. Each forum is organised in boxes. Members contribute posts in threads of boxes. In some “hot” threads, posts spread over many pages. The problem is all pages of the same thread have same titles, so it is difficult for users to find a specific post irrespective of the strategy they use.

Do not remember: This was when people did not remember anything about target pages. Sometimes, participants expressed their frustration of knowing a target webpage was somewhere but could not revisit it.

4.2.4 Pattern of causes of failure for revisiting strategies

After underlying causes of failure had been identified, all unsuccessful trials were reviewed to see whether there was any pattern of causes of failure for each revisiting strategy. Table 4.8 shows the number of failures of each cause for revisiting strategies.

Table 4.8 The number of failures of each cause for revisiting strategies

| Strategies Causes | Search again | Search in history | Browse | Bookmark | Mixture | Did not try |
|---|-----------------|----------------------|----------|----------|----------|----------------|
| Topic | 22 | 2 | 2 | 0 | 3 | 0 |
| Search results | 19 | 2 | 3 | 1 | 3 | 0 |
| Known website | 3 | 0 | 9 | 1 | 1 | 0 |
| Deleted link | 5 | 0 | 1 | 0 | 0 | 2 |
| Hidden information | 5 | 0 | 2 | 0 | 0 | 0 |
| Search on specific website | 3 | 0 | 3 | 0 | 2 | 0 |
| Inappropriate page title | 0 | 5 | 0 | 0 | 1 | 0 |
| Links from email, forum & social networks | 0 | 2 | 1 | 0 | 0 | 0 |
| Multi-page thread | 2 | 0 | 0 | 0 | 0 | 0 |
| Do not remember | 6 | 0 | 1 | 0 | 0 | 6 |

The numbers in bold of Table 4.8 emphasise the main cause(s) of failure of each revisiting strategy. There were some clear patterns. More than 63% of unsuccessful trials revisited by the “Search again” method belonged to either the “Topic” cause or the “Search results” cause. It was understandable when the main cause of failure for the “Search in history” strategy was “Inappropriate page title” because this method relies on a webpage title. Similarly, participants did not try to revisit target pages because they mainly did not remember them. Another pattern was that participants often did not browse effectively when target pages belonged to a known but complex website.

4.3 Discussion

In contrast to many previous studies of revisit (Hightower et al., 1998; Robertson et al., 1998; Wexelblat and Maes, 1999; Mayer and Bederson, 2001; Ceri et al., 2006), this study asked participants to revisit pages they had previously visited during their own day-to-day web usage. However, any user study has limitations, and the main ones that concerned this study were the method used to select/describe the target pages, the amount of time participants were given to revisit each target, the fact that logfiles were only recorded on participants' main computer, and the number and background of the participants who were used. The fact that participants successfully revisited 80% of the targets, coupled with the patterns that were found in the underlying causes of failure, indicates that the method used to select/describe the target pages was broadly appropriate. In terms of the amount of time that was allowed, participants either thought they had found a target page within the three minute limit or decided to give up. Only recording logfiles on participants' main computer means that it is possible that targets may have been visited more frequently or recently than assumed. However, this widens the implications of our findings, rather than limiting them. With only 12 participants being used, there are clearly limits to the extent that the results can be generalised to web users as a whole. However, the number of participants is similar to other research of the same type (Jones et al., 2001; Mayer and Bederson, 2001; Teevan et al., 2004; Won et al., 2009). The rather uniform background of the participants, if anything, strengthens the findings because they demonstrate that even with young, well-educated and familiar WWW users, they have difficulties in revisiting.

There was not a significant effect of failure rate between the cues although the slightly lower rate for thumbnail cue was consistent with the predicted advantage it provided in terms of recognition (Kaasten et al., 2002). The lack of a significant difference is due partly to the revisiting strategies that participants adopted in 79% of trials: using search queries to search for the target either again from scratch or in their history (see Table 4.3).

Thumbnails helped participants to recognise targets, but they still had to navigate to them. The words contained in the anchor text were often also part of a page's description (indeed, this should be the case if anchor text provides "scent" for a page's content), and navigational paths placed participants in the vicinity of a target, but finding it remained difficult if the site

was complex or the participant had previously visited many different pages. These reasons were analysed in Section 4.2.3.

Overall, the failure rate of 20% did not differ between pages visited on one week vs. one month previously, which indicates that even a time delay of a week is sufficient for researchers to study revisiting phenomena, which will be useful when tools designed to assist revisiting are to be evaluated. Data in Tables 4, 5, 6, and 7 show that in most of the cases there was a fairly even split between unsuccessful trials of 1 week and 1 month. However, in some circumstances, there were sizeable differences. As shown in Table 4.3, it is likely that it was more difficult for participants to browse to webpages they visited a month ago than a week ago (the “Browse” strategy in the table) or participants gave up more with webpages visited a month ago (the “Did not try” strategy in the table). The results in Table 4.4 are generally consistent with those of Table 4.3. Most of the unsuccessful trials of the “Known website” cause belonged to the 1 month recency (accounting for 79%) and participants often adopted the “Browse” strategy for the trials that fell into this cause (see Table 4.8). With the “Search in history” method, the proportion of failures of 1 week trials was higher than that of 1 month trials. This could be explained by the fact that the main cause of failure of this strategy was “Inappropriate page title” (see Section 4.2.4). As webpages of the “Deleted link” cause belonged to websites which were frequently updated (see Section 4.2.3), it was understandable when all the failures were visited a month before. Similarly, it was more difficult for participants to re-access “Hidden information” after a month (accounting for 86%). Other cases had too few failures to reveal reliable patterns.

It would also be worthwhile educating people about existing browser functionality for searching with its history, because surprisingly few participants knew about that functionality, even ones with computing backgrounds, and a search in history strategy proved to be quicker and more successful than search again. Broadly speaking, the revisit methods observed in the experiment sessions were similar to a previous study (Bruce et al., 2004), albeit with participants in this study relying more on searching again and searching in history, rather than direct entry of URLs or Bookmarks. This can be explained by the difference of the targeted category general revisit. Browsing was still one of the most common methods, with participants exploiting an “orienting” strategy (Teevan et al., 2004). It is understandable that some failures were caused by incorrect queries,

because 20% of search-based re-finding uses queries that are different to those originally used to find a webpage (Tyler and Teevan, 2010).

Given that the logfile data showed that only 18% of revisits were to pages that participants had previously visited neither frequently nor recently, and participants successfully found 80% of the pages they were asked to revisit in the experiment, the practical significance of this research might be questioned. However, although the overall revisiting failure rate (3.6%) is small, the frustration that our participants expressed has also been noted in previous studies (Bruce et al., 2004; Teevan, 2007b). This study is the first to attempt to dig deeper into these failures to investigate their underlying causes. These problems are hypothesised to be largely alleviated if participants had been able to interactively explore their web history, and filter it to zero in "orienteering" (Teevan et al., 2004) to the webpage they wish to revisit.

4.4 Summary

Revisiting webpages is difficult when they have been visited neither frequently nor recently. This chapter describes an empirical study into this type of webpage to investigate the difficulties that people encounter during occasional revisiting. The participants' logfiles in this study revealed that almost one fifth of the revisited pages were in this group, and the failure rate of 20% when revisiting them did not differ between pages visited one week vs. one month previously. Ten causes of failure were identified by analysing unsuccessful revisiting trials of a controlled laboratory experiment, data about participants' navigational actions during the experiment, video/audio of participants' thinking aloud and related data from participants' logfiles. The three main causes (accounting for 61% of the failures) were: (1) participants visiting a large number of pages on a particular topic, (2) webpages that had originally been accessed via search results, (3) participants knowing which website contained a page but that site itself being large. These causes of failure need to be taken into account when designing future web browser functionality and web history tools. The next chapter describes our own such a web history tool.

Chapter 5. The new visualization history tool

As highlighted in Chapter 2, visualization presentation of the web history tools was proved to be superior to conventional list-based presentation in supporting webpage revisiting. The findings from the user study described in Chapter 4 showed: (1) the limitations of current history support (20% trials failed when revisiting non-frequently and non-recently visited webpages) and (2) the increased effectiveness of revisiting using a web history compared with searching again on the WWW as a whole. This chapter proposes a new visualization history tool that addresses some of these limitations. First, the requirements of the tool are identified. Following this, the design of the tool is described. As mentioned in Chapter 1, a user centred design was adopted to develop the new history tool. Therefore, the design of the history tool is presented through several iterations. Finally, the technical implementation of the tool is described.

5.1 Requirements

The goal of the new visualization history tool is to address the five main causes of failure: 1) Topic, 2) Search results, 3) Known website, 4) Deleted link, and 5) Links from email & social networks. These causes account for almost 70% of failures in the user study. Consistent with established processes for interaction design (Preece et al., 2002), the requirements of the tool are divided into four aspects: functional, data, environmental, and usability.

5.1.1 Functional requirements

The common problem in all the five causes of failure that are mentioned above is that there are often a few candidate webpages, but users cannot decide which webpage they want to revisit until they actually see it.

Therefore the tool should provide users with three key functions:

- 1) The tool will allow users to navigate their history to select a small set of possible webpages from the whole history.
- 2) The tool will then present the selected set of webpages in a visual way so users can recognise and choose the right page.
- 3) In case users still find it difficult to find the target page, the tool should support filters or other ways of navigating the selected set.

5.1.2 Data requirements

To provide the above functions, the tool must have access to data captured by the logging tool (see Chapter 3). These data include information about webpages such as URL, title, description, thumbnail, frequency, referrer, visited time, dwell time.

5.1.3 Environmental requirements

The data stored by the tool should only be accessible by a given user because it displays personal information. Users should be able to install and run the tool on any computer (e.g., Windows, Mac, and Linux) without any other technical support, libraries, frameworks or configurations. Last but not least, the tool should be seamlessly integrated into web browsers so users do not need to remember to start it separately when browsing the WWW.

5.1.4 Usability requirements

To be used by a wide range of WWW users, the tool needs to be simple and intuitive so that new users can use it immediately with minimal training. As suggested in Chapter 4, informational webpages are likely to be revisited more often than navigational webpages, so the tool should emphasise them. To allow users to navigate in their history, the tool should provide navigation information such as letting users know where they are, where they have been (e.g., so they could go back or forward), and where they can go from where they are. To support long-term revisiting, the tool needs to retrieve and present data interactively, to ensure a positive user experience. Last but not least, the tool should have minimal effect on the performance of the web browser.

5.2 The first iteration: paper-based prototype

In the first iteration, a paper-based prototype was used to get quick feedback from potential users. The main goal of this iteration was to establish key components and functionality for the future history tool.

5.2.1 Design

The prototype was rapidly designed with Microsoft PowerPoint. Some familiar components of existing tools such as the Google Web History's Calendar and the Firefox's History were used to illustrate initial design ideas. Figure 5.1 shows the layout of the future visualization history tool and describes its main functionality.

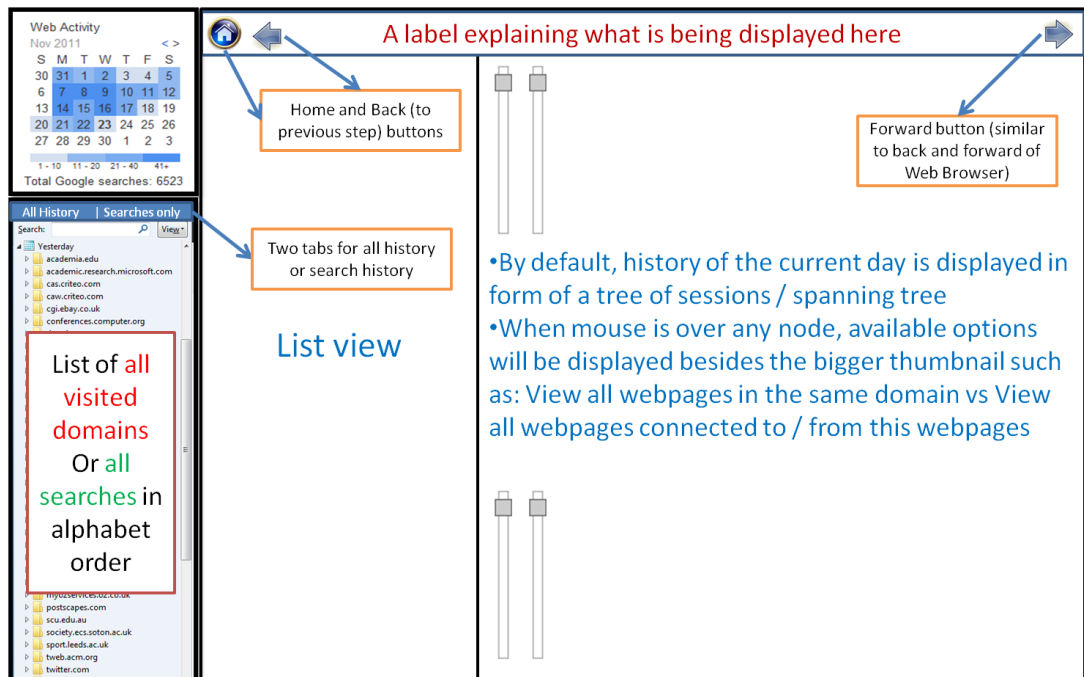


Figure 5.1 The paper-based prototype of the visualization history tool includes: the Global Navigation at the left with a heat map calendar and a tab view; the Result View at the right with a list view and a tree view; and the Toolbar at the top containing buttons.

Having a similar layout to Microsoft Outlook, which is used by millions of people, the tool has three main components:

- The *Global Navigation* at the left side of the window includes a calendar, a search box and a tab view. This is the main component which allows users to navigate within their web history based on how they remember target pages and how they start revisiting. The *Global Navigation* is designed to meet the first functional requirement.
- The *Result View* at the right side of the window displays results of every navigation in both a list view and a tree view. When the mouse is over any node in the tree, the corresponding item in the list is highlighted by a blue bar. The *Result View* addresses the second functional requirement.

- The *Toolbar* at the top of the window with buttons allows users to perform actions such as going back to the default state (home), going back/forward.

5.2.1.1 Global navigation

This is the main area for users to navigate within their history. Elements of the *Global Navigation* are designed mostly based on the ways participants started their navigation when revisiting webpages. Two components are provided.

A *month view calendar* enables users to navigate by date in their web history (top left of Figure 5.1). When a date is selected, it is highlighted in red while others are in black and the web history on that day is displayed in the *Result View*. The calendar is enhanced by a heat map which uses different background colours for different days. The heat represents the number of either webpage visits or search queries on each day depending on which tab is active (see the *Tab View* below). Both colour hue and different shades of a colour can be used to encode the heat. However, here the same information with different frequency needs representing so different shades of blue are chosen. Only five different shades of blue (including white) are selected so users can remember and distinguish ranges (e.g., [1 - 25], [26 - 49], [50 - 75], and [76 - ∞] for webpage visits, and [1 - 5], [6 - 10], [11 - 15], and [16 - ∞] for search queries). The heat map provides a good overview of a user's history in each month. When the mouse is over any date, a label displays the number of visits/search queries on that day.

A *tab view* with two tabs (*All History and Searches Only*) provides users with other ways of navigating. By default, the *All History* tab is a list of all domains visited in the history. If a date is selected on the calendar, only domains visited on that date are listed. When users know the domain of a target webpage, selecting it from the list shows all visited webpages of that domain in the history (or on a selected date) in the *Result View*. As the list becomes long over time, a filter box is provided. Users can type some characters there to filter out unmatched domains. Users can always go back to the full list of domains by clicking "All Domains" at the top of the list. Similar to the *Domains* tab, the *Searches Only* tab allows users to select a search query launched in Google search to view its search trails.

5.2.1.2 Result view




The main purpose of the *Result View* is to display a filtered set of a user's history, from which they can recognise the page they wish to revisit. By

default, the *Result View* displays the web history of the current date. A list view can display more information in a given amount of display real estate than tree view, but a tree view is better at showing the relationship between webpages. That's why this design employs both types of view. In some navigation, a selected set of a user's history still has many webpages, filters are required. Four sliders are added to the tree view to enable users to filter webpages in the *Result View* by dwell time, number of visits (frequency), recency, and number of days visited (e.g., a webpage might have been visited several times but only on one or two days). These sliders partly support the third functional requirement. When a filter is applied, unqualified entries are removed from the list view. In the meanwhile, if an unqualified node is deleted from the tree view, its navigational path is distorted. The solution for this is to reduce the size of such a node (see Figure 5.9).

Users can zoom in/out the tree view with a mouse wheel and pan by dragging and dropping. If the cursor is over a node in the tree then: (1) the list view automatically scrolls to the corresponding entry and highlights it with a blue bar, so that users can see both the detailed information about a webpage and its relation to others; and (2) a dialog is displayed to provide some other navigation options from the node such as:

- View all webpages visited in the same domain (similar to when users select a domain in the domain tab).
- View all webpages visited from this webpage. This function would be useful when links of a certain webpage are updated regularly. For example, users go to the BBC homepage one day and click some links there to read the latest news. The next day, those links might be no longer there.
- View all external webpages visited from this domain. Today, links sent and shared by emails, forums and social networks are very popular. If a user remembers the wanted webpage was shared from a certain domain, this feature would help.
- View all webpages visited on the same day/session. This option allows users to explore all webpages that have been visited at the same time with the current webpage.
- View all webpages visited on the same topic. A keyword of the webpage will be extracted and of course users can refine this keyword. All webpages whose title or description contain that keyword will be displayed.

5.2.1.3 The toolbar

The *Toolbar* also offers extra functionalities via some buttons. To give an overview of a filtered set of a user's history, initially the tree view is always fitted in the tree view area. The "Home"  button resets all components of the visualization history tool to the states when the tool is opened. Similar to the "back" and "forward" buttons of web browsers, two buttons ( and ) in the *Toolbar* let users go back and forward navigation actions.

5.2.2 Initial feedback from users

The paper-based prototype and the description in Section 5.2.1 were shown to three potential users individually (an academic, a researcher, and a PhD student). The academic concerned about the way the sliders for filtering features were add to the tree view. As there were four sliders, labels would be required. The sliders with their labels might occlude the tree view. He also suggested that the tool should let users switch between the list view only and both the list view and the tree view.


The researcher thought displaying a dialog whenever the cursor is over a node was not practical because it would occlude the tree view and slow down users' navigation. The PhD student suggested renaming the tabs of the tab view to reflect what they really do and wondered if the tool could split the web history of a day into sessions, e.g. morning and afternoon.

5.3 The second iteration: the visualization design details

The second iteration took into account the feedback of users from the previous iteration to refine the key components and functionality of the history tool. It also focused on the detailed design of the *Result View* of the history tool. After that, the history tool was implemented as a Firefox add-on. The first version of the history tool was then presented and demonstrated to the author's research group. Again, the three previous users agreed to try the history tool on their computer for one month while the author prepared for a formal user evaluation. The aims of this try were: (1) to test the tool in terms of installation, compatibility, functionality, and usability; and (2) to see if any new functions are required when users actually use the history tool.

5.3.1 Addressing the feedback from the previous iteration

To address the feedback from the previous iteration, the following changes were made: (1) the three sliders were moved to the top of the tree view (see Figure 5.2); the slider for filtering by recency was eliminated because users

could select to view the visited webpages of any day on the month view calendar; (2) a new button  was added to the Toolbar to enable users to switch between the list view only and both the list view and the tree view (see Figure 5.2); (3) the dialog providing other navigation options will be displayed when users right click on a webpage; (4) the tabs of the tab view were renamed to “Domains” and “Google searches” (because currently the tool only supports catching search queries submitted to Google search engine); and (5) a new tab, called Sessions, was added to divide each result set into sessions based on the 25.5 minute pause of browsing (Catledge and Pitkow, 1995).

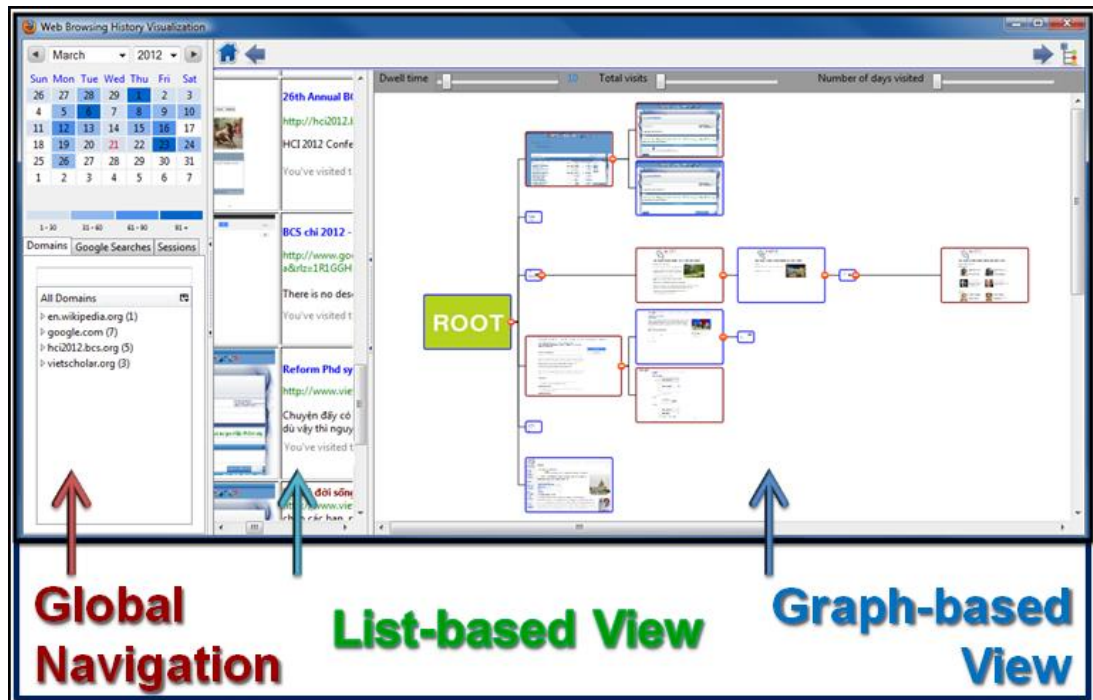


Figure 5.2 The visualization design details of the history tool.

5.3.2 Detailed design of the *Result View*

As described in Section 5.2.1.1, the *Result View* displays a filtered set of a user’s history in both a list view and a tree view.

5.3.2.1 The list view

Taking advantage of ideas used in both WebNet (Cockburn and Jones, 1996) and SessionGraph (Mayer and Bederson, 2001), a full list view with detailed information about each page is employed to complement the tree view. As Google search has become so familiar with WWW users, the design adopts the style of Google search results to present a webpage’s title, URL, description, frequency (“You’ve visited this webpage X times”) and recency (“Last visited ...”). The basic listing is enriched by adding a small

thumbnail, using the same approach as CWH (Won et al., 2009). This conveys the layout of a webpage, which proved to be useful for users' recognition (Kaasten et al., 2002; Won et al., 2009), and is becoming increasingly common with search engines such as Google. Considering the trade-off between the amount of recognisable detail vs. thumbnail size, which dictates how many thumbnails can be displayed at the same time, a thumbnail height of 148 pixels has been chosen (Kaasten et al., 2002; Won et al., 2009). Today different monitors have different aspect ratios and webpages are often displayed in this ratio. Therefore, rather than a fixed aspect ratio for all monitors, the width of a thumbnail is calculated by the formula (1) below.

$$\text{Thumbnail width} = \frac{\text{Screen width}}{\text{Screen height}} * \text{Thumbnail height (1)}$$

To distinguish between pages of interest (dwell time \geq 30 seconds) and other pages, maroon and blue are used for their titles respectively. Although the width of the list view is user-adjustable, horizontal and vertical scrollbars are provided so users can browse through a long list easily. The example of a list view entry is shown in Figure 5.3.

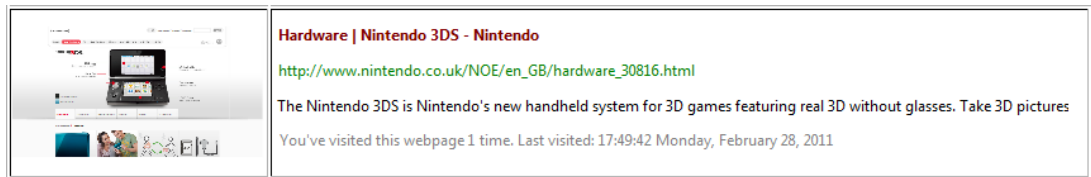


Figure 5.3 An example of the list view entries, with a bold maroon title for a webpage of interest (dwell time \geq 30 seconds) and normal title for another webpage.

In a web history, a page might have been visited several times. Displaying it several times with associated webpages might preserve contextual information. On the other hand that might confuse users too as they may think they are scrolling back to the same place of the list view. Our design displays a page only once in the list view, even if it has been visited several times. This makes the presentation consistent and compact.

In the list, webpages are ordered by when they were first visited, which provides users with contextual information by grouping webpages that were first visited together. It would also be possible to order the list by another attribute, for example, dwell time, frequency or recency.

5.3.2.2 The tree view

A tree view is used because it reflects the manner of user navigation on the WWW. A tree is the type of graph that is most familiar to people in general,

have been used in Webmap (Dömel, 1995), Domain Tree Browser (Gandhi et al., 2000), and SessionGraphs (Mayer and Bederson, 2001), and is fast to compute. The tree view in this design is built by reconstructing a user's actual navigational paths in a tabbed browser rather than based on the visited time of webpages like previous studies (*ibid*). The tree is presented in a horizontal orthogonal layout (see Figure 5.4).

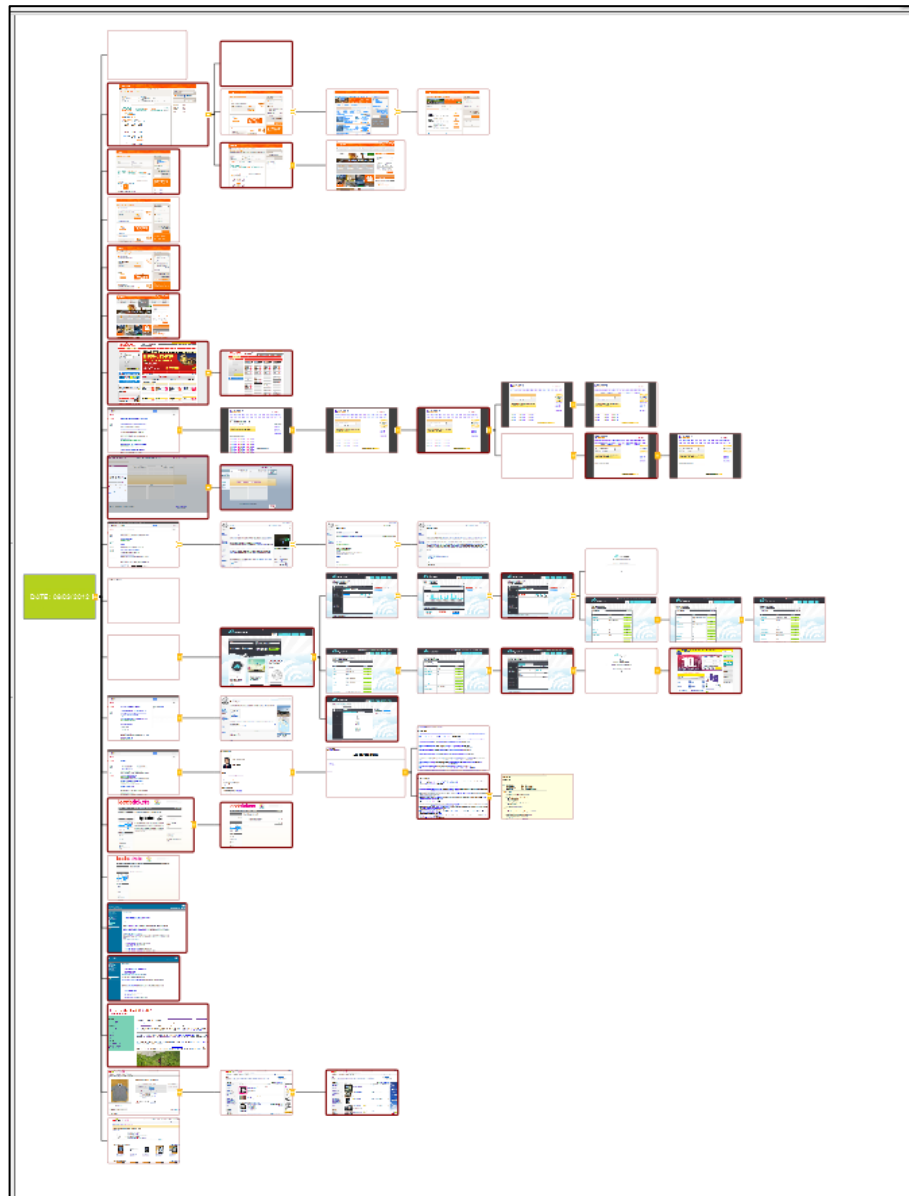


Figure 5.4 A horizontal orthogonal tree view with 63 nodes. Each node represents a webpage by its thumbnail. Frequency of visits to a page is encoded by node size, and edges are connected based on a user's navigational path.

As discussed in Chapter 2, there are several ways of representing a webpage as a node in a tree. This design uses a thumbnail because that is useful for user recognition. Other information is highlighted in the list view

when the mouse is over a node. Each node has a border that is the same as in the list view (maroon/blue).

Similar to WebNet (Cockburn and Jones, 1996) and SessionGraphs (Mayer and Bederson, 2001), the frequency of visits to a page is encoded by node size. The default width and height size of a node is the same as one in the list view. If the number of visits to a node is less than six, its width and height is calculated by the formula (2):

$$Node\ width(or\ height) = \left(1 + \frac{Frequency-1}{10} \right) * Default\ width(or\ height) \quad (2)$$

Otherwise its size is the same as a node with frequency of five. The root node always has the default width and height. This lets users have a standard to refer to. The original size of the thumbnail is 70% size of real webpage to save some hard disk space while ensuring the quality is sufficient for users to read the page content when the tree is zoomed in.

Relations between webpages are represented by tree edges. For example, if users go to a new webpage B (no matter if in the same or in a new tab) by clicking on a hyperlink in webpage A, there is an edge from A to B in the tree. If a webpage is visited by direct entry, there is an edge between it and a nominal root node R. However, as a webpage might have been reached from more than one webpage or even by direct entry, the question is which edge should be presented in the tree. The following steps are used to create the tree:

- 1) Create the overall network for the filtered set of webpages (e.g., all webpages visited on a date selected by a user) based on the definition of edge above (see an example in Figure 5.5).
- 2) Calculate weight for each edge to a node. If a webpage is visited by both a hyperlink and a direct entry, this design preserves the edge by hyperlink because it provides more contextual information. To do so, the weight of an edge to a node is created by 1 if it is visited by a hyperlink and by 0.1 if it is visited by a direct entry.
- 3) Create a spanning tree from the overall network using the Dijkstra's shortest path algorithm. To do so, another weight, named *DWeight*, for each edge is calculated from the weights in step 2. *DWeight* is calculated by the formula (3):

$$DWeight = \left(\frac{No.of\ visits\ to\ node}{Weight\ of\ that\ edge} \right)^2 \quad (3)$$

The example below of a browsing session is used to explain the solution. In the graph, each node represents a different webpage and the number above

each link represents the number of times a user travels via that link. Let's care about nodes with more than one incoming link: nodes 2, 5 and 8.

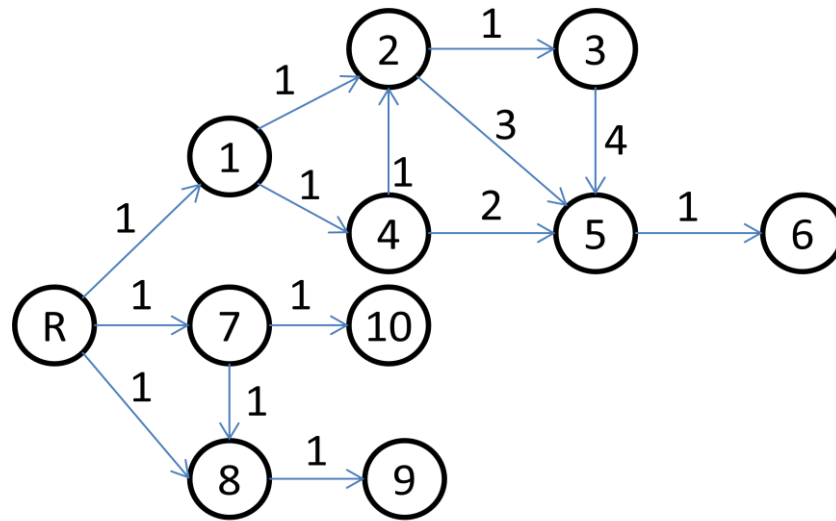


Figure 5.5 An example of the overall network for the filtered set of webpages: nodes connected to the nominal node R are visited by direct entry, nodes connected to other nodes are visited by hyperlinks; the number on a edge presents the number of times users use that edge to visit a webpage.

With node 2:

- Number of visits to node 2 = 2.
- Weight of link from node 1 to node 2 = $(2/1)^2 = 4$. So the length of the path $R \rightarrow 1 \rightarrow 2 = 0.1 + 4 = 4.1$.
- Weight of link from node 4 to node 2 = $(2/1)^2 = 4$. So the length of the path $R \rightarrow 1 \rightarrow 4 \rightarrow 2 = 0.1 + 1 + 4 = 5.1$.

So the shortest path from R to node 2 is not via node 4. So the link from node 4 to node 2 is removed in the spanning tree (see Figure 5.6).

With node 5:

- Number of visits to node 5 = 9.
- Weight of link from node 2 to node 5 = $(9/3)^2 = 9$. So the length of the path $R \rightarrow 1 \rightarrow 2 \rightarrow 5 = 0.1 + 4 + 9 = 13.1$.
- Weight of link from node 3 to node 5 = $(9/4)^2 = 5.0625$. So the length of the path $R \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 5 = 0.1 + 4 + 1 + 5.0625 = 10.1625$.
- Weight of link from node 4 to node 5 = $(9/2)^2 = 20.25$. So the length of the path $R \rightarrow 1 \rightarrow 4 \rightarrow 5 = 0.1 + 1 + 20.25 = 21.35$.

The path $R \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 5$ is the shortest one although it doesn't have the least nodes (see Figure 5.6). Squaring the link weight is important as it puts more weight on an important link. Without it, the lengths of the paths

would be 5.1, 5.35, and 5.6 respectively and the shortest path would be $R \rightarrow 1 \rightarrow 2 \rightarrow 5$.

With node 8:

- Number of visits to node 8 = 2.
- Weight of link from node 7 to node 8 = $(2/1)^2 = 4$. So the length of the path $R \rightarrow 7 \rightarrow 8 = 0.1 + 4 = 4.1$.
- Weight of link from node R to node 8 = $(2/0.1)^2 = 400$. So the length of the path $R \rightarrow 8 = 400$.

Due to this, the link from node 7 to node 8 is kept in the tree, reflecting the fact that webpages 7 and 8 have been visited together. The final spanning tree would look like Figure 5.6.

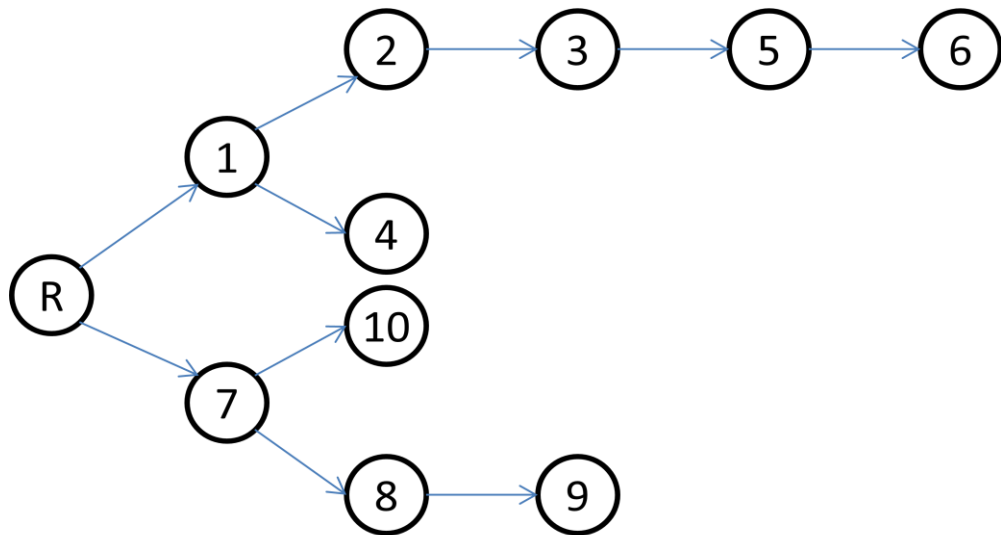


Figure 5.6 The final spanning tree for the browsing session in Figure 5.5 after applying the Dijkstra algorithm.

Right clicking on a node displays a dialog with detailed information about that webpage and navigation options (see Figure 5.7).

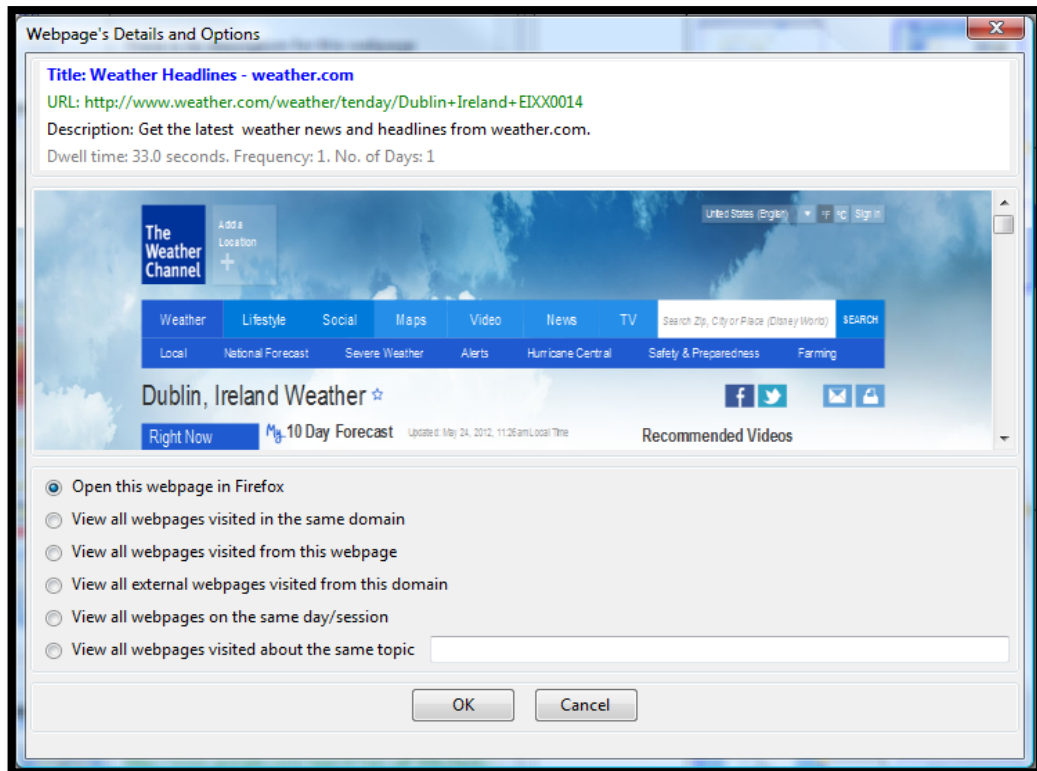


Figure 5.7 The dialog opened by right clicking on a node in the tree view displays a webpage's details and navigation options.

Once a desired webpage is recognised in the *Result View*, users can double click on its thumbnail either in the list view or tree view to open it in a new tab of the Firefox browser. Double clicking is chosen over left clicking because left clicking is a part of dragging and dropping action which is used very often for zoom and pan.

In the list view only mode, when the mouse is over a small thumbnail, its full size thumbnail is displayed in the tree view area so users can glimpse through pages easily.

5.3.3 The visibility of the tool as an add-on in Firefox

Visibility is one of the key principles in user computer interaction design (Norman, 1988). Controls need to be visual so icons of the tool are placed in the status bar of the Firefox browser (see Figure 5.8).

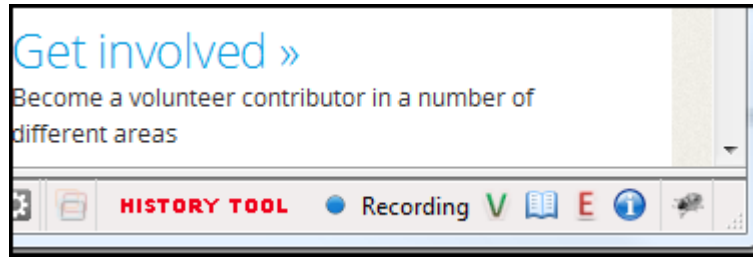


Figure 5.8 Icons of the visualization history tool, situated in the status bar of the Firefox web browser.



5.3.4 Feedback from users

After the presentation to the research group and one month of testing with three users, the first version of the history received some positive feedback. Regarding installation, the history tool was easily integrated into Firefox on Windows, Linux, and Mac. It was compatible with different versions of Firefox (from 3.6 to 10.0 as in April 2012¹⁶) and ran smoothly without any noticed error. In general, all the three users agreed that it was straightforward to use the tool. However, some new suggestions about usability and functionality were also collected. First, the researcher remarked: “It is good to have the tree view fitted in the tree view area initially. But after I zoom and pan the tree, I would like to quickly get back to that initial presentation but not the ‘home’ state”. He also suggested making use of the ROOT node rather than just displaying the word ROOT. The academic and another person in the presentation said they were a little confused about the blue border of some webpages. They thought those webpages were selected and did not know why. The academic also suggested that the three sliders should be placed on the *Toolbar* as they apply on both the list view and tree view. Finally, the PhD student thought a search feature would be useful in case people remember neither the domain, visited date nor the search query of a webpage.

5.4 The third iteration: the refined visualization history tool

Suggestions from the research group and the three users were carefully considered. Then the following changes were made accordingly: (1) the three sliders were moved to the *Toolbar* (see Figure 5.9); (2) a *Search box* was provided under the *month view calendar* so a search query can be keyed in; Webpages whose titles or descriptions contain that search query

¹⁶ See Firefox release history:
http://en.wikipedia.org/wiki/Firefox_release_history

are displayed in the *Result View*; (3) the “Fit to Screen”  button was added to take the tree back to its initial view. In addition, the “Month view”  button allows users to view their web history of a selected month; (4) one colour (maroon) is used for all titles/borders, and this is made bold to emphasise pages of interest (see Figure 5.10); and (5) the ROOT node was enhanced to provide navigation information. It tells users where they are and where they have been. For example, if users click on a date in the calendar, the ROOT node displays that date. If the users click on a domain within that date, the domain is displayed with the date (see Figure 5.11). The final design of the visualization history tool is shown in Figure 5.12.

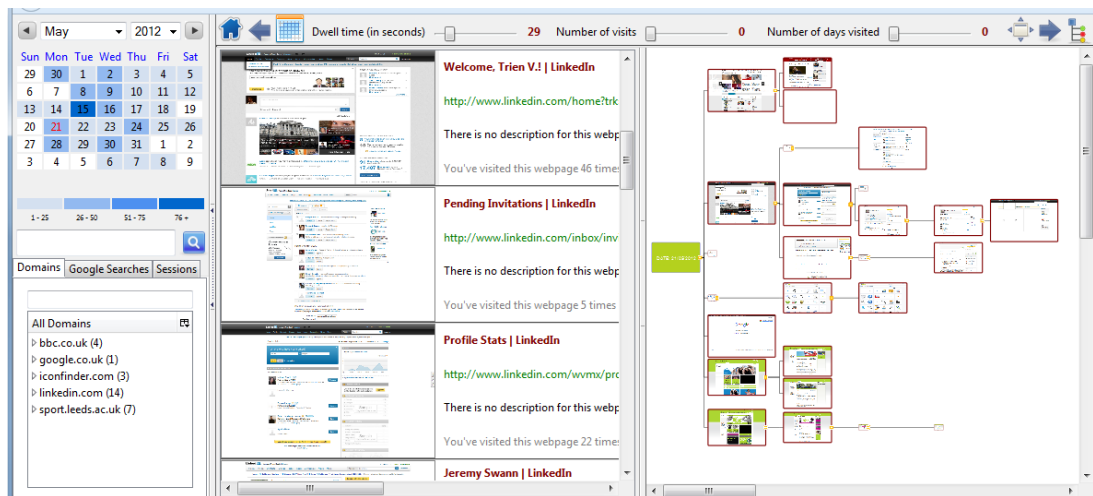


Figure 5.9 Sliders are placed in the Toolbar. Webpages that do not satisfy the filter criterion are removed from the list view but are shown in the tree view with reduced size.



Figure 5.10 An example of the list view entries, with a bold maroon title for a webpage of interest (dwell time \geq 30 seconds) and normal title for another webpage.

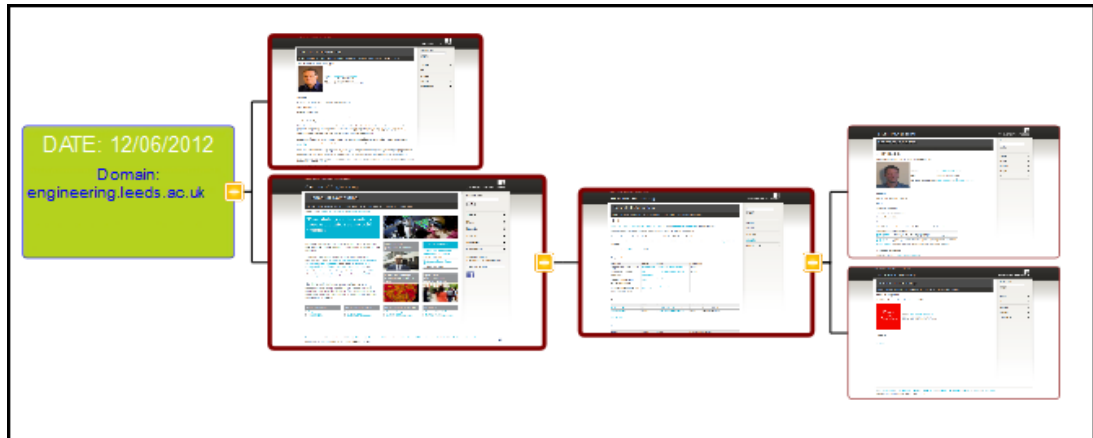


Figure 5.11 The tree's ROOT node is used to provide general information.

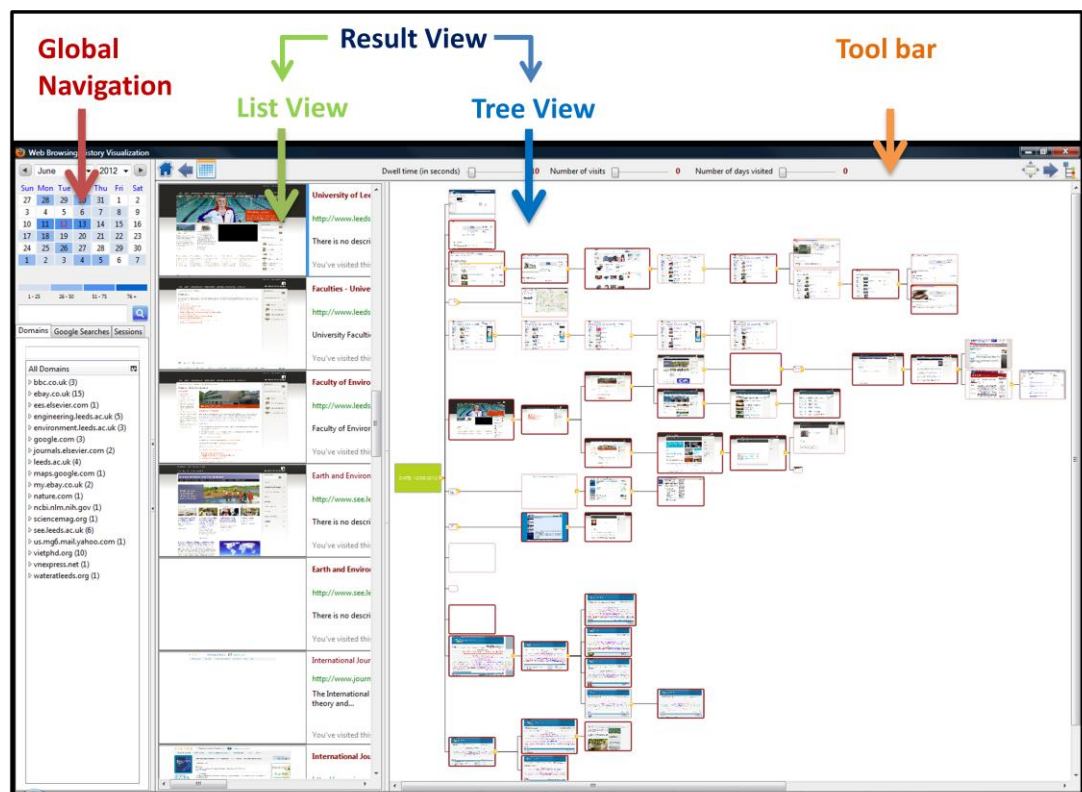


Figure 5.12 The final visualization history tool with: the Global Navigation at the left with a heat map calendar, a search box, and a tab view; the Result View at the right with a list view and a tree view; and the Toolbar at the top containing buttons and sliders.

5.5 Technical implementation

To meet the data and environmental requirements, the visualization history tool has been developed as another module of the logging tool (see Chapter 3). This section explains the technologies that have been chosen and describes important aspects of the implementation.

5.5.1 Technologies chosen

As it is implemented as a module of the logging tool, the visualization history tool uses all the technologies discussed in Chapter 3. To render the tree view, both HTML5 Canvas and SVG are good candidates to create graphics inside the browsers. However Canvas has better performance when many objects are redrawn frequently (Microsoft, 2012). That is why HTML5 Canvas¹⁷ has been chosen. Canvas tags can be inserted easily and seamlessly into an XUL element as they are both mark-up languages. Finally, Cascading Style Sheets (CSS) are utilised to create styles for both the XUL and HTML components of the tool. This makes the user interface consistent and easy to maintain. Other options like Flash, Silverlight or Java Applet could have been used, but they all require users to install other plug-ins.

Besides that, Dreamweaver has been used to write source code and Firebug and Chromebug are used to debug directly on the Firefox browser. Although there are several libraries available for visualization with HTML Canvas such as Processing, Raphaël, JavaScript InfoVis Toolkit, and Protovis (Wiederkehr, 2009), none of them was used. A simple implementation of a tree layout (Cl, 2006) is modified to implement the design.

5.5.2 Tool architecture

Figure 5.13 shows the architecture of the history tool. As described in Chapter 3, the logging module is integrated into the browser to keep track of all visited webpages and store information about them in a SQLite database. The visualization module retrieves data from the SQLite database and presents to users.

¹⁷ See http://www.w3schools.com/html/html5_canvas.asp

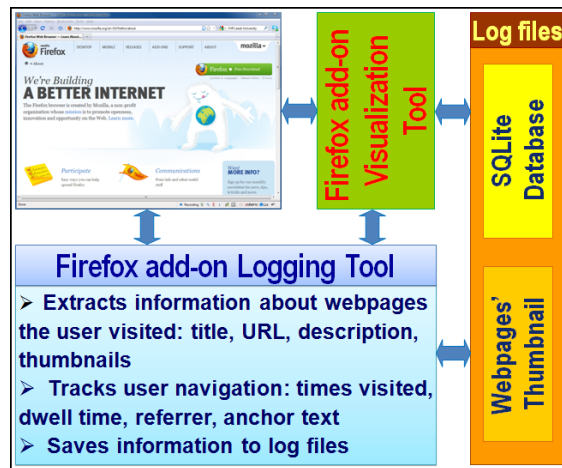


Figure 5.13 The architecture of the visualization history tool: The logging module tracks and saves data about user web history in a SQLite database then the Visualization module retrieves and presents them to users.

5.5.3 Implementation details

This section highlights key details of the implementation of the visualization history tool. Some of the details are described as a list of steps and pseudo code is provided for others.

5.5.3.1 The heat map calendar

As a month can span six different weeks (see Figure 5.14), a frame for six weeks is needed.

| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|-----|-----|-----|-----|-----|-----|-----|
| 26 | 27 | 28 | 29 | 30 | 31 | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| 30 | 1 | 2 | 3 | 4 | 5 | 6 |

Figure 5.14 An example of a month view calendar which spans six different weeks.

The algorithm for drawing the heat map calendar has two steps:

1. Create a frame for the calendar: To improve the performance the frame is drawn only once then its content is updated later in response to user's interaction. A table is created with 6 rows x 7 columns. Content of each cell is embedded in the span tag so it can be accessed and updated later.

2. Fill data into the calendar: When a month or a tab is selected, the algorithm in Pseudo code 5.1 is run:

Pseudo code 5.1 Updating the calendar.

```
function updateCalendar()
{
    //Fill dates into the month calendar
    //get selected month and year by users
    var selMonth = getSelectedMonth();
    var selYear = getSelectedYear();
    var objDate = new Date();
    objDate = firstDate(selMonth, selYear);
    //e.g: if September of 2012 is selected, objDate = 01/09/2012

    //Update data for the first row of the calendar
    //e.g: from 26 - 31
    var objDayOfWeek = getDayOfWeek(objDate);
    //objDayOfWeek = Saturday
    var maxDayOfPreviousMonth = getMaxDay(selMonth=1?12:selMonth-1);
    //maxDayOfPreviousMonth = 31 as August has 31 days
    printLastDatesOfPreviousMonth();
    //As the first day of September is Saturday, From Sunday to Friday of
    //The first row of the calendar are of August
    //So 26, 27, 28, 29, 30, 31 must be printed first
    var noOfItem = getNumberOfVisits_SearchesForDate(aDate);
    updateBackgroundColourForCell(noOfItem);

    //Update data for the current month
    //e.g from 1 to 30 of September
    var maxDayOfMonth = getMaxDay(selMonth);
    printAllDatesOfSelectedMonth();
    noOfItem = getNumberOfVisits_SearchesForDate(aDate);
    updateBackgroundColourForCell(noOfItem);

    //Update data for the rest of the calendar
    //e.g from 1 - 6 October
    printAllDatesOfNextMonth();
    noOfItem = getNumberOfVisits_SearchesForDate(aDate);
    updateBackgroundColourForCell(noOfItem);
}
```

5.5.3.2 The Google searches tab

When a page is loaded in the browser, a check is made to see the URL is a Google search page. If yes then the search query is extracted from the URL. The Google search page has the syntax like:

http://www.google.com/search?q=golden+triangle+of+europe&ie=utf-8&oe=utf-8&aq=t&client=firefox-a&rlz=1R1GGHP_en-GB__GB462

The search query is between the first token “?q=” and the next “&”, and the search query can be extracted by string operations. The search query is stored as a field for each webpage. An SQL statement retrieves all the search queries and fills in the Google Searches tab.

5.5.3.3 Search session reconstruction

Tracking webpages clicked from Google search is a bit more challenging. Normally, if a search query such as “how to create update firefox add-on” is typed into Google search bar, the results page will be something like:

http://www.google.com/search?q=how+to+create+update+firefox+add-on&ie=utf-8&oe=utf-8&aq=t&client=firefox-a&rlz=1R1GGHP_en-GB_GB462

If a webpage is selected from this results page, its referrer attribute should be the URL above. However, wanting to keep track of what results have been clicked, Google directs these results via another webpage. One of the pages clicked from the above results page has the following referrer:

http://www.google.com/url?sa=t&rct=j&q=create%20update%20firefox%20add-on&source=web&cd=18&ved=0CGEQFjAHOAo&url=http%3A%2F%2Fstackoverflow.com%2Fquestions%2F6484749%2Fxp-create-update-rdf-for-previous-version&ei=ZZ7tT-yvNIO30QXG_oHQDQ&usq=AFQjCNGS2SL8348TtkvoGtyVO5NvAAAYNA&cad=rja

This is rather different to the original referrer. Such a page can be determined when its referrer contain the two tokens “?sa=” and “&q=”, and needs further analysis. The search query extracted from this referrer is then compared with the search query of the current Google search to find out the exact referrer.

In addition, after clicking on a Google search results, users might browse further to other pages. To reconstruct the whole search trail, the following Pseudo code 5.2 is employed:

Pseudo code 5.2 Reconstructing the search session of a search query.

```
function reconstructSearchTrail(selectedQuery)
{
    //A query might be launched on more than one day
    var queryDates = getDatesQueryLaunched(selectedQuery);
    //Get all webpages on those dates, order by visited time asc
    var allPages = getAllWebpageOfDates(queryDates);

    //Identify if a page belongs to the search session
    var searchSession;
    var parentID;
    for (each objPage in allPages)
    {
        //check if it is a google search page of the selectedQuery
        if (objPage.URL.indexOf("google") > 0
            && objPage.searchTerm == selectedQuery)
        {
            //a google search page
            searchSession.push(objPage);
            //store its visitedTime which is used as a pageID
            //if other pages have referrer ID is this pageID
            //it was visited via hyperlink from this page and
            //it is of the search session too
            parentID.push(objPage.visitedTime);
        }
        else if(objPage.referrerID in the parentID)
        {
            searchSession.push(objPage);
            parentID.push(objPage.visitedTime);
        }
        else if(objPage.referrer.indexOf("google")
            //a page clicked on google search page
            //and directed via another page
            //containing the two tokens "?sa=" and "&q="
            {
                if(extractQuery(objPage.referrer) = selectedQuery)
                {
                    searchSession.push(objPage);
                    parentID.push(objPage.visitedTime);
                }
            }
    }
    return searchSession;
}
```

5.5.3.4 Tree construction

The Pseudo code 5.3 implements the steps of building the tree described in Section 5.3.2.2.

Pseudo code 5.3 Creating a tree for a filtered set.

```
function createTree(pageSet)
{
    //STEP1: Create the overall network for a filtered
    //set of webpages : pageSet (order by visitedTime asc)
    var Root = createNewNode();
    var visitedURLs;parentID;treeNodes;
    for(each objPage in pageSet)
    {
        var newNode;
        if(objPage.URL in visitedURLs)
        {
            //This URL has been visited more than once
            //In the filtered set --> Dont create a new node
            newNode = getNodeOfURL(objPage.URL);
        }
        else
        {
            //New URL --> create new node
            newNode = createNewNode();
            treeNodes.add(newNode);
            visitedURLs.push(objPage.URL)
        }
        parentID.push(objPage.visitedTime);
    }
    //Find parent
    if(objPage.referrerID in parentID)
    {
        var parentNode = getParentForNode(objPage);
        parentNode.addChild(newNode)
    }
    else
        Root.addChild(newNode)

    //STEP2: Calculate weight for each edge to a node
    // 1 for hyperlink - 0.1 for direct entry
    for(each objNode in treeNodes)
    {
        var uniqueNode;
        for(each objChild in objNode.children)
        {
            if (objChild not in uniqueNode)
            {
                objChild.weight = 1;
                uniqueNode.push(objChild);
            }
            else
            {
                var tmpNode = getExistingNode(objChild);
                tmpNode.weight += 1;
                objNode.children.remove(objChild);
            }
        }
        for(each objChild in Root.children)
        {
            objChild.weight = 0.1;
        }
    }

    //STEP3: Calculate Deight for each edge and apply Dijkstra
    // Calculate weights
    for(each mNode in treeNodes)// Loop over each node
    {
        var sum = 0;
```



```
// Calculate total number of visits to node mLoop,  
// and then use that sum to set the weight  
for(var mLoop = 0; mLoop < 2; mLoop++)  
//Trick:loop1 to calculate sum,loop 2 to calculate weight  
{  
    for(each mNodeRun in treeNodes) //With each node  
    {  
        for(each child in mNodeRun.children)  
        {  
            if(child = mNode)// Link from child to mNode  
            {  
                if (mLoop == 0)  
                    sum += child.weight;  
                else  
                {  
                    var ww = (10 * sum) / child.weight;  
                    child.weight = ww * ww;  
                }  
            }  
        }  
    }  
}  
}  
applyDijkstra();  
}
```

5.5.3.5 Tree layout algorithm

The tree layout algorithm by Walker (Walker, 1990) is implemented. The algorithm is designed to occupy as little space as possible while satisfying the following aesthetic rules:

- Nodes at the same level of the tree are aligned, and the straight lines defining the levels should be parallel.
- A parent should be centred over its children.
- A tree and its mirror image should produce drawings that are reflections of one another. Consistently, any subtree should be drawn in the same way.

5.5.3.6 Filters

The tool allows filtering by three attributes: dwell time, frequency, and number of visited days. As explained in Section 5.2.1.2, the size of unqualified nodes is reduced in the tree view. Pseudo code 5.4 summaries how the filtering functions have been implemented.

Pseudo code 5.4 Filtering the tree view by resizing nodes.

```
function filterTree(time, freq, noD)
{
  for(each objNode in treeNodes)//But not Root
  {
    if(objNode.dwellTime < time or objNode.frequency < freq
      or objNode.noD < noD)
    {
      //A node might be not qualified in a previous filter
      if(objNode.isUnqualified = false)
      {
        //Store its old pos to restore when it becomes qualify
        objNode.oldX = objNode.X;
        objNode.oldY = objNode.Y;
        //Calculate new position for it
        calculateNewPosition(objNode);
      }
      objNode.width = REDUCED_WIDTH;
      objNode.height = REDUCED_HEIGHT;
      objNode.isUnqualified = true;
    }
    else
    {
      CalculateNodeSize(objNode);//Based on its frequency
      if(objNode.isUnqualified = true)
      // not qualified in a previous filter
      {
        //Restore its position
        objNode.X = objNode.oldX;
        objNode.Y = objNode.oldY;
      }
      objNode.isUnqualified = false;
    }
  }
  RedrawTree();
}
```

5.5.3.7 Back and forward

Back and Forward buttons allow users to navigate through the five most recent history visualizations. To improve the performance, all data of each visual presentation are stored such as nodes, their parent nodes, selected options with values (e.g., domain, domain name). Note that data of the visualization generated again by Back/Forward themselves are not stored.

5.6 How the new tool addresses the causes of failure

The above sections describe the design and implementation of each component of the visualization history tool. To illustrate how the tool would work, this section describes some user cases. The first five cases show how the five underlying causes of failure stated in the requirements of the tool (see Section 5.1) are addressed. The last case is about when users

remember the date (e.g., earlier today, yesterday, last weekend) that they last visited the target page.

5.6.1 Known website

Scenario: A user remembered that she read news on her university website and browsed to information about the “Turing centenary conference”. She did not remember much about the name of the conference except the name “Turing”. A couple of days later, she talked to her friend and this friend mentioned about this conference again. When she got back to her desk, she wanted to look at the information about the conference again.

What she should do: Remembering the university website, she types “leeds.ac.uk” in the filter box of the domain list. However, she realises that the domain has some sub-domains (e.g., accommodation.leeds.ac.uk, comp.leeds.ac.uk). So she makes a guess that the information should be in either school of computing or maths. She clicks on the “comp.leeds.ac.uk” first but does not recognise the desired webpage. Then she tries “mathcomp.leeds.ac.uk” and finds all webpages about the conference there (see Figure 5.15).

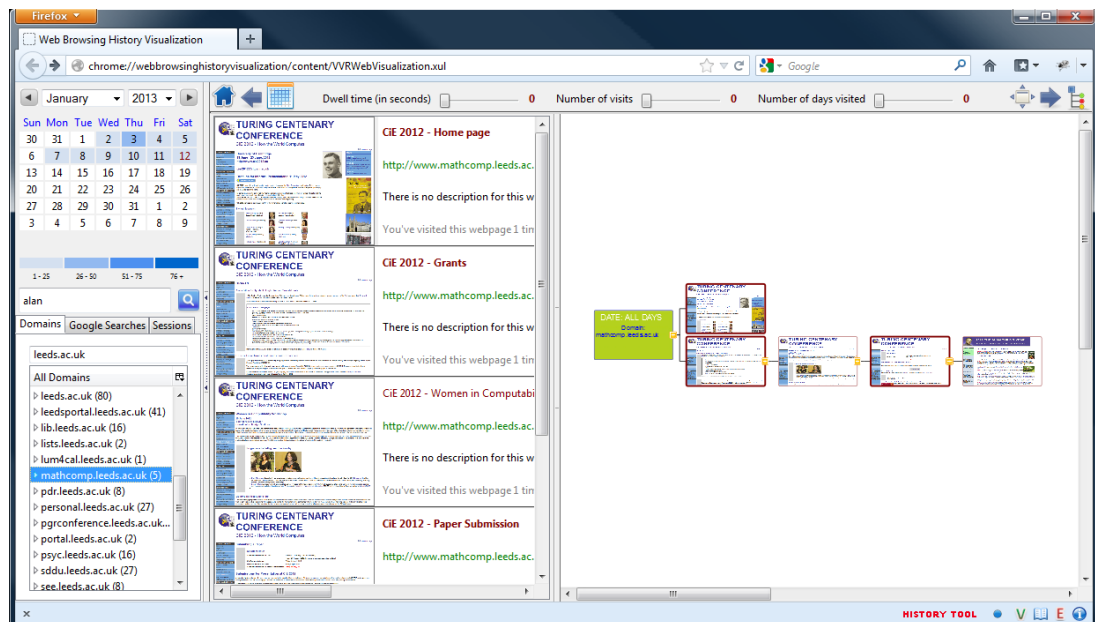


Figure 5.15 A user selects the domain of the target page from the domain list, to see all webpages within that domain.

5.6.2 Search results

Scenario: A user remembered that he searched information about the best places to work a while ago and wanted to go back to a page about Google of that search.

What he should do: He remembers his previous search query containing something like “best place”. As the list of search queries is too long he types in the filter box some characters and stops with the word “best” as there are only seven matched search queries in the list (see Figure 5.16). Then he sees the search query “best places to work”, he clicks on that and observes the visualization of its search trail. He immediately notices three bold nodes and easily recognises the top right page is what he wants to revisit.

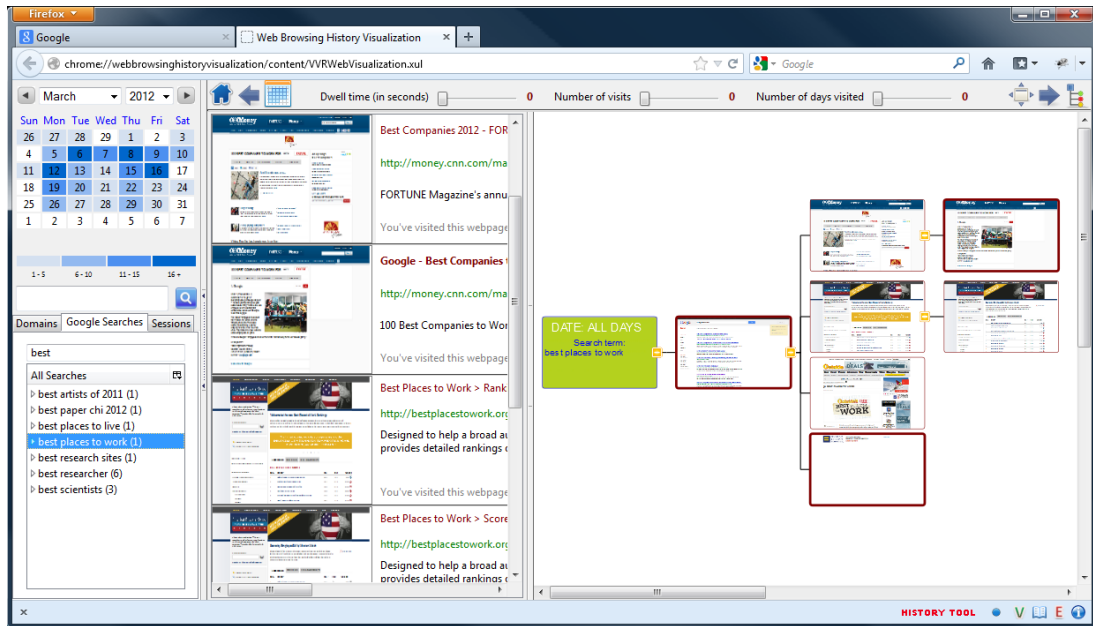


Figure 5.16 If a user knows the target page belonged to a search session, they can select the relevant search query from the list of searches.

5.6.3 Deleted links

Scenario: A user remembered that she read news about a study to discover how dust particles in the solar system interact with the Earth’s atmosphere, which was featured on the homepage of her university website. Several days later, she wanted to read that information again but when going back to that homepage, new stories had replaced old ones.

What she should do: She opens the visualization history tool. She had visited the homepage of the university again that day so she should easily locate it from the default visualization. Right clicking on that node in the tree view, she selects the option “View all webpages visited from this webpage” (see Figure 5.17). In this case, the tree view looks rather similar to the list view (see Figure 5.18), so she switches the tool to the “List view only” mode to see full thumbnails of webpages in the list view area (see Figure 5.19). Scrolling through the list view, she then recognises the webpage she is looking for again.

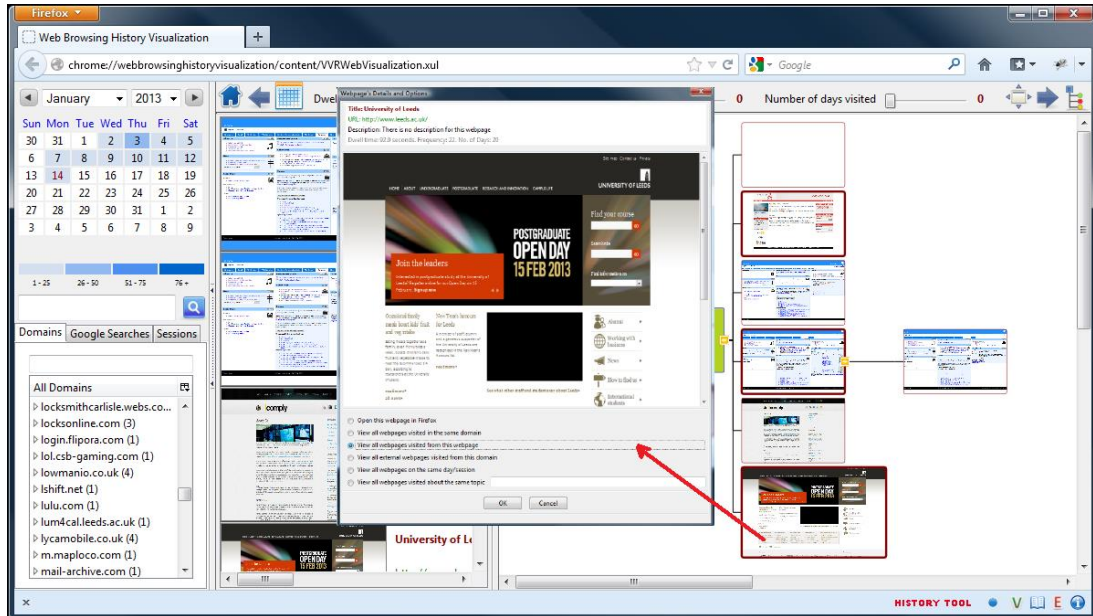


Figure 5.17 Right clicking on a node to select the “View all webpages visited from this webpage” option.

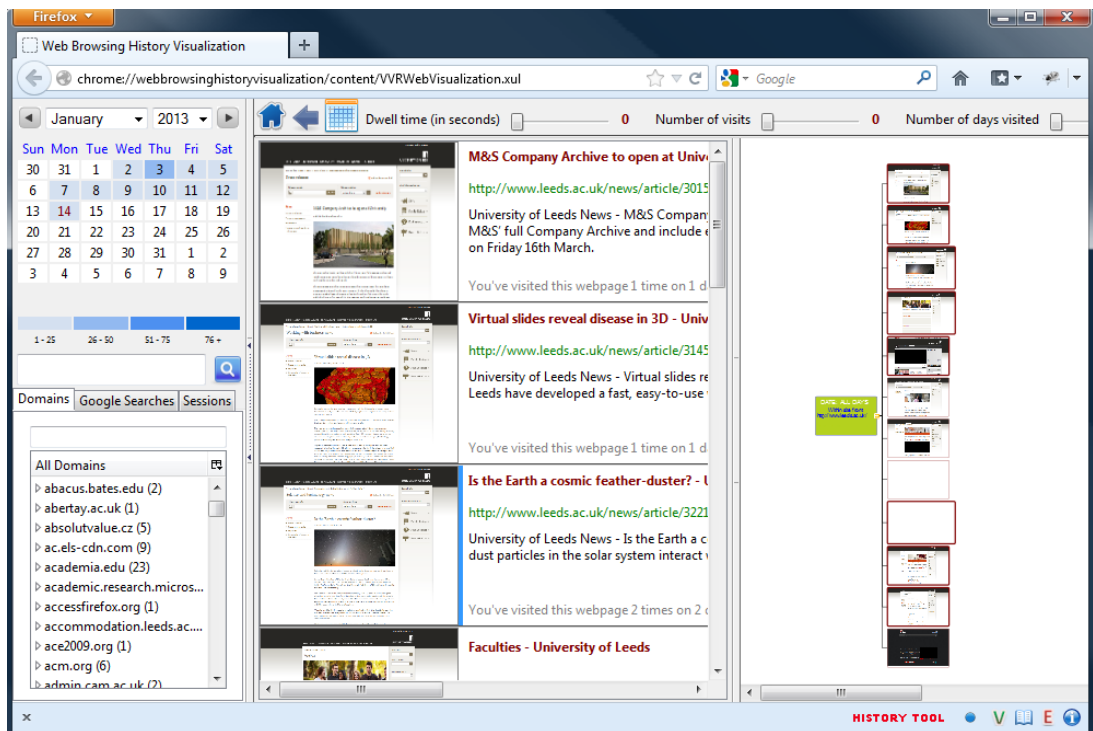


Figure 5.18 Example of a tree view that looks similar to the list view.

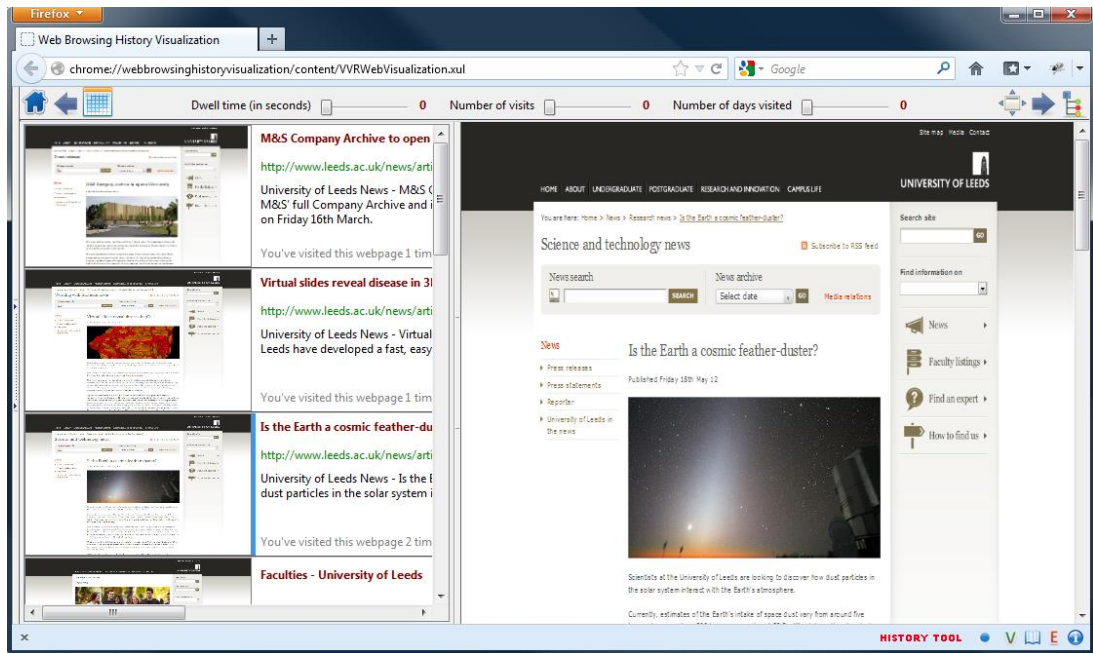


Figure 5.19 The “List view only” mode displays the full thumbnail of a webpage.

5.6.4 Links from email & social networks

Scenario: A user often visited a forum and occasionally clicked on links shared by other members. Once, he needed to revisit one of those shared webpages. However, there were so many new threads and new posts on the forum since then.

What he should do: He opens the visualization history tool and right clicks on any webpage of the forum in the tree view then selects the option “View all external webpages visited from this domain”. Similar to the “Deleted Links” case above, he switches to the “List view only” mode and quickly finds the desired webpage.

5.6.5 Topic

Scenario: A user spent two weeks researching bikes before deciding to buy one. He had viewed a lot of webpages. He finally decided to order one of the bikes he had seen and needed to go back to that webpage. The problem was all the bike webpages had similar content and he could not identify the webpage he wanted until he could see it.

What he should do: He just types “bike” in the search box to see all the webpages he had visited that were about bikes (see Figure 5.20). Then he scrolls the list to find the bike he wants.

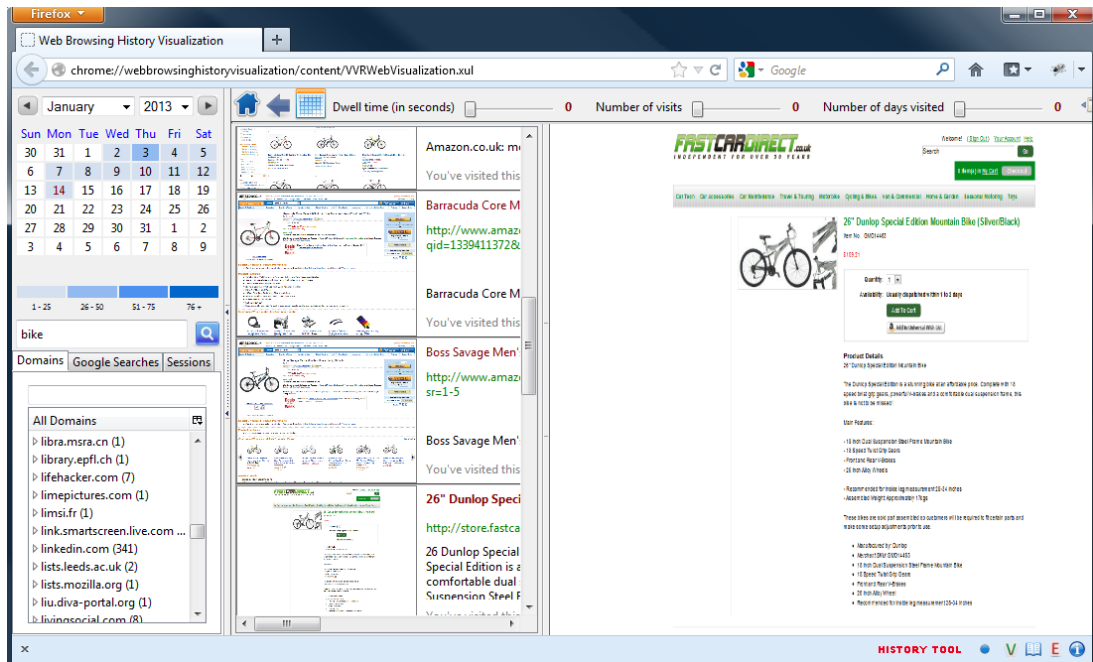


Figure 5.20 To display all visited webpages that are about the topic “bike”, a user types “bike” in the search box.

5.6.6 Knowing the visited date

Scenario: A user read an online article earlier today but it was too long to finish at one go. So, she decided to leave it there for lunch time. However, while working, she accidentally closed that tab of the web browser as she had opened too many tabs. At the lunch time she could not find that tab again.

What she should do: She just opens the visualization history tool. By default, the tool displays the web history of that current date. She then could easily recognise the article from the visualization.

The underlying cause addressed: Although this case study does not directly address any causes of failure, if users can remember a specific date that they last visited a target webpage, the calendar would be useful. The visualization history tool has been designed to address the difficulties when people revisit webpages of the 4th group (see Chapter 1), however it can also be used to revisit webpages of other groups.

5.7 Summary

This chapter describes the requirements, design, and technical implementation of a new visualization history tool that addresses the five main causes of failure: 1) Topic, 2) Search results, 3) Known website, 4) Deleted link, and 5) Links from email & social networks.

Three key functional requirements of the tool are: (1) allowing users to navigate their history to select a small set of possible webpages from the whole history, (2) presenting the selected set of webpages in a visual way so users can recognise and choose the right page, and (3) supporting filters or other ways of navigating the selected set in case users still find it difficult to find the target page.

Using the automatically recorded web history approach (see Section 2.3), the tool exploits visualization techniques to support both browsing and searching mechanisms in revisiting an individual's complete web history (see Section 2.2 and 2.4). The user interface has three main components: the *Global Navigation*, the *Result View*, and the *Toolbar*. The *Global Navigation* lets users navigate within their web history with a heat map calendar, a tab view with a list of web domains/search queries/sessions, and a search box. These different navigation techniques enable the tool to manage a complete web history. This is one of the main novelties of the present visualization history tool. The *Result View* displays results of every navigation in both a list view and a tree view. The *Toolbar* allows users to perform actions like going back to the default state (home), going back/forward navigation actions, fitting the tree to the tree view area, and filtering.

The tool encodes and presents important data (e.g., user interest in a page (dwell time ≥ 30 seconds), frequency, recency, and associations (links between webpages)) from a user's web history in both a list-view and a tree view. Drawing on the well-known presentation style of Google's search results, the tool enhances the list view by adding a small thumbnail, colour-codes titles of webpages to indicate whether they were navigational or informational. The tree view is created from the user's navigational paths (even crossing different tabs) rather than the sequence of visits over time. Edges of the tree are also weighted to help ensure that the tree includes the links that the user most often traversed. The way the tree view is created is another main novelty of the present visualization history tool.

The design has been developed as a Firefox add-on. Six user cases illustrating how the tool would help users in revisiting are described. Although these user cases address the five underlying causes of failure, the visualization history tool might also solve other causes of failure. For example, the approach for the Known website cause can be applied to the Search on specific website and the Multi-page thread cause, the thumbnails of webpages enable users to recognise webpages with inappropriate page titles, and the navigational paths in the tree view might reveal hidden

information. The argument is whether real users would adopt the tool as it is intended. The next chapter presents an evaluation of the tool.

Chapter 6. The user evaluation of the history tool

This chapter presents a three month field study of how participants used the visualization history tool. The goals of this user evaluation were to: (1) explore how participants actually used the tool, (2) investigate whether such a tool solved the underlying causes of failure, as designed, and (3) learn what participants thought about the history tool. To do that, participants were asked to browse the WWW as usual, and use an electronic diary methodology to record occasions when they revisited webpages both with and without the tool. A participant's web navigation, usage of the tool, and diary entries were recorded electronically and saved in a logfile. At the end of the study, a semi-structured interview was conducted to clarify aspects of the diary entries and to learn what people thought about the tool.

6.1 Method

6.1.1 Participants

The study was approved by the University Research Ethics Committee. All the participants gave their informed consent. Participants were not paid for taking part in this study. Twenty-four individuals (seven females) commenced the study but two of them never responded after receiving the history tool, two others installed the tool but never used it, and another did not finish the study after buying a new laptop.

The data reported in the following sections were from the 19 participants (5 females) who completed the whole study. They were either students or employees of different universities and companies in the UK. Two participants were academic staff in the School of Computing, two were researchers in Computing, three were software engineers, one was a manager, and the rest were PhD students (one studying Earth and Environment, one Biology, and nine Computing). Four of them previously took part in the study described in Chapter 4. All participants had at least one year of experience with Firefox and three years of experience with navigating the WWW. Three of them only used their laptop during the course of the study. The rest used more than one device to access the WWW (one at home, the other at work, and maybe a mobile device (e.g., smart phone, tablet)). In this case, the tool was installed on the computer chosen by these

participants. One participant installed the tool on both of his computers. Two logfiles sent by this participant were merged for the final analysis.

6.1.2 Procedure

At the start of the study, participants were sent the visualization history tool, a user manual, an information sheet, and a consent form. The participants read the documents, signed and sent back the consent form, installed the tool, and were encouraged to ask questions. A quick guide (see Figure 6.1) was displayed to them after a successful installation.

A week later, a follow-up email was sent to each participant to ensure that they had no problem with installing and using the tool. During three months of the study, every two weeks, another email was sent to remind participants to fill in the diary form when they wanted to report a revisit. At the end of the study, participants were instructed how to send the logfiles back. Logfile of each participant was then briefly reviewed for a follow-up interview. Depending on participants' preferences, the interviews were in person, or via Skype.


| Quick guide | |
|--------------------|---|
| Zoom In/Out | Mouse wheel |
| Pan | Mouse Drag and Drop |
| Open a webpage | Double click on a node in tree or list |
| Fit tree to Screen |  |
| Other options | Right click on a node in the tree |

Figure 6.1 The quick user guide.


6.1.3 Logging participants' activities

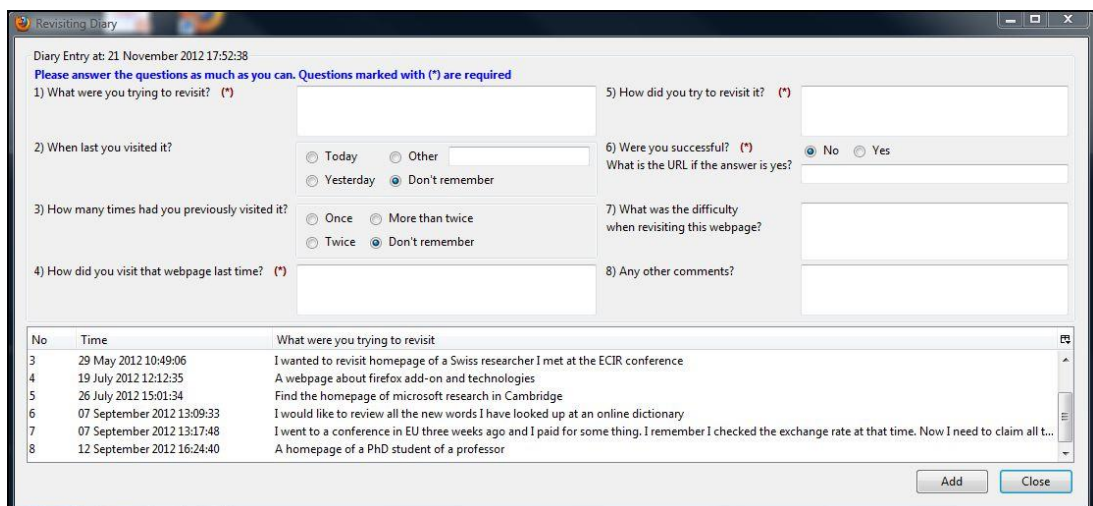
The tool logged participants' web history as described in Chapter 4, and participants' usage of the visualization history tool. For the latter, recorded actions were: clicking on a date in the calendar, clicking on an entry of the domain list, search list, session list of the tab view, typing in filter boxes of the domain and search list, typing in the search box, opening a webpage from the list view and tree view, right clicking on the tree view to open the overview dialog and each option selected there (see Chapter 5), clicking on the collapse/expand buttons in the tree view, filtering (by dwell time, frequency, days visited), and clicking on buttons (e.g., Home, Back, Forward, Visualize whole month, Fit to screen, and switch List/Tree mode). Figure 6.2 shows an excerpt of a participant's history.

| usageTime | humanTime | actionType | actionData |
|---------------|--------------------------|--------------------|----------------------------|
| 1339767530223 | Fri Jun 15 14:38:50 2012 | PAN | VVRVVRVVR |
| 1339767532756 | Fri Jun 15 14:38:52 2012 | PAN | VVRVVRVVR |
| 1339767787003 | Fri Jun 15 14:43:07 2012 | CHANGETAB | 0 |
| 1339767787008 | Fri Jun 15 14:43:07 2012 | CHANGEDATASET | { "dataSet": [{"nodeInfo": |
| 1339767787022 | Fri Jun 15 14:43:07 2012 | changeCalendar | 6#2012 |
| 1339767794816 | Fri Jun 15 14:43:14 2012 | OpenPageFromTree | http://pngu.mgh.harva |
| 1339767801727 | Fri Jun 15 14:43:21 2012 | OPENOVERVIEWDIALOG | http://pngu.mgh.harvar |
| 1339767830626 | Fri Jun 15 14:43:50 2012 | CHANGETAB | 0 |
| 1339767830631 | Fri Jun 15 14:43:50 2012 | CHANGEDATASET | { "dataSet": [{"nodeInfo": |
| 1339767830644 | Fri Jun 15 14:43:50 2012 | changeCalendar | 6#2012 |

Figure 6.2 An excerpt of the logfile of a participant's activity, recorded by the history tool.

6.1.4 The diary form

The diary form was automatically displayed when a participant opened a webpage using the tool (see Figure 6.3). Participants could also manually activate the form by clicking the  icon, e.g., if they were in the middle of something when revisiting a webpage and decided to fill the diary form later. Participants were asked to provide the diary information in as much detail as possible. Questions marked with (*) were compulsory. Answers for questions 2, 3, 4, and 5 could also be derived from the tool usage log if participants always filled in the diary form immediately when they revisited a webpage. Question 1, 6 and 7 provided insights into what a participant was looking for, if they were successful and any difficulties encountered. Based on this information, further data could be retrieved from the logfile and analysed.



Diary Entry at: 21 November 2012 17:52:38
Please answer the questions as much as you can. Questions marked with (*) are required

1) What were you trying to revisit? (*)

2) When last you visited it?
 Today Other
 Yesterday Don't remember

3) How many times had you previously visited it?
 Once More than twice
 Twice Don't remember

4) How did you visit that webpage last time? (*)

5) How did you try to revisit it? (*)

6) Were you successful? (*) No Yes
What is the URL if the answer is yes?

7) What was the difficulty when revisiting this webpage?

8) Any other comments?

| No | Time | What were you trying to revisit |
|----|----------------------------|---|
| 3 | 29 May 2012 10:49:06 | I wanted to revisit homepage of a Swiss researcher I met at the ECIR conference |
| 4 | 19 July 2012 12:12:35 | A webpage about firefox add-on and technologies |
| 5 | 26 July 2012 15:01:34 | Find the homepage of microsoft research in Cambridge |
| 6 | 07 September 2012 13:09:33 | I would like to review all the new words I have looked up at an online dictionary |
| 7 | 07 September 2012 13:17:48 | I went to a conference in EU three weeks ago and I paid for some thing. I remember I checked the exchange rate at that time. Now I need to claim all t... |
| 8 | 12 September 2012 16:24:40 | A homepage of a PhD student of a professor |

Add Close

Figure 6.3 The diary form.

6.1.5 The follow-up interview

Prior to the interview, a participant's diary entries were reviewed. The follow-up interview was 10 minutes long, and semi-structured using the questions shown in Figure 6.4 below. The purpose of the interview was to explore participants' thoughts about the tool and, if necessary, to clarify the meaning of diary entries.

Follow-up Interview

1) Did you read the user manual? Yes No

2) How would you rate the ease of use of the tool?
Difficult 1 2 3 4 5 Easy

3) Did you notice the information encoded

- Colours of the heat map calendar Yes No
- Bold title or border of nodes Yes No
- Thumbnail size Yes No
- Tree reconstruction Yes No

4) How would you rate your satisfaction with the tool?
Not satisfied 1 2 3 4 5 Very satisfied

5) Will you keep using the tool? Yes No

6) What was the reason you didn't fill in the diary form?

7) Did you report all cases that you could not find a wanted webpage with the tool?
Yes No

8) Do you have any difficulties/suggestion?

Thank you for your help with our research

Figure 6.4 The follow-up interview sheet

6.2 Results

The evaluation results are presented in six sections. First, similar to the study in Chapter 4, the logfile data are analysed to summarise participants' web navigation activity. Next, the usage of the history tool is analysed. Then how the tool solved the underlying causes of failure are examined. After that, what participants thought about the history tool is described and the diversity of participants is considered. Finally, some other comments and reflections from participants are included to complete the section.

6.2.1 Logfile data

From the web history logfiles, our participants' activity was slightly different from the first use study in Chapter 4. The average number of webpages visited per day per participant reduced almost to a half (see Table 6.1) and the revisiting rate was 8% lower. There were two reasons for this: (1) the user study took place during the summer time when many of our participants went on holiday for several weeks, and (2) most of them used more than one device to access the WWW (see Section 6.1).

Table 6.1 Comparison of web activity with the study in Chapter 4.

| | Study in Chapter 4 | This study |
|---------------------------|---------------------------|----------------------------------|
| Year of study | 2010-2011 | 2012 |
| Tools | Add-on for Firefox 3.x | Add-on for Firefox 3.x and later |
| Number of participants | 12 | 19 |
| Duration (days) | 50-97 | 21 – 112 |
| Recurrence rate | 36% | 28% |
| No. of URL visits per day | 86 | 55 |
| Informational visits | 31% | 35% |

Although the number of pages visited by participants reduced substantially, the percentages of informational vs. navigation webpages that were revisited for each combination of recency and frequency stayed almost the same (see Table 6.2).

Table 6.2 The percentage of informational vs. navigational pages that were revisited for each combination of recency and frequency.

| Page Type | Frequent | | Not Frequent | |
|---------------|----------|------------|--------------|------------|
| | Recent | Not Recent | Recent | Not Recent |
| Informational | 1.2% | 1.6% | 23.5% | 7.8% |
| Navigational | 11.4% | 5.6% | 34.9% | 14.0% |
| Total | 12.6% | 7.2% | 58.4% | 21.8% |

6.2.2 How participants used the history tool

The logfiles (see Section 6.1.3) were analysed to explore how participants used the visualization history tool, what navigation patterns were adopted, and which components and functionality were used most. This information may be useful for future history tools (including improvements for the presented tool).

Navigation on the visualization history tool was divided into sessions. A session started when the tool was opened and finished either when the tool was closed or after a long period of user inactivity (a 25.5 minute timeout was used, as in Section 4.2.1). On average, there were 16 ($SD = 13$) navigation sessions per participant during the course of the study. These sessions were divided into three categories: exploring the tool, revisiting a webpage, and reviewing browsing history (see Table 6.3). Within the first few sessions, participants explored functionality of the tool by clicking on different components without opening many webpages. After exploring the tool, participants used the tool to either revisit a webpage or review their web history. These two types of pattern were differentiated by whether any webpages were opened in that session. It can be argued that participants did not open any webpage in that session because they could not find the page they needed. However, from the follow-up interview, participants confirmed that there was no such case or they had already filled in the diary form.

Table 6.3 The percentage of exploring, revisiting and reviewing sessions.

| Session Types | No. of Sessions | % Sessions |
|---------------|-----------------|------------|
| Exploring | 65 | 22% |
| Revisiting | 145 | 48% |
| Reviewing | 92 | 30% |

There was no particular pattern of interaction when participants explored the tool. To review their history, participants opened the history tool to see their

history on that day, navigated from date to date, clicked on different domains, switched to Google searches tab then selected search queries, searched to view webpages on a topic, went through months, selected a date then clicked on domains, and occasionally combined different methods. Table 6.4 shows how often these patterns were employed.

A participant who mostly used the tool for reviewing wrote: *“I have looked at the visualization, which I find useful for getting an overview of my browsing behaviour. The tree of sites visited, for example, is a nice visualization which tells me whether I was in a site that had a poor structure, as I can see a wide tree of links I followed. Seeing the tree and clicking on one of the nodes to see a larger snapshot made me remember that I was looking for a specific piece of information on a site. I followed several links, but in the end couldn't find the information on that site. Other trees are more like linked lists: they are very deep but don't branch. Inspecting one such tree made me remember I was just browsing a new site which had lots of interesting content, so I kept browsing on that site”*. Another participant commented: *“The tool made me aware of search habits and memory issues. Instead of typing in keywords, I often type in full questions. Also I sometimes look for the same things a month later without fully remembering that I have asked that question before”*. These were examples of how the tool was used for the reviewing purpose. Sometimes people simply wanted to review their activity rather than revisiting a particular webpage. The visualization history tool is assumed to encourage this new type of usage.

Table 6.4 Percentage of sessions that employed for each reviewing pattern.

| Reviewing Pattern | No. of Reviewing Sessions | % Reviewing Sessions |
|--|----------------------------------|-----------------------------|
| Reviewing history of the current date | 28 | 30% |
| Reviewing history of different dates | 19 | 21% |
| Reviewing history of different domains | 11 | 12% |
| Reviewing different searches | 8 | 9% |
| Reviewing webpage on different topics | 8 | 9% |
| Reviewing different months | 5 | 5% |
| Reviewing different domains on a specific date | 5 | 5% |
| Other | 8 | 9% |

Similarly, several patterns were adopted for revisiting (see Table 6.5). The most common pattern was that participants selected a domain in the domain list and then picked a desired webpage. Another common pattern was when participants opened the tool just to select a webpage either in the list or tree

view of the default result view (history of that current date). These revisits were straightforward because participants remembered that they visited those webpages earlier on that day. Searching for a webpage on a specific topic was utilised fairly often, as was participants remembering the approximate date they last visited a target page and clicking on several dates to see if they could find it. Other patterns were selecting a date and then a domain, selecting a search query in the search queries tab then choosing a webpage. Occasionally, participants employed other navigation such as switching to the search queries tab to browse by date, or using the session tab. On average, it took participants 55 seconds ($SD = 57$) to revisit a webpage. The revisiting time of a webpage was measured from the point the visualization tool was activated to the point the webpage was open in the [web browser](#). Overall, if participants selected a specific date, then 33% of selected dates were 1 or 2 days before (~recently), 25% were of 3-5 days before, 25% were of 6-8 days before (~a week ago), and the rest were for other dates.

Remarking about how they used the tool, a participant said: *“I did start to change my way of using the tool once I realised that if I clicked on an item in the domain list, then this gave me thumbnails from which I could select a particular one. I found this very useful because I often knew the domain name, but was unsure where the page I wanted lay within the domain. It gave me easy access deep within a domain, without needing to keep lots of bookmarks.”* This remark matched with the *Know website* cause of failure reported in Chapter 4.

Table 6.5 Percentage of sessions that employed each revisiting pattern.

| Revisiting Patterns | No. of Revisiting Sessions | % Revisiting Sessions | Average revisiting time (seconds) |
|--|----------------------------|-----------------------|-----------------------------------|
| Open → Select a Domain → Click | 35 | 24% | 46 |
| Open → Click | 32 | 22% | 34 |
| Open → Search → Click | 27 | 19% | 58 |
| Open → Select a Date → Click | 23 | 16% | 67 |
| Open → Select a Date → Select a Domain → Click | 16 | 11% | 76 |
| Open → Select Search Tab → Select a Search → Click | 7 | 5% | 85 |
| Other | 5 | 3% | 89 |

In terms of the tool’s views, participants tended to primarily use either the tree view (42%) or the list view (47%) to open webpages. One participant

explained: “*I prefer the list view because its thumbnails were often much bigger than ones in tree. I just needed to scroll the list to find my wanted pages*”. In the tree view, Pan and Zoom were main interaction (97% of all actions in the tree). Other functions were rarely used: Right clicking to open the overview dialog (2%) and Collapsing/Expanding a branch (1%).

The *Toolbar* was rarely used. Features such as filters, viewing all webpages of a month, switching between Visualization and List mode, back/forward, and home were used in less than 2% of the sessions.

6.2.3 How the tool solved the underlying causes of failure

This section examines how the tool solved the underlying causes of failure when participants employed it to revisit webpages. The analysis was mainly based on diary entries. Analysis of the logfile data indicated that, on the occasions of diary entries, participants chose to use the tool and did not try to revisit the page in the same way as they previously visited it.

The content analysis method (Berelson, 1952; Krippendorff, 1980; Weber, 1990; Lazar et al., 2010) from qualitative research was utilised because data which described revisiting occasions came from free-text fields in the diary form. This analysis involved two steps. First, diary entries were classified into potential coding categories. The *a priori coding* approach (Weber, 1990) was employed rather than the *emergent encoding* approach (Haney et al., 1998) because the underlying causes of failure had been identified in Chapter 4. When diary entries could not be classified into any causes of failure, new categories were created. *Data source triangulation* was used to help ensure high-quality analysis (Erlandson et al., 1993), by checking the diary entries against the logfile data for the tool usage and a participant’s everyday navigation. Secondly, as recommended by Weber (1990), to make valid inferences from the diary entries in the first step, both stability (*intra-coder reliability*) and reproducibility (*inter-coder reliability*) were checked. Regarding stability, the author repeated the first step after one week. To check the reproducibility, another coder (the supervisor of the author) independently classified again these diary entries into the underlying causes of failure. Then the reliability was measured through the Cohen’s Kappa coefficient (Cohen, 1960).

Although 224 webpages were opened with the tool, only 143 diary entries were filled in. 32 of these 143 entries were made after webpages had been revisited using purely Firefox. These entries were analysed to investigate if there were any new difficulty that the user study described in Chapter 4

might have missed, but no new issues were found. The rest of the entries (111 occasions) reported revisiting cases with the new history tool. On average, each participant made 6 entries for tool revisiting ($SD=8$). The follow-up interview revealed that the main reasons for not filling the diary form were: busy, too trivial to report, and were exploring the tool. Participants were unsuccessful on four occasions, indicating that the overall effectiveness of the tool was 96%.

Diary entries of the 107 successful cases were then analysed to be classified into which type of difficulties participants encountered and why they needed the history tool. Data for this analysis mainly came from fields 1, 4, 5, 7, and 8 of the diary form in Figure 6.3. The validity of each diary entry was checked against the logfile of (1) everyday web navigation to confirm data input in fields 1 and 4, and (2) tool usage to verify data input in field 5 of the diary form. There were two reasons for these checks. First, sometimes participants were in the middle of something important or busy so they filled the diary form after a time delay. The second reason was that participants were not always sure about how they previously visited a webpage to fill in the field 4 (i.e., they typed there “not sure” or “search possibly”). All the successful diary entries were consistent with the logfile, except for one case which is discussed later. Diary entries were then classified into causes of failure based on the description of each cause in Chapter 4. For example, diary entries were classified into the *Search results* cause when participants mentioned in the field 4 that they previously visit a webpage through a search engine (e.g., “Google search”, “Search from Google”, or “much searching”), and then stated in either the field 5, 7, or 8 that they preferred to use the history tool rather than search again with a search engine (e.g., “I preferred to use the history tool rather than rerun my Google search”, “useful tool for getting to a site that I had had difficulty searching before”, or “I couldn’t search from web browser’s bar and Google returned so many results that I couldn’t find which exactly I need. So I used the history tool”).

The stability of the author between two times of coding was 94%. The second coder agreed with the author on 75% cases. The Cohen’s Kappa was 0.61 indicating satisfactory reliability (threshold = 0.6). There were two main reasons for differences in coding. First, the boundary between the *Topic* and *Search results* categories was somewhat grey because when participants read on a special topic, they could also access information via any of the three navigation mechanisms (see Section 2.2) including search.

Second, a similar problem happened with the *Known website* and *Search on specific website* categories, but these cases could be distinguished by analysing the whole logfile of each participant. The final classification of diary entries is presented in Table 6.6.

Table 6.6 Classification of diary entries into the underlying causes of failure

| Cause | No. of diary entries | % of diary entries |
|---|----------------------|--------------------|
| Topic | 10 | 9% |
| Search results | 55 | 51% |
| Known website | 4 | 4% |
| Deleted link | 3 | 3% |
| Hidden information | 0 | 0% |
| Search on specific website | 4 | 4% |
| Inappropriate page title | 1 | 1% |
| Links from email, forum & social networks | 5 | 5% |
| Multi-page thread | 0 | 0% |
| Direct entry | 21 | 19% |
| No information about a webpage but its appearance | 3 | 3% |
| Retrieval of the old version of a webpage | 1 | 1% |

Three new categories appear at the end of the table. In a substantial number of cases (19%), participants stated that they visited webpages previously with the direct entry method (e.g., typing URL or accessing bookmarks), but since they had the history tool, they preferred to use the tool to revisit those webpages. They knew those webpages well and had no difficulty revisiting them, but the history tool made revisiting much easier. Overall, in 29 out of 107 diary entries (27%) participants explicitly stated that they chose to use the tool to visit a webpage because they thought it would be easier than using other methods. There were three occasions where participants had no idea about the webpage they wanted to revisit, apart from what the webpage looked like (e.g., “I saw X viewing that webpage on my computer and I wanted to find it again”). The only way they could revisit those webpages was to recognise their thumbnails. The last case was when a participant wanted to see the exact content of a webpage on some day in the past. The content of this webpage changed regularly but the thumbnails captured by the tool preserved this information. In short, in more than 100 occasions, the visualization history tool helped participants succeed in revisiting webpages which participants thought it would have been difficult or time-consuming to

re-access without the tool. To this extent, the history tool has solved the identified causes of failure.

For the diary entries where participants stated that they were not successful, further analysis was done to investigate whether the tool could not help or participants had not fully exploited the tool's capabilities. The first failure was when a participant tried to find a webpage within a website he knew well. He noticed that no webpages of that website had been recorded then filled in the diary form and sent an email reporting the problem. It turned out that website used the https protocol which was automatically ignored by the tool during the user study for all pages (see Section 3.1). On another occasion, a participant picked a wrong thumbnail but he reported that he was still able to navigate to the desired page from there. Examining the logfile of the tool usage showed that the participant recalled correctly the domain then he viewed all webpages within that website. The problem was all the webpages of this website had the same template so it was easy to pick a wrong page. The logfile also revealed that this participant never used the zoom function during the study. This failure might have been avoided if the participants had exploited the zoom function. The third case was when a participant tried to find a scientific paper again amongst many other papers on Web of Knowledge he visited before. He remembered arriving at this paper from citations of another paper. Reviewing the participant logfile, no webpage within the Web of Knowledge was found. During the follow-up interview, the participants admitted that he might have looked at the paper on the computer at the university rather than his laptop where the visualization history tool was installed. A more challenging failure was when a participant remembered what she was looking for with Google search but she did not remember the search term because it was a strange word when she was reading something. She tried to filter domains, searched with a keyword, browsed through some dates and opened some webpages but she could not find it.

Successful revisits were also analysed further to measure the tool's effectiveness and efficiency, and classify them into categories according to recency and frequency. To rate the efficiency of the tool, the number of the steps taken by participants for each revisit were examined from the logfile of the tool usage. On average, including the click to open the tool, participants needed to take 3 steps ($SD=1$) to retrieve a desired page. Some participants also commented that: *"I knew I could have found the webpages with Google search again but I preferred the tool because it was much easier and faster"*.

Examination of the web navigation logfiles allowed each of the diary entries to be classified in terms of recency and frequency (see Table 6.7). These percentages were broadly similar to the overall frequency/recency of revisits in Table 6.2, and participants particularly used the visualization history tool to revisit webpages which had been visited neither frequently nor recently.

Table 6.7 Percentage of revisit diary entries that fell into each combination of recency and frequency.

| Frequent | | Not Frequent | |
|----------|------------|--------------|------------|
| Recent | Not Recent | Recent | Not Recent |
| 13% | 10% | 26% | 51% |

In many diary entries participants remarked how the visualization history tool was useful to them and partly how the tool addressed the causes of failure. Some examples were:

About Known website:

“I know there is a website in my country that provides the same service, but I couldn’t google it again. I tried various keywords with Google but all the results were globally famous websites. Then I realised I had this tool and I found this website easily.”

“Quite useful to go directly to a part of a large website (rather than via main page)”

About Search results:

“This search (the search functionality of the visualization history tool) was faster than re-doing the Google search because I remember that it took me quite a while to find what I was looking for.”

“I suspected that a Google search would help me find the page, but I chose the visualization history tool as I couldn’t remember where in the Google results the relevant webpage was - I thought it might be quicker.”

“This was an extremely useful tool for getting to a site that I had had difficulty reaching before. Best use so far!”

“I’d printed the map, logged out, and then found the printout hadn’t scaled right, so I needed to get the page again. Far easier to go to the visualization history tool, selected ‘today’ (maybe 10 pages browsed), and from there in a few seconds I could see from the thumbnails the page I wanted. That was far, far easier than trying to remember which of the pages I’d visited from google search results page was the one I actually wanted.”

About Inappropriate page title:

"I often type some characters in the address bar to see suggestions from the history, however in this case I could remember what the page was about but not what the title was."

About Bookmarks:

"I couldn't remember the name of the page or the google search query that originally found it for me. When I revisited the page with the visualization history tool I realised that I had actually added it to my bookmarks on my previous visit. That would have been quicker had I remembered that I had done that."

About revisiting a webpage on the same day:

"I used the visualization history tool to reach the page. It was easy to find as I have only visited a handful of pages today and was able to see the webpage I wanted from the tree view"

About user recognition:

"In the first place, I googled for online Gantt chart tools, and I reached it from a blog that reviewed such tools. When I needed to revisit it, I couldn't remember the name entirely but I knew if I saw the webpage I'd know it. That's why I used the visualization history tool."

"I closed the tab then realised I wanted to look at the page again. It would have been a nightmare to find the page again by searching, because I'd looked at many images to find one that was suitable for an illustration. The list view made it easy to revisit - just recognise the image I wanted from near the bottom of the list"

6.2.4 What participants thought about the tool

The follow-up semi-structured interview gave insights into how participants used the tool and what they thought about it. Only 5% of the participants read the user manual and 10% said they glimpsed it. Most of them explained they seldom read user manuals and this case was not an exception. However, they all found it easy to use the tool without reading anything. On average, participants rated the ease of use of the tool is 3.8/5.0.

Referring to the manner in which information was encoded, not all participants noticed the different colours of the heat map calendar, different types of titles (bold/normal) of webpages in the list view, and different types of borders (bold/normal) and different sizes of nodes in the tree view (see

Table 6.8). Some participants said “*I did notice those differences but to be honest I didn’t know why*”. However, all participants figured out how the tree view was created based on navigational paths. They were often excited when saying something like “It was interesting. From the tree, I can recall how I reach to a specific webpage.”

Table 6.8 Percentage of participants who noticed each type of information encoded by the tool.

| Information Encoded | % Participants noticed |
|------------------------------|-------------------------------|
| Heat map calendar | 79% |
| Bold title or border of node | 47% |
| Thumbnail size | 53% |

The heat map calendar was useful for those who wanted to have an overview of their browsing in a month. For the revisiting purpose, they often just clicked on a date they had in mind and the heat map was not much helpful. When viewing results either in tree or list views, participants tended to rely only on the thumbnails of webpages and the thumbnails were much bigger than their borders so it was understandable when participants did not notice this encoded information of the border. Five different sizes of thumbnails were designed for nodes at the original size however they were often zoomed out to fit the whole tree in the screen. This zooming made participants difficult to notice the size difference. When zoomed in, not many nodes were displayed in the screen so participants could not notice this encoded information either. And again, participants often tried to recognise the content of the thumbnails rather than comparing sizes of them.

Overall participants were satisfied with the tool and rated it 4.1/5, and 84% of participants wanted to keep using the tool after the evaluation. One participant said: “*I don’t really have any problems with using the tool as it is quite straight forward and the graphic interface is really useful. I don’t use it every day but when I need to look up something in the history I always use your tool.*” The logfile of the tool usage revealed that this participant used the tool twice a week on average.

Participants who did not want to keep using the tool explained that they had been always able to revisit webpages easily either by using search in history or search from scratch strategies (see Section 4.2). One participant said he just used the tool to get an overview of his web navigation behaviour to maybe get some insights into how to improve certain task.

Another complained that the loading time was a bit slow. To investigate this, the performance of the tool was measured on Firefox 16 running on Windows Vista 64-bit, Intel(R) Core(TM)2 Quad CPU Q9550 @ 2.83GHz 2.83GHz, 4.00GB RAM. The participant's logfile data showed that he often visited more than 100 webpages per day, which the measurements showed would take several seconds to load (see Table 6.9).

Table 6.9 Tool loading time for different numbers of pages.

| Number of Pages | Loading time (seconds) |
|-----------------|------------------------|
| 0 - 49 | 1 – 2 |
| 50 - 99 | 2 – 3 |
| 100 - 150 | 3 – 7 |
| 150 - 200 | 7 – 15 |
| 200 - 1000 | 15 – 30 ⁺ |

Another participant reported an incident when he tried to visualize the history of a month and Firefox crashed.

6.2.5 Diversity of participants

This section examines how participants differed in the way they used the tool. As most participants installed the history tool on their work machine, their job was used as a criterion to classify participants into different groups. The 19 participants were divided into five groups of user: academic staff (A), researcher (R), software engineer (S), manager (M), and PhD student (P). Table 6.10 shows the number of participants, the average numbers of diary entries, and types of sessions of each group. There was little difference between groups except that academic staff used the tool for revisiting webpages more than other types of participants.

Table 6.10 Number of participants in each group and average number of diary entries, and usage sessions.

| Category | No. of Participants | Average No. of diary entries | Average No. of Sessions | | |
|-------------------|---------------------|------------------------------|-------------------------|------------|-----------|
| | | | Exploring | Revisiting | Reviewing |
| Academic | 2 | 22 | 4 | 30 | 4 |
| Researcher | 2 | 4 | 2 | 6 | 4 |
| PhD student | 11 | 4 | 4 | 4 | 5 |
| Software Engineer | 3 | 5 | 2 | 5 | 4 |
| Manager | 1 | 4 | 3 | 4 | 4 |

The patterns of how each group used the history tool were also examined, and are presented in Table 6.11. It seemed that each group adopted a different main pattern of using the tool and all groups made use of the calendar (the last two patterns).

Table 6.11 The revisiting patterns used by each group of participants

| Revisiting pattern | Number of sessions | | | | |
|--|--------------------|---|----|---|---|
| | A | R | P | S | M |
| Open → Select a Domain → Click | 28 | 0 | 3 | 3 | 1 |
| Open → Click | 14 | 0 | 17 | 1 | 0 |
| Open → Search → Click | 4 | 5 | 12 | 6 | 0 |
| Open → Select a Date → Click | 10 | 2 | 7 | 3 | 1 |
| Open → Select a Date → Select a Domain → Click | 4 | 4 | 4 | 2 | 2 |

Regarding difficulties when revisiting webpages, the *Search results* cause was the main problem for all groups. Table 6.12 shows the distributions of the underlying causes of failure experienced by each group of participants.

Table 6.12 Causes of failure distribution of each group of participants

| Cause | Number of diary entries | | | | |
|---|-------------------------|---|----|----|---|
| | A | R | P | S | M |
| Topic | 3 | 1 | 5 | 0 | 1 |
| Search results | 16 | 6 | 23 | 10 | 1 |
| Known website | 2 | 0 | 0 | 1 | 1 |
| Deleted link | 1 | 0 | 0 | 2 | 0 |
| Hidden information | 0 | 0 | 0 | 0 | 0 |
| Search on specific website | 2 | 1 | 1 | 0 | 0 |
| Inappropriate page title | 0 | 0 | 0 | 0 | 0 |
| Links from email, forum & social networks | 3 | 0 | 2 | 0 | 0 |
| Multi-page thread | 0 | 0 | 0 | 0 | 0 |
| Direct entry | 16 | 0 | 5 | 0 | 0 |
| No information about a webpage but its appearance | 1 | 0 | 0 | 1 | 1 |
| Retrieval of the old version of a webpage | 0 | 0 | 1 | 0 | 0 |

6.2.6 Other comments and reflections from participants

Occasionally, participants sent emails telling their experience when using the tool. This section summarises comments and reflections by participants from emails, diary entries and the follow-up interview. First some unpredicted usages of tool are described. Then comments for improving the tool are discussed. Next, suggestions for the future work are presented. Finally, excerpts from two participants' reflections are provided.

Participants used the tool for other purposes. For example, one participant mentioned that she often looked up words on an online dictionary, and showing all the webpages within that website enabled her to review all words she had learnt. Thanks to the tool, she also realised her browsing patterns at home and work. As the tool displays navigational path, a participants said: "*I often used the tool to get a shortcut to a webpage that I normally had to browse several steps from its homepage*". Another participant explained: "*The tool is more than just a revisiting tool. It is also the archive of my web history. The other day I needed to claim my conference expenses. The tool helped me get the exact foreign exchange rate in the past*".

Participants also gave useful comments to improve the tool. Firstly, they would have preferred to open the visualization history tool in a new tab rather than in a new window. Secondly, though it was easy to use the tool,

participants had not exploited the tool thoroughly (e.g., did not notice/understand some of the information that was encoded, and did not know about certain functionality). The reason for this under-exploitation was partly the way the user manual distributed to the user. No one wanted to read a long attached PDF file in an email, the quick guide was automatically shown only once when the tool was first installed. Participants prefer small tip balloons next to components explaining what they are; and the quick guide should be displayed like “Tips of the day” whenever the visualization history tool is opened. One of the participants reported that he could not see all the components of the tool when using it on a netbook (he later installed it on his PC), and it was difficult to zoom in/out without a scroll button of a mouse.

Regarding functionality, participants expected something to happen when left-clicking on thumbnails, and the tab view to be updated according to filtering. It would be convenient if double clicking on a node in the tree made it full size. Dynamic ranges for the heat map calendar to reflect better individual’s navigation pattern were also required. Summary the content of a webpage should be provided as the title and description are not always well assigned. A participant confused between the Google search tab and the search box. Another comment was to display “something” if no result is returned.

For the future work, some participants suggested a distributed version for the tool, and other versions for other web browsers (e.g., Safari, Chrome), mobile devices (e.g., smart phones, tablets). Another participant said *“It may be interesting to get more high-level feedback about browsing behaviour and maybe allow the interface to prompt me if it detects that I’m doing the same, or similar thing I did on a previous session.”*

To end this section, the following reflections are quoted from two participants telling their own experience when using the tool.

Participant 1 – a PhD student in Biology:

“I had a really amazing incidence where I could use your tool. I was spending the Saturday night in with a friend and we were watching a movie. We heard about something in the movie that was interesting and we did not know about so we started googling it. On that homepage (A) we found something (B) that was super interesting as well so we searched for 'B' online, too. From B we then came to C and so on and so forth. We were basically rambling on all night. After literally hours of online searching we

asked ourselves 'how did this happen, how/where did we start? '. Obviously I had the answer to that question and opened your tool and showed the visualization screen to my friend. She was so amazed by the fact that we could trace our steps and that she could even see the little thumbnails. She was immediately asking where I got the tool from because she saw how useful and easy to handle it was.

And she was actually right. When I look back at my online searching history (the little calendar at the top left corner) it is like I have written a diary. I can see when I was searching for things for my fiends wedding, when I was watching BBC iPlayer, the Olympics etc. It is like my personal history. It was also 'shocking' to realise how often I looked up words in an online dictionary (I thought being in this country for 3 years would have helped a little with the language). What is even more shocking was to realise how little time I spend on working on my PhD when I am at home. Although your tool would be super useful for me at work, I am sure! So, I actually learned a lot about myself from your tools. Might sound ridiculous but it is true. I can also see your tool as a tool for parents who want to see what their children are up to online all day”.

Participant 2 – a Professor with expertise in visualization:

Robustness: The software worked flawlessly. It “survived” an upgrade to Firefox, and gave no problems at all during the evaluation period. It appears to be a very robust piece of software.

Searching: I found it to be generally useful although these days I have a restricted set of regularly visited sites (e.g., golf club tee booking, online banking, BBC Sport, etc.). These are bookmarked and indeed I have placed some (e.g., e-mail) on the menubar so they are particularly easy to reach. I think if I was still working I would find it more useful – for example, as a quick route to a site that I had found difficulty locating initially.

Audit Trail: I found it fascinating to be able to review where and when I had visited sites over the past month. This was the most useful aspect for me – for example, there was a cluster of activity when I was looking for spare Olympic tickets. If I was writing a monthly diary, as many retired people do, then this would be an extremely valuable resource. For a research academic, I imagine it could be useful as an aide-memoire when preparing reports.

Novelty: I enjoyed some of the novel (to me) features. For example, being able to pick out sites with a long dwell time was very interesting when I reviewed the month's activity.

Different interfaces: I found the domain list useful because in most cases I knew the URL I wanted and this was a way of saving me typing! The tree was nice visually, and I used it quite a lot – but generally by recognising the thumbnails rather than exploiting the history structure (this would probably have been different if I had been working). I did not seem to use the list view so often, perhaps because it needed more screen width than was available and needed me to use the slider bar. However I did find the list view very useful when I reviewed my month's use, as it told me frequency of visit to each site (i.e., my obsessions!).

Improvements: I think I might have explored more of the features if I had been prompted from time to time. When working at home, there is not the “coffee room” interaction that you get at work, so it is easy just to learn the basics and go no further – even if as in this case there is a perfectly good user manual.

Evaluation: The diary form was easy to complete, but it was probably a deterrent to making full use of the tool since I felt some duty to complete it (which took time!). I could not understand why I needed to enter the URL I was re-visiting, since I felt the tool should know that information.

Features not used: I never switched off reporting, nor did I edit my history, as I never seemed to have any reason to do so.

6.3 Discussion

From the comments and logfiles of participants, the evaluation study revealed that besides revisiting webpages, participants also wanted to review their web navigation history (this accounted for almost 30% off all sessions on the history tool). The goals of reviewing sessions were to help people get the overview of their browsing behaviour, to realise their “obsession” (e.g., spending too much time on something), and to recall where and when they looked at something. Reviewing activity might enable people to manage their time better, improve their WWW navigation strategy, or even to support report/diary tasks. In that respect, it is likely that the visualization history tool encouraged participants to perform this type activity. This finding suggests that further assistance for this activity in future history tools would benefit WWW users.

The main ways participants started their revisiting were selecting a domain (24% of occasions) or a date (27%). These patterns provided evidence that participants often roughly remembered the domain or the date of previous visit of a wanted webpage. On the other hand, on less than 20% of occasions people chose search as their revisiting strategy with the history tool. This implies that participants preferred browsing rather than searching with the tool. Therefore, supporting browsing or searching should be taken into account in designing future history tools.

The rare use of the *Toolbar* can be explained by the fact that even without it the history tool was efficient enough for participants to revisit webpages. On average, only 3 mouse clicks were needed to revisit a webpage.

More than 50% of revisits with the tool were neither frequent nor recent. This highlights people's need for revisiting this category of webpage and supports the author's decision to focus on this category in his research.

The distribution of the underlying difficulties when people revisited webpages in this evaluation study broadly aligned with the finding of the user study presented in Chapter 4. As the evaluation study was carried out during the summer, without any undergraduate and master students, the *Topic* cause happened less often.

More than 50% of diary entries fell into the *Search results* cause. This number indicates that participants often did not want to repeat previous searches and confirms the need of supporting refinding as in (Capra, 2006; Elswailer, 2007; Teevan, 2007b). The question is why they employed the *Google Searches tab* only 7 times out of 56 *Search results* occasions (less than 13%). There were several reasons: most of participants did not read the tool's user manual, so they might not have known about this feature; Exploring sessions were often done at the beginning of the user study when there were not much data to recognise how it worked; the "out of sight, out of mind" problem; Participants might have forgotten the existence of the *Google Searches tab* as the *Domain* tab was the default tab. A short demonstration explaining main features for each participant before the user study may have avoided this problem.

Although participants used the tool different ways, an order of magnitude more participants would be needed to rigorously investigate individual differences. Nonetheless, participants were successful with more than 95% of revisiting occasions, and on more than 30% occasions they explicitly stated that they chose the tool over other methods because revisiting

webpages would be easier with the tool. On 20% of occasions, participants had even switched from using direct entry to the tool. These results indicate that such a visualization tool could solve the underlying difficulties when people revisit webpages. These results also encourage the application of visualization on other areas of information retrieval as suggested by Zhang (2007).

6.4 Summary

This chapter reports the results of the user evaluation of the visualization history tool. The goals of this user evaluation were to: (1) explore how participants actually used the tool, (2) investigate whether such a tool solved the underlying causes of failure, as designed, and (3) learn what participants thought about the history tool. An electronic diary methodology was employed for three months with 19 participants. The participants were asked to browse the WWW as usual, and use an electronic diary methodology to record occasions when they revisited webpages both with and without the tool. At the end of the study, a follow-up semi-structured interview was conducted to clarify aspects of the diary entries and to learn what people thought about the tool.

Navigation with the visualization history tool was divided into sessions based on when the tool was opened, closed, and left inactive. On average, there were 16 navigation sessions per participant on the visualization history tool during the course of the study. These sessions were divided into three categories: exploring the tool (22%), revisiting a webpage (48%), and reviewing browsing history (30%). The goals of reviewing sessions were to help people get the overview of their browsing behaviour, to realise their "obsession", and to recall where and when they looked at something. Reviewing activity might enable people to manage their time better, improve their WWW navigation strategy, or even to support report/diary tasks. This finding suggests that further assistance for this activity in future history tools would benefit WWW users.

There was no clear pattern when participants explored the tool. To review their history, three most common patterns were participants 1) opened the history tool to see their history on that day (30%), 2) navigated from date to date (21%), and 3) clicked on different domains (12%). Similarly, there were three main patterns adopted for revisiting: (1) selecting a domain in the domain list to pick a desired webpage (24%), (2) opening the tool just to select a webpage visited earlier on that day (22%), and (3) searching for a

webpage on a specific topic (19%). These patterns provided evidence that participants often roughly remembered the domain or the date of previous visit of a wanted webpage (51%). Less than 20% of occasions, people chose search as their revisiting strategy with the history tool. This number implies that participants preferred browsing rather than searching with the tool. On average, it took participants 55 seconds ($SD = 57$) to revisit a webpage.

Using the diary form, participants reported 111 cases using the visualization history tool to revisit webpages. Participants often used the tool to revisit webpages which had been visited neither frequently nor recently (more than 50% of occasions). They were unsuccessful on four occasions, indicating that the overall effectiveness of the tool was 96%. On average, including the click to open the tool, participants needed to take 3 steps to retrieve a desired page.

The content analysis method was employed to analyse how such a visualization history tool solved the underlying causes of failure when revisiting webpages. On more than 50% of occasions, the tool helped participants deal with the *Search results* cause. This number indicates that participants often did not want to repeat previous searches and confirms the need of supporting refinding.

With the follow-up interview, participants rated the ease of use of the tool is 3.8/5.0, and rated 4.1/5.0 for their satisfaction. 84% of them wanted to keep using the tool after the evaluation. Participants also commented how the visualization history tool was useful to them.

Chapter 7. Conclusions and future work

Addressing the general topic of “keeping found things found”, this research first investigated how people revisited webpages and the difficulties they encountered. Then a new history tool has been designed, developed and evaluated to address these difficulties. This last chapter completes the thesis by concluding the presented research and suggesting some work for the future.

7.1 Conclusions

The overall aim of the research was to design, develop and evaluate a web history tool that helps people revisit webpages more easily. Existing history tools were designed mostly based on users’ revisiting patterns, classification and management of webpages, potentially useful cues, and enhancing current support of web browsers (see Section 2.7). This research adopted a new approach. It proposed a new design based on findings of an investigation into the difficulties that people encountered when revisiting webpages.

An empirical study has been conducted to investigate what difficulties people encounter when they revisit webpages. Participants recorded their web navigation for three months using a Firefox add-on, and then took part in a controlled laboratory experiment to revisit webpages they had visited neither frequently (on only one day) nor recently (1 week or 1 month ago). The participants’ logfiles revealed that almost one fifth of the revisited pages were in this category, and the failure rate of 20% when revisiting them did not differ between pages visited one week vs. one month previously. This failure rate was higher than the one of revisiting webpages with low frequency of visits (Bruce et al., 2004). An explanation is that target pages of the study in this thesis had been neither frequently nor recently visited. The similar failure rates of one week vs. one month supported the categorizing of revisiting based on recency by Mayer (Mayer, 2009). The frustration that participants expressed in this study has also been noted in previous studies (Bruce et al., 2004; Teevan, 2007b).

One of the main contributions of this work is the investigation of the underlying causes of failure when people tried to revisit those webpages. Ten causes were identified by analysing unsuccessful revisiting trials of a

controlled laboratory experiment, data about participants' navigational actions during the experiment, video/audio of participants' thinking aloud and related data from participants' logfiles. The three main causes (accounting for 61% of the failures) were: (1) participants visiting a large number of pages on a particular topic, (2) webpages that had originally been accessed via search results, (3) participants knowing which website contained a page but that website itself being large. The second cause of failure can be explained by challenges of re-finding such as new ranking algorithms or updated databases (Aula et al., 2005; Teevan et al., 2007), recalling search queries, recognising the pages clicked on the results pages and effectively browsing further from those pages (Obendorf et al., 2007).

From the findings of the empirical study, a novel visualization history tool which supports people in revisiting webpages has been designed and developed. This is another of the main contributions of this thesis. Using the automatically recorded web history approach, the tool exploits visualization techniques to supports both browsing and searching mechanisms in revisiting an individual's complete web history. The new tool is designed to address the main causes of failure identified above, and two more minor causes (Deleted link, and Links from email & social networks). Having a similar layout to Microsoft Outlook, which is used by millions of people, the tool has three main components: (1) the *Global Navigation* lets users navigate within their web history by providing a heat map calendar, a tab view with a list of web domains and a list of search queries, and a search box; (2) the *Result View* displays results of every navigation in both a list view and a tree view; and (3) the *Toolbar* allows users to perform actions like going back to the default state (home), going back/forward navigation actions, fitting the tree to the tree view area, and filtering.

One of the main advantages of the visualization history tool is that it provides users with flexible ways of navigating a web history based on how users remember a target webpage. The tool brings together individual ideas that are included in a number of other tools, and provides a novel visualization approach. Users can jump to any date like in Google history, select a domain like in Domain Tree Browser (Gandhi et al., 2000), review a search query like in SearchBar (Morris et al., 2008), and use a search capability like the history list of web browsers. By providing different navigation techniques, the history tool enables users to revisit webpages within their long-term web history. This is one of the main novelties of the tool.

Similar to WebNet (Cockburn and Jones, 1996) and SessionGraph (Mayer and Bederson, 2001), a full list view with detailed information about each page is employed to complement the tree view. The list view adopts the style of Google search results to present a webpage's information and enriches it by adding a small thumbnail like CWH (Won et al., 2009). As thumbnails are the most important cue for user recognition of visited webpages (Kaasten et al., 2002), they are also used to represent webpages in the tree view like PadPrints (Hightower et al., 1998) and Domain Tree Browser (Gandhi et al., 2000). Like WebNet (Cockburn and Jones, 1996) and SessionGraphs (Mayer and Bederson, 2001), the frequency of visits to a page is encoded by node size. Based on dwell time, this design also differentiates informational and navigational webpages using bold or normal border for nodes.

Employing the same approach as Webmap (Dömel, 1995) and PadPrints (Hightower et al., 1998), the list view use a spanning tree to presents a web history. However, the tree in this design is built by reconstructing a user's actual navigational paths (even crossing different tabs in a tabbed browser) rather than based on the visited time of webpages like previous studies (e.g., Webmap (Dömel, 1995), Domain Tree Browser (Gandhi et al., 2000), and SessionGraphs (Mayer and Bederson, 2001)). Edges of the tree are weighted to help ensure that the tree includes the links that the user most often traversed. This makes the design different from previous ones and reflects more precisely users' navigation on the WWW.

As the Back and Forward buttons of web browsers have become essential to users' navigation, the visualization history tool provides them in the *Toolbar*. Besides letting users filter their web history by frequency like Webnet (Cockburn and Jones, 1996), the *Toolbar* also allow filtering by dwell time. Especially, a button is provided to make the tree fit in the tree view area.

The visualization history tool has been evaluated in a three month field study. The results showed that such a visualization history tool enabled users to navigate effectively within their long-term history to find webpages again. This is another contribution of the present research and encourages the application of visualization for supporting revisit. The evaluation showed that participants could use the tool immediately without any difficulty. They used the tool to not only revisit a specific webpage but also review their web history. On average, they rated the ease of use of the tool was 3.8/5.0. Succeeding in 96% of revisiting occasions, participants especially used the tool to revisit webpages which had been visited neither frequently nor recently (more than 50% of occasions). Including the click to open the tool,

participants needed to take 3 steps on average to retrieve a desired page, implying the excellent efficiency of the tool. The most important thing is they were satisfied with the tool and 84% of them wanted to keep using the tool after the evaluation. Based on these results, it is predicted that the tool will be a practical history tool and could be well adopted by a large number of public audience.

The visualization history tool has some limitations. First, if a webpage named A is automatically redirected from another webpage named B, its navigational path is broken. The reason is that the *load* event is not triggered by webpage B so information about it is not recorded to match with the *referrer* attribute of webpage A. Second, the web history of a user is not up-to-date because information about webpages is not written to the database until they are closed. Third, the performance of tool needs improving when dealing with a large number of webpages.

In short, in the realm of webpage revisitation, the findings of our investigation into difficulties of webpage revisiting can have implications for other researchers considering new history tools. Similar methodical investigation could be performed to explore human factors in re-accessing other personal digital information. The success of the new history tool produced by this research should encourage the wider community of PIM and information retrieval to exploit user-centred design method and visualization techniques for the development of more interactive and effective tools.

7.2 Future work

There are several ways in which this research could be continued. In the short term, some modifications would enhance the visualization history tool. In the longer term, the present research suggests other directions for future work.

7.2.1 Enhancing the visualization history tool

The visualization history tool was highly rated by the participants, and we plan to release a public version in the future. For this, the tool could be improved in the following ways.

Simplifying the tool: From the analysis of the logfile of the tool usage in Chapter 6, unimportant and rarely used functionality could be removed (e.g., Back/Forward, Month View) to improve the performance of the tool and save the real estate (e.g., so the tool can be display properly on a small screen like one of a netbook).

Improving the usability of the tool: Based on users' comments and suggestions in Section 6.2.6, small changes should be made. Firstly, the tool will be opened as a new tab of the Firefox browser, and a new slider will be provided for zooming functionality. Secondly, a notification which indicates what the tool is doing or the result of an action will be shown to the users. Instead of using only a PDF user manual file, tip balloons will be displayed to explain functionalities time to time.

As quantified in Section 6.2.4, the visualization history tool was a little slow when creating a tree of more than 150 webpages (more than 7 seconds). The performance of the tool during the evaluation was partly affected by logging functionalities. However, over time viewing all webpages of an "everyday" website (e.g., bbc.co.uk) could involve dealing with thousands of nodes. Three solutions are being considered. The simplest one is to use the list view only when the number of webpages exceeds a certain threshold (e.g., 200 webpages) because in that case nodes in the tree view become too small for user recognition. Another idea is to divide the tree into sub-trees like the pagination technique of the conventional list. The final solution would be optimising the algorithm of creating the tree.

Finally, if time and funding are allowed, versions of the visualization history tool could be developed for other web browsers (e.g., Google Chrome, Safari, and Internet Explorer).

7.2.2 Other directions for future work

There are a number of promising directions for future work suggested by the research presented here.

Better understanding the difficulties that people may encounter when revisiting personal information (including webpages, files, and emails): the findings of the empirical study presented in Chapter 4 were derived from only 12 participants through 3 x 1 hour controlled laboratory session. Other longitudinal studies with more participants may reveal new difficulties. To do this, a diary method is suggested as it is more natural and suitable with the unpredictable manner of revisiting.

Applying the user-centred design method and visualization techniques in information retrieval in general and in PIM in particular: "User-Centred Design (UCD) offers businesses a number of critical advantages. It enables them to develop easy-to-use products, satisfy customers, decrease expenditures on technical support and training, advertise ease-of-use successes, and ultimately increase market share. Despite these advantages,

many organizations do not practice UCD. Instead, technologically savvy developers often assume they understand the needs of common users and that UCD is implicit in their designs. These assumptions often allow the technology itself to guide the development of products. The difficulty of adopting UCD within such environments requires attention.”¹⁸ Research projects should consider this method when designing and developing new interactive tools. Zhang (2007) well discusses the seven benefits of applying visualization to information retrieval ranging from using human perceptual ability, to reducing cognitive workload, and to enhancing new retrieval effectiveness. The success of the history tool presented in this thesis is another example to encourage the application of visualization.

Exploring other possibilities of visualization to support revisiting: a visual approach should exploit all the possibilities that information visualization offers. Animation could be applied to improve the users’ interaction with their web history as Bederson and Boltman (1999) found that “animation improves users’ ability to reconstruct the information space, with no penalty on task performance time.” Other ways of visual presentation (e.g., 3D or map-based) should be also explored.

Using history information to support web navigating: Recommender systems are popular nowadays. Suggesting an individual webpage has been exploited by the function URL auto-completion of web browsers. The idea of integrating search trails into search engine results pages was suggested in a previous study (White and Huang, 2010). The future history tool could track the address bar of the Firefox browser for this recommending purpose. For example, when a user revisits a webpage the tool can detect all previous navigational paths which contain this webpage and display them in a widget at the right side of a browser. This functionality would provide shortcuts for web navigation as commented by a participant (see Section 6.2.6). Similarly, the tool can detect re-finding queries to show their search trails.

From anywhere and by any devices: With the spread of cloud computing and the high speed internet today, a history tool should be a distributed system that let users integrate their web history across all devices (e.g., computer, tablet, and smart phone).

¹⁸ See <http://www-01.ibm.com/software/ucd/ucd.html#whatisucd>

Appendix A: Research ethics approval

Research Support

3 Cavendish Road
University of Leeds
Leeds LS2 9JT

Tel: 0113 343 4873
e-mail: j.m.blaikie@adm.leeds.ac.uk



MEEC Faculty Research Ethics Committee University of Leeds

5 July 2010

Mr Trien Do
VVR Group
School of Computing
University of Leeds

Dear Trien

Title of study: Making it easier for you to revisit webpages
Ethics Reference Number: MEEC 09-031

The above project was reviewed by the MEEC Faculty Research Ethics Committee at its meeting on 2nd July 2010

The following documentation was considered:

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|--|----------------|-------------|
| MEEC 09-031 Ethical_Review_Form.pdf | 1 | 22/06/10 |
| MEEC 09-031 Information_Sheet.pdf | 1 | 22/06/10 |
| MEEC 09-031 Participant_Consent_Form.pdf | 1 | 22/06/10 |
| MEEC 09-031 Software's User manual.pdf | 1 | 22/06/10 |

On the basis of the information provided, the Committee is happy to approve the project subject to the following conditions:

- The Committee requests clarification of the source of funding for the participant payments
- Please specify the exact number of hourly sessions the participants would be expected to participate in on the information sheet.

Please respond by email or letter to these two points for our records.

Yours sincerely

Jennifer Blaikie
Research Ethics Administrator, Research Support
On Behalf of Professor Richard Hall, Chair, MEEC FREC.

Appendix B: Participant consent form

Participant Consent Form

Title of Research Project: **Visually Browsing an Individual's Long-Term Web History**

Name of Researcher: **Trien Van Do**

Initial the box if you agree with the statement to the left

- 1 I confirm that I have read and understand the information sheet dated 15th October explaining the above research project and I have had the opportunity to ask questions about the project.
- 2 I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. *Insert contact number here of lead researcher/member of research team (as appropriate).*
- 3 I understand that my responses will be kept strictly confidential. I give permission for members of the research team to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.
- 4 I agree for the data collected from me to be used in future research
- 5 I agree to take part in the above research project and will inform the principal investigator should my contact details change.

| | | |
|--|-------|-----------|
| _____ | _____ | _____ |
| Name of participant (or legal representative) | Date | Signature |

| | | |
|---------------------------------|-------|-----------|
| _____ | _____ | _____ |
| TRIEN VAN DO Lead researcher | Date | Signature |

To be signed and dated in presence of the participant

Copies:

Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/pre-written script/information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project's main documents which must be kept in a secure location.

Date: _____ Name of Applicant: _____

Appendix C: Participant information sheet for the user study described in Chapter 4

Participant Information Sheet

Research Student: Trien Do (sctvd@leeds.ac.uk)
Supervisor: Roy Ruddle (r.a.ruddle@leeds.ac.uk)

Address: School of Computing
University of Leeds
Leeds, LS2 9JT

Telephone: 0775 9794 788

I am a researcher in the School of Computing at the University of Leeds, focusing on how people find and revisit web pages. This research is subject to ethical guidelines set out by the British Psychological Society. These guidelines include principles such as obtaining your informed consent before research starts, notifying you of your right to withdraw at any time, and protection of your anonymity. This sheet will hopefully provide you with enough information about the study to allow you to make an informed decision about participation. However, if you have any questions or would like to discuss anything with me please let me know.

The purpose of this experiment is to investigate how people find web pages they have previously visited. You will be asked to find 16 target web pages. For each target:

- You will be given a description of the target
- For some targets, you will also be given the anchor text on the hyperlink to the page, the URL of a page you looked at before the visiting the target, or a thumbnail image of the target
- Using any method you like (search, browse, etc.) find the target webpage (you will be stopped if you have not found it after 3 minutes)

The whole experiment should last less than 1 hour.

Your actions will be recorded in a log file and filmed for later analysis. The research may be reported at academic conferences and in academic journals, but you will remain anonymous. No-one should be able to identify you and at no point will your identity be divulged.

Appendix D: Information sheet for the user study described in Chapter 4

Information Sheet

Research title: Making it Easier for you to Revisit Webpages

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

D.1 Project's purpose

It is predicted that an “average” person will look at approximately one million web pages during their lifetime, but finding a particular page again can be very difficult. The overall goal of this research is to develop a tool which makes it much easier for people to revisit webpages. This particular study investigates the navigation involved in revisiting and the cues (e.g., text vs. pictures) that help.

D.2 Why have I been chosen?

For this study, we need 30 participants who use Firefox for the majority of their web browsing.

D.3 Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason.

D.4 What will happen to me if I take part?

You will be receiving a Firefox add-on, a small piece of software which can be integrated with the Firefox web browser. We will explain how to install and use the add-on, and it will capture your web browsing history for two months. The history will be encrypted and stored by us, so that your navigation and revisiting can be analysed, and we will ask you to take part in a small number of sessions where you find web pages that you have previously visited. You will be paid £7/hour for those sessions.

D.5 What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will allow us to identify cues that are useful for revisiting, so we can improve on current web history tools.

D.6 Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential. You will not be able to be identified in any reports or publications.

D.7 What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?

Details of the webpages you visit, and when you visit them, will be recorded. To protect your privacy we will:

- Not record and https:// ("secure") web pages.
- Let you block individual websites and pages you do not want to be recorded.
- Allow you to turn on/off recording at any time, by clicking on a button.
- Allow you to view your web browsing history, and delete pages from it if you wish.
- Encrypt your history before we store it.

D.8 Will I be recorded, and how will the recorded media be used?

The revisiting sessions will be videoed, to help us analyse the data and illustrate conference presentations and lectures. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the original recordings.

D.9 What will happen to the results of the research project?

Results of the research are will be published at conferences (e.g., the annual ACM SIGIR conference). The proceedings are publicly available, but you will not be identified in any report or publication. Your web browsing history may also be used to inform our follow-on research on revisiting.

D.10 Who is organising and funding the research?

This research is a part of my PhD which is funded by the University of Leeds.

D.11 Contact for further information

Research Student

Mr Trien Van Do
School of Computing
University of Leeds
Email: sctvd@leeds.ac.uk
Cell phone: 0775 9794 788

Student's Supervisor

Dr Roy Ruddle
School of Computing
University of Leeds
Email: R.A.Ruddle@leeds.ac.uk
Telephone: 0113 343 1711

If you decide to take part in this user study, you will be given a copy of this information sheet and, a signed consent form to keep.

Thank you very much for taking part in the project.

Appendix E: Participant information sheet for the user study described in Chapter 6

Participant Information Sheet

Research Student: Trien Van Do (sctvd@leeds.ac.uk)
Supervisor: Roy Ruddle (r.a.ruddle@leeds.ac.uk)

Address: School of Computing
University of Leeds
Leeds, LS2 9JT


Telephone: +44 (0) 113 343 5823

I am a researcher in the School of Computing at the University of Leeds, focusing on how people find and revisit web pages. This research is subject to ethical guidelines set out by the British Psychological Society. These guidelines include principles such as obtaining your informed consent before research starts, notifying you of your right to withdraw at any time, and protection of your anonymity. This sheet will hopefully provide you with enough information about the study to allow you to make an informed decision about participation. However, if you have any questions or would like to discuss anything with me please let me know.

The purpose of this experiment is to evaluate how users would use our visualization web history tool which helps people find web pages they have previously visited. You should have been using Firefox Browser as your main web browser. You will:

- Install our Firefox add-on
- Read the user manual attached
- Browse the web as usual for a month.
- After the first week you will be contacted either by email or in person to discuss any problem you might have.
- During the period of the study, you will be asked to fill in diary entries. You can open the diary form (how to open the form is described in the next section) whenever you would like to report a revisiting experience no matter which tool you use (Firefox history mechanisms, our visualization tool, your own method...). Sometimes, the form will automatically pop up when you revisit a webpage which your last visit was a long time ago. Each diary entry will take a few minutes (Please see Figure 1)
- At the end of the study, you will be asked to anonymously send us your data recorded by our add-on which include the diary and the browsing history. Just send your data from another email (or create a new email account) we don't know.

The research may be reported at academic conferences and in academic journals, but you will remain anonymous. No-one should be able to identify you and at no point will your identity be divulged.

To open the Diary dialog, click the button  (See Figure 1) at the bottom right of the browser where the visualization tool icons are displayed. Then please fill in all the fields as detailed as possible. Below is an example

1) What were you trying to revisit?

I wanted to revisit the homepage of a Swiss researcher I met at the ECIR conference

2) When last you visited it?

Other - About a month ago, when I prepared my talk reporting my experiences at the conference.

3) How many times had you previously visited it?

More than twice

4) How did you visit that webpage last time?

I had the conference proceedings and type his name in Google search and clicked on one result.

5) How did you try to revisit it?

Because his name was not easy for me to remember, I could not form the query for Google again. So I decided to use the visualization tool rather than using Google search or Firefox history mechanisms. First I clicked the Google search tabs of the tab view and scanned through it but it was too long and I could not filter it by typing some characters as instructed in the User manual because I could not remember anything about that name. Fortunately I knew visited it within three days before my talk so I clicked the Search tab of those days and I recognised the name from the list of searches on one of those three days.

6) Were you successful?

Yes and the URL is <http://www.hrtabci.net/>

7) What was the difficulty when revisiting this webpage

I knew the webpage belonged to a search session but the keyword was a name which was not easy to remember for me. I could have gone to the website of the conference/looked at the paper proceedings but I was too lazy to do that.

8) Any other comments?

I was very happy with the visualization tool, however I would like a filter for list of domains or searches within a month because they would become very long by time. The filter by typing some characters didn't work for me because I could remember anything about that name.

Diary Entry at: 30 May 2012 17:32:43
Please answer the questions as much as you can. Questions marked with (*) are required

1) What were you trying to revisit? (*)

2) When last you visited it?
 Today Other
 Yesterday Don't remember

3) How many times had you previously visited it?
 Once More than twice
 Twice Don't remember

4) How did you visit that webpage last time? (*)

5) How did you try to revisit it? (*)

6) Were you successful? (*)
What is the URL if the answer is yes?
 No Yes


7) What was the difficulty when revisiting this webpage?

8) Any other comments?

| No | Time | What were you trying to revisit |
|----|----------------------|---|
| 1 | 28 May 2012 16:15:03 | abc |
| 2 | 29 May 2012 10:46:58 | Roy number |
| 3 | 29 May 2012 10:49:06 | I wanted to revisit homepage of a Swiss researcher I met at the ECIR conference |

Add Close


Figure 1 The Diary form

At the end of the study, please click the button  to view the About us dialog (see Figure 2)

About us

Thank you for participating in our research!

Quick guide

| | |
|--------------------|---|
| Zoom In/Out | Mouse wheel |
| Pan | Mouse Drag and Drop |
| Open a webpage | Double click on a node in tree or list |
| Fit tree to Screen |  |
| Other options | Right click on a node in the tree |

Display Path to Data Folder Open User Manual Close

Figure 2 The About us dialog

Click on the button *Display Path to Data Folder*, copy the path and go to that folder. You will see a file named *VVRWHT.sqlite*. Please compress it if necessary then send us through email as an attachment.

Appendix F: Information sheet for the user study described in Chapter 6

Information Sheet

Research title: Visually Browsing an Individual's Long-Term Web History

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

F.1 Project's purpose

It is predicted that an “average” person will look at approximately one million web pages during their lifetime, but finding a particular page again can be very difficult. The overall goal of this research is to develop a tool which makes it much easier for people to revisit webpages. This particular study evaluates how participants would use our visualization web history tool.

F.2 Why have I been chosen?

For this study, we need 30 participants who use Firefox for the majority of their web browsing.

F.3 Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason.

F.4 What will happen to me if I take part?

You will be receiving a Firefox add-on, a small piece of software which can be integrated with the Firefox web browser. We will explain how to install and use the add-on, and it will capture your web browsing history for one month. During this period, you will be required to fill in diary entries. Each entry will take you a few minutes.

F.5 What are the possible benefits of taking part?

The visualization tool would be useful for your everyday web browsing activities. It helps you easily revisit information on the web.

F.6 Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept ***strictly confidential***. You will not be able to be identified in any reports or publications.

F.7 What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?

Your diary entries, details of the web pages you visit, and when you visit them will be recorded. To protect your privacy we will:

- Not record https:// ("secure") web pages.
- Let you block individual websites and pages you do not want to be recorded.
- Allow you to turn on/off recording at any time, by clicking on a button.
- Allow you to view your web browsing history, and delete pages from it if you wish.
- Encrypt your history before we store it.

F.8 What will happen to the results of the research project?

Results of the research are will be published at conferences (e.g., the annual ACM SIGIR, ACM CHI conference). The proceedings are publicly available, but you will not be identified in any report or publication. Your web browsing history may also be used to inform our follow-on research on revisiting.

F.9 Who is organising and funding the research?

This research is a part of my PhD which is funded by the University of Leeds.

F.10 Contact for further information

Research Student

Mr Trien Van Do
School of Computing
University of Leeds
Email: sctvd@leeds.ac.uk
Cell phone: 0775 9794 788

Student's Supervisor

Dr Roy Ruddle
School of Computing
University of Leeds
Email: R.A.Ruddle@leeds.ac.uk
Telephone: 0113 343 1711

If you decide to take part in this user study, you will be given a copy of this information sheet and, a signed consent form to keep.

Thank you very much for taking part in the project

Appendix G: User manual for the logging tool of the user study described in Chapter 4

WebBrowsingHistoryRecorder Add-on's User Manual

This document explains how to install and use the WebBrowsingHistoryRecorder Firefox add-on.

G.1 Requirements

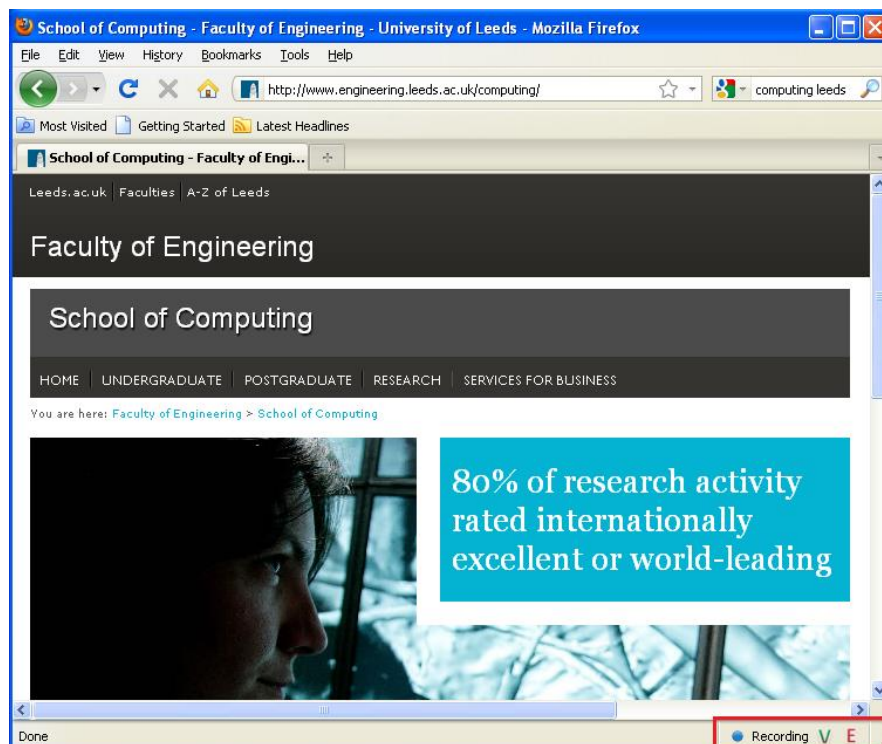
- Firefox browser (on any operating systems (Windows, Linux, Macs))
- 2GB hard disk free (on the drive where Firefox browser is installed)

G.2 Installing the WebBrowsingHistoryRecorder Firefox add-on

WebBrowsingHistoryRecorder Firefox add-on is a small piece of software which can be integrated with the Firefox browser. After being installed, this add-on will record information of webpages loaded by the browser. *Note that, the add-on works only with Firefox browser.*

Each participant will be receiving a file named **WebBrowsingHistoryRecorder.xpi**. To install the add-on, please:

- Launch the Firefox browser; then go to menu File → Open File → Browse to the WebBrowsingHistoryRecorder.xpi. The browser will install the add-on.
- Restart the browser, make sure that the *Status Bar* is visible (Go to menu *View* of the browser → check on *Status Bar*), if some icons are displayed at the bottom right of the status bar like the figure below, the installation is successful.



G.3 Using the Firefox add-on

Stopping/Starting recording web browsing history

To stop/start recording progress, click the button on /off (●/●).

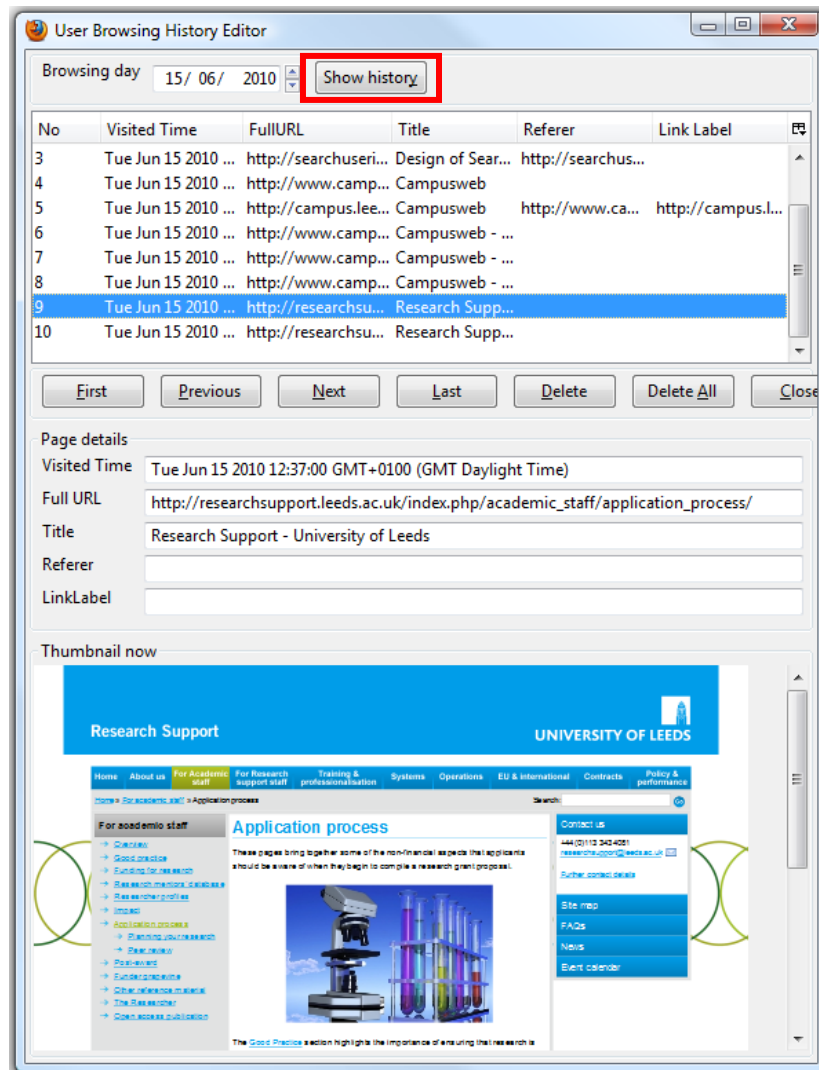
Note that the add-on does not capture https web pages

Where will the log files be stored?

The add-on will create a folder named *logs* on your computer. To view the path to this folder, please click the button **V**.

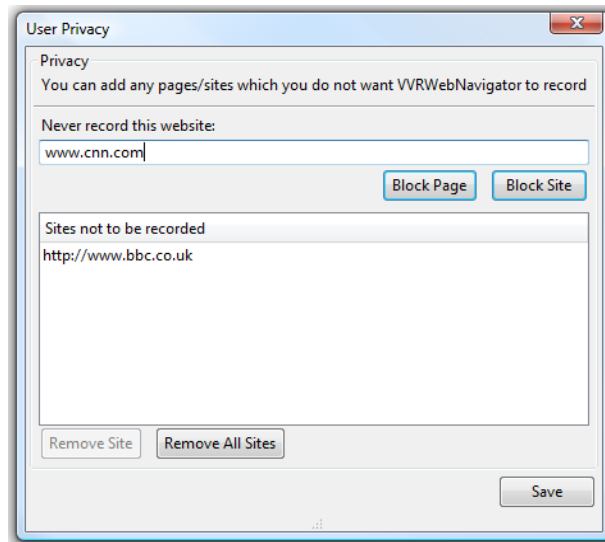
Viewing and deleting web browsing history

To view or delete entries from your web browsing history, click the button **E**. *Note that, if you decide to delete a URL, the tool will search thorough out your history and delete all entries which have the same URL. Visited pages will not be written to logfile until they are closed. Some pages will not be recorded until the browser is closed. In case you cannot find expected URLs on the history dialog, please close the browser and click the button **Show History***



Blocking websites/web pages

If you want to always protect certain websites or web pages from being recorded, go to the menu *Tools* of the browser → *WebBrowsingHistoryRecorder Settings...* A dialog like the figure below will appear to help you do this.



G.4 Contact for further information

Mr Trien Do
School of Computing
University of Leeds
Email: sctvd@leeds.ac.uk
Cell phone: 0775 9794 788

Appendix H: User manual for the logging tool of the user study described in Chapter 6

WebBrowsingHistoryVisualization Add-on's User Manual

This document explains how to install and use the Web Browsing History Visualization Firefox add-on.

H.1 About the Tool

WebBrowsingHistoryVisualization is a visualization tool which captures an individual's web history then visualizes it to support webpage revisiting.

H.2 Requirements

- Firefox browser (on any operating systems (Windows, Linux, Macs))
- 1GB hard disk free (on the drive where Firefox browser is installed)

H.3 Installing the WebBrowsingHistoryVisualization Firefox add-on

WebBrowsingHistoryVisualization Firefox add-on is a small piece of software which can be integrated with the Firefox browser. After being installed, this add-on will record information of webpages loaded by the browser. *Note that, the add-on works only with Firefox browser.*

Each participant will be receiving a file named **WebBrowsingHistoryVisualization.xpi**. To install the add-on, please:

- Launch the Firefox browser. Go to menu *File* → *Open File* → *Browse* to the **WebBrowsingHistoryVisualization.xpi**. The browser will install the add-on. (With some version, go to *Firefox* → *New Tab* → *Open File ...*)
- Restart the browser, make sure that the *Status Bar* is visible (Go to menu *View* of the browser → check on *Status Bar*. With some version, please go to *Firefox* → *Options* → *Add-on Bar* or *View* → *Toolbars* → *Add-on bar*), if some icons are displayed at the bottom right of the status bar like the Figure 1 below, the installation is successful.

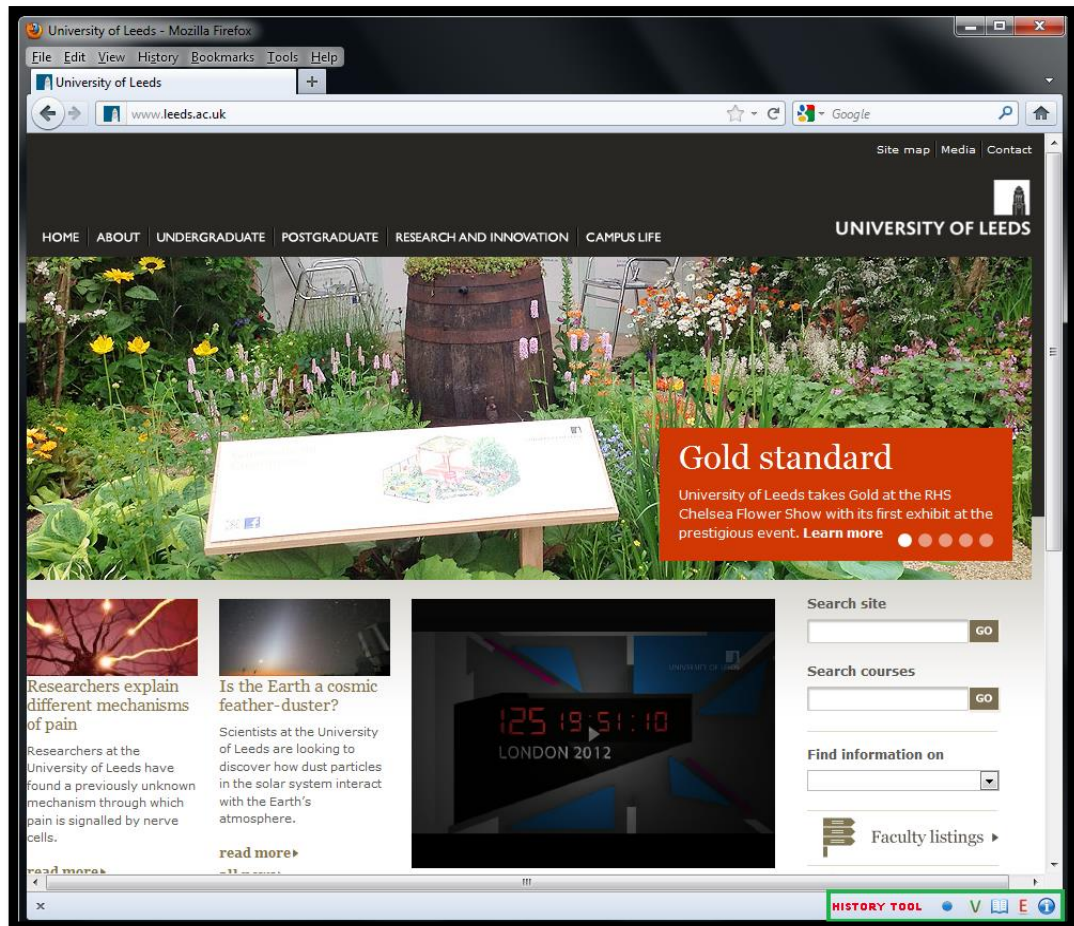


Figure 1 The add-on after successfully installed on Firefox


H.4 Using the Firefox add-on

H.4.1 Stopping/Starting recording web browsing history


To stop/start recording progress, click the button on/off (● / ●).

Note that the add-on does not capture https web pages

H.4.2 Where will the log files be stored?

The add-on will create a *file* named VVRWHT.sqlite to store information about visited webpages and a folder named logs storing webpages' thumbnails on your computer. To view the path to this folder, please click the button . *You will be asked to send us the VVRWHT.sqlite file at the end of the study.*

H.4.3 Viewing and deleting web browsing history

To view or delete entries from your web browsing history, click the button . *Note that, if you decide to delete a URL, the tool will search thorough out your history and delete all entries which have the same URL. Information about visited pages will not be written to logfiles until they are closed. Some pages will not be recorded until the browser is closed. In case you cannot find expected URLs on the history dialog, please close the browser and click the button **Show History** (see Figure 2)*

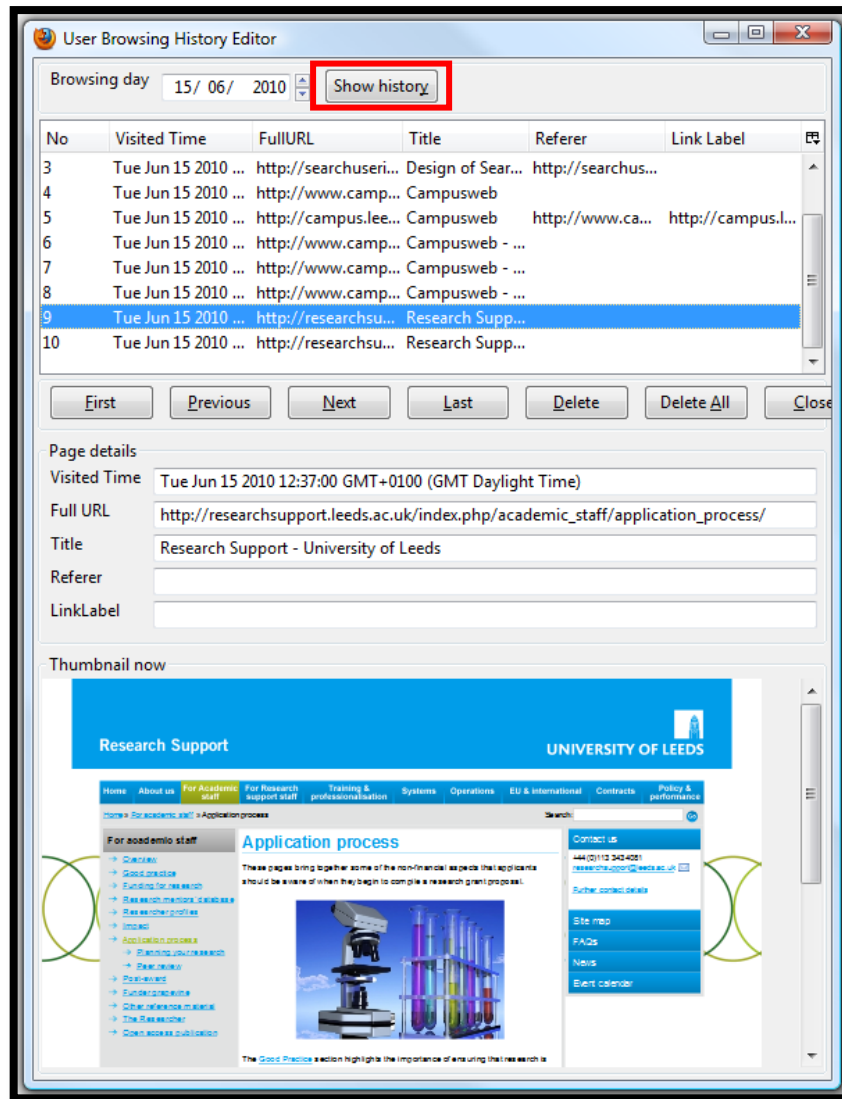


Figure 2 The History Editor Dialog

H.4.4 Blocking websites/web pages

If you want to always protect certain websites or web pages from being recorded, go to the menu *Tools* of the browser → *WebBrowsingHistoryVisualization Settings...* A dialog will appear to help you do this (see Figure 3).

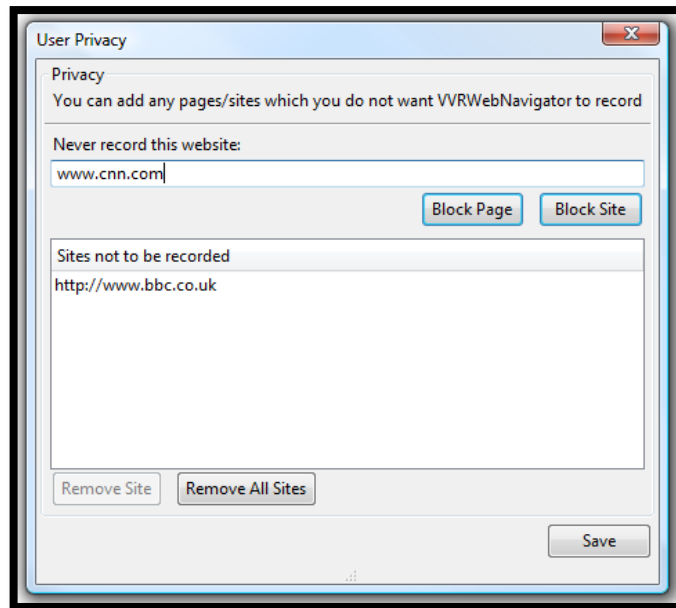



Figure 3 The Privacy Dialog

H.4.5 Using the Visualization Tool

To open the Visualization tool, click the button . The default history of the current date is displayed (see Figure 4).

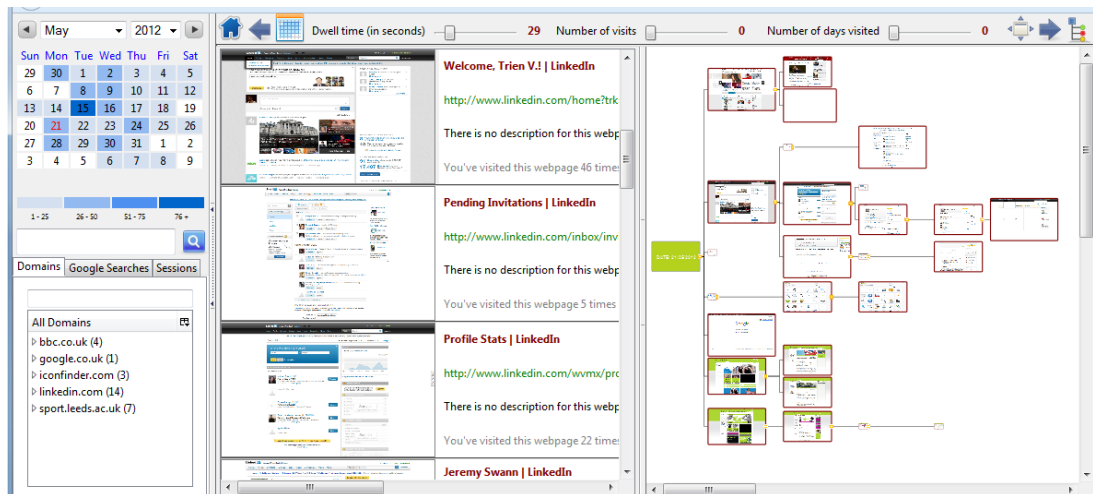


Figure 4 The visualization tool

H.4.5.1 Overview


The tool has three main components:

- Global navigation includes a calendar, a search box, and a tab view. This is the main component which helps you navigate within your web browsing history.
- Result view contains a list view and a tree view. Results of your every navigation step will be presented in both list-based and tree-based presentation. When the mouse is over any node on tree, the corresponding item in the list will be highlighted by a blue bar.

- Tool bar with buttons and sliders allows user to perform quick actions like going back to the default state (home), back, forward, fit to screen and filter functions.

H.4.5.2 How to start your revisitation

There are a few ways to start your revisitation depending on how you last visited the webpage, your memory about the page, and your navigation habit. These are some suggestions.

- If you *remember the date* you last visited the webpage, go to that date using the calendar. The blue background on each day represents the number of webpages/searches you visit/launch on that day. The darker the background the more webpages/searches you have visited/launched.
- In case you *know the domain* of the webpage, you can select (If the domain list is too long you can filter it by typing some characters in the textbox above it) the domain from the tab domain of the tab view. All webpages of the selected domain which have been visited will be displayed in the result view. By default, this list contains all domains you have been to. If you select a specific date on the calendar, only domains visited on that date will be shown. You can always go *back to the full list* of domains by clicking “All Domains” on top of the list or click button Home .
- When you remember the webpage belonged to a *search session* using Google search, you can start with the Google Searches tab of the tab view. This tab is similar to the domain tab. When you click on a search term, the whole search session will be displayed on the result view which includes the Google result list and clicked results then further browsing from them. If you believe that the webpage belonged to a search but cannot find it from the result, right click on any page on the tree and select option *View all webpages on the same day/session* from the dialog.
- Each result set is always divided into sessions based on the 25.5 minute pause of browsing. By selecting the Session tab you choose to view only webpages belonging to a certain session. *Note that the session tab displays sessions of the previous navigation action. For example, if you select a date on the calendar, all sessions on that date will be shown; if you select a domain or a search, only sessions belong to that domain/search will be listed.*
- If you cannot remember anything, just type what you would like to revisit again in the *search box*.

H.4.5.3 How to exploit the result view area

Any of your interaction with the visualization tool will be reflexed on the result view area. The list view and the tree view encode some information to present webpages.

H.4.5.3.1 List view

- The list view adopts the style of Google search results to present a webpage's title, URL, description, frequency (“You've visited this webpage X times”) and recency (“Last visited ...”).



- A basic listing is enriched by adding a **small thumbnail** which conveys the layout of a webpage for better user recognition.
- In search engine results pages, colour-coding is used to distinguish visited webpages from others. With a history tool it is not necessary anymore, so the same colour **blue** is used for all titles but **bold** for pages viewed more than 30 seconds and **normal** for the rest.
- One page might be visited several times, but will be displayed only once in the list. Because each page can be displayed only one, the dwell time will be of the longest time

H.4.5.3.2 Tree view

- The tree view is constructed based on user **navigational paths** (the sequences/branches of links clicked by a user). If a node has been reached by different paths, the shortest path will be presented in the tree.
- Each node of the tree is represented by a webpage thumbnail and, using the same colours as list-based presentation, the borders are in **bold or normal blue**.
- Size of each node tells the number of visits to that webpage. The default size of a node is 151 x width (calculated based on screen resolution). If the frequency of visit to a node is less than 6, its width and height are scaled by $(1 + (\text{Frequency}-1)/10)$, otherwise its size is the same as node with frequency of 5. The root node always has size of page with frequency = 1.

H.4.5.3.3 Interaction

Users can interact with the result view area with

- Zoom in/out the tree view by mouse wheel.
- Pan the tree by drag and drop operations
- Collapse or expand children of a node  
- Move the mouse over a node in the tree view and see detailed information in the list item highlighted by a blue bar.
- Double click on a thumbnail of a list item or a node in tree to open the webpage in a new tab of Firefox browser.
- Right click on a node to open a dialog with detailed information about the webpage and options (see Figure 5)

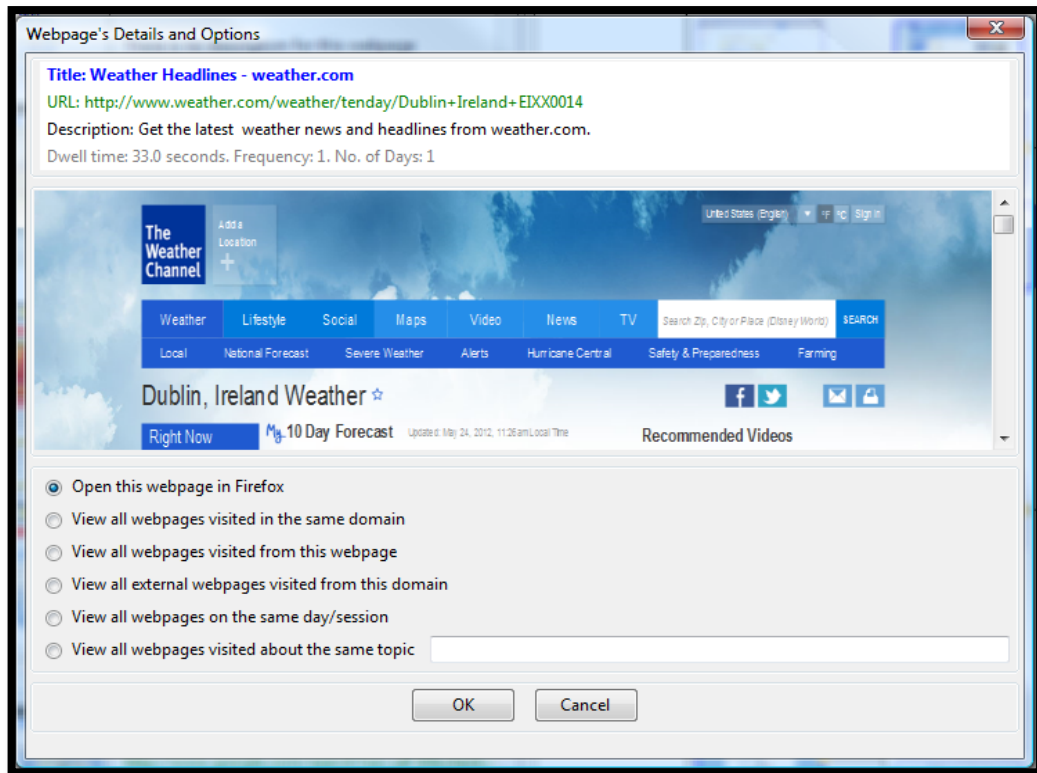






Figure 5 The webpage's details and options dialog



The dialog provides several options

- *View all webpages visited in the same domain* (similar to when you select a domain in the domain tab)
- *View all webpages visited from this webpage.* This function would be useful when links of a certain webpage are changed regularly. For example, you go to the BBC homepage today and click some links there to read latest news. The next day, those links might be no longer there.
- *View all external webpage visited from this domain.* Today, links sent and shared by emails, forums and social networks are very popular. If a user remembers the wanted webpage was shared from a certain domain, this feature would help.
- *View all webpages visited on the same day/session.* This option allows users to explore all webpages that have been visited at the same time with the current webpage.
- *View all webpages visited about the same topic.* A main keyword of the webpage will be extracted and of course users can refine this keyword. All webpages whose title or description contain that keyword will be displayed.

H.4.5.4 The tool bar

The tool bar has 6 buttons for quick operations

- Home : go back to the default state of the visualization tool.
- Back  and Forward : go back or forward to the state of the visualization results.
- Month view : visualize history of the current selected month.

- Fit to Screen : fit the tree to the visible area.
- List/Tree : switch between the List view only and the List with tree modes. In the list view only mode, when the mouse is over a small thumbnail, the full size thumbnail will be displayed.

And 3 sliders for filters:


- Filter by dwell time: when navigation, a user might spend quite much time on some pages and just few seconds on others so dwell time would be a good filter.
- Filter by number of visits: if users can recall how many times they have been to a webpage like one, twice or more, this filter would be useful.
- Filter by number of days visited: a webpage might have been visited several times but only on one or two days. If the users can estimate this number, the result set will be reduced significantly.

Please note that, when filtering, unqualified nodes will be removed from the list view but will be only reduced size in the tree view to reserve contextual information (see Figure 4).

H.5 Diary form

To open the Diary dialog, click the button .

H.6 About us dialog

To view the “About us” dialog, click the button . From there, you can open this user manual document, and view the path to your data file.

H.7 Contact for further information

Mr Trien Do
School of Computing
University of Leeds
Email: sctvd@leeds.ac.uk
Tel: +44 (0) 113 343 5823

Bibliography

- Abrams, D., Baecker, R., et al. (1998). Information archiving with bookmarks: personal Web space construction and organization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Los Angeles, California, United States. ACM, New York, NY, USA, pp. 41-48.
- Adar, E., Teevan, J., et al. (2008). Large scale analysis of web revisitation patterns. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy. ACM, New York, NY, USA, pp. 1197-1206.
- Anisfeld, M. and Knapp, M. E. (1968). Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology*. 77, pp. 171–179.
- Atkinson, R. C. and Shiffrin, R. M. (1968). Human Memory: A proposed system and its control processes. *The Psychology of Learning and Motivation: Advances in Research and Theory*. 2.
- Aula, A., Jhaveri, N., et al. (2005). Information search and re-access strategies of experienced web users. In: Proceedings of the International Conference on World Wide Web, Chiba, Japan. ACM, New York, NY, USA, pp. 583-592.
- Aula, A., Khan, R. M., et al. (2010). A comparison of visual and textual page previews in judging the helpfulness of web pages. Proceedings of the International Conference on World Wide Web. Raleigh, North Carolina, USA, ACM: 51-60.
- Ayers, E. Z. and Stasko, J. T. (1995). Using graphic history in browsing the world wide web. In: Proceedings of the International Conference on World Wide Web, Boston, Massachusetts, USA. ACM, Newyork, NY, pp. 11-14.
- Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and visual similarity. *Quarterly Journal of Experimental Psychology*. 18, pp. 362–365.
- Baddeley, A. D. and Hitch, G. J. (1974). *Working Memory*. Academic Press.
- Badesh, H. and Blustein, J. (2012). VDMs for finding and re-finding web search results. In: Proceedings of the 2012 iConference, Toronto, Ontario, Canada. ACM, New York, NY, USA, pp. 419-420.

- Baeza-Yates, R. and Castillo, C. (2001). Relating web structure and user search behavior. In: Proceedings of the International Conference on World Wide Web.
- Bahrack, H. P., Clark, S., et al. (1967). Generalization gradients as indicants of learning and retention of a recognition task. *Journal of Experimental Psychology*. 75, pp. 464–471.
- Barreau, D. and Nardi, B. A. (1995). Finding and reminding: file organization from the desktop. *SIGCHI Bulletin*. 27(3), pp. 39-43.
- Bates, M. J. (1989). The design of browsing and berrypicking techniques for the on-line search interface". *Online Review*. 13(5), pp. 407-431.
- Bederson, B. B. and Boltman, A. (1999). Does animation help users build mental maps of spatial information? In: Proceedings of Information Visualization Symposium, San Francisco Airport Hyatt, San Francisco, California, USA. IEEE Computer Society Washington, DC, USA, pp. 28-35.
- Bederson, B. B. and Hollan, J. D. (1994). Pad++: a zooming graphical interface for exploring alternate interface physics. In: Proceedings of the ACM Symposium on User Interface Software and Technology, Marina del Rey, California, United States. ACM, New York, NY, USA, pp. 17-26.
- Beitzel, S. M., Jensen, E. C., et al. (2004). Hourly analysis of a very large topically categorized web query log. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom. ACM, pp. 321-328.
- Belkin, N. J., Marchetti, P. G., et al. (1993). Braque: design of an interface to support user interaction in information retrieval. *Information Processing and Management*. 29(3), pp. 325-344.
- Bellotti, V., Ducheneaut, N., et al. (2005). Quality versus quantity: e-mail-centric task management and its relation with overload. *Human-Computer Interaction*. 20(1), pp. 89-138.
- Berelson, B. (1952). *Content Analysis in Communication Research*. Free Press.
- Bergman, O., Tucker, S., et al. (2009). It's not that important: demoting personal information of low subjective importance using GrayArea. In: Proceedings of the SIGCHI Conference on Human Factors in

Computing Systems, Boston, MA, USA. ACM, New York, NY, USA, pp. 269-278.

Boardman, R. and Sasse, M. A. (2004). "Stuff goes into the computer and doesn't come out": a cross-tool study of personal information management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria. ACM, New York, NY, USA, pp. 583-590.

Brewer, W. F. and Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*. 13, pp. 207–230.

Brodie, K., Poon, A., et al. (1993). GRASPARC-A problem solving environment integrating computation and visualization. In: Proceedings of the IEEE Conference on Visualization, San Jose, California, USA. IEEE Computer Society, Washington, DC, USA, pp. 102-109.

Brown, N. R., Rips, L. J., et al. (1985). The subjective dates of natural events in very long term memory. *Cognitive Psychology*. 17, pp. 139–177.

Bruce, H. (2005). Personal, anticipated information need. *Information Research*. 10(3).

Bruce, H., Jones, W., et al. (2004). Keeping and re-finding Information on the Web: What do people do and what do they need? In: Proceedings of the American Society for Information Science and Technology, Providence, Rhode Island, USA. American Society for Information Science and Technology, pp. 129–137.

Bruder, G. and Silverman, W. (1972). Effects of semantic and phonetic similarity on verbal recognition and discrimination. *Journal of Experimental Psychology*. 94(3), pp. 314–320.

Bruza, P. D. and Dennis, S. (1997). Query reformulation on the Internet: empirical data and the hyperindex search engine. In: Proceedings of the RIAO conference, pp. 488-499.

Cacheda, F. and Vina, A. (2001). Experiences retrieving information in the world wide web. In: Proceedings of the IEEE Symposium on Computers and Communications, pp. 72-79.

Cacheda, F. and Vinã, Á. (2001). Understanding how people use search engines: A statistical analysis for e-business. In: Proceedings of the Ee-business and E-work Conference and Exhibition, CheshireHenbury, Macclesfield, UK, pp. 319–325.

- Capra, R. and Pérez-Quñones, M. (2005). Using Web search engines to find and refind Information. *Computer*. 38(10), pp. 36-42.
- Capra, R. G. (2006). An investigation of finding and refinding information on the web. Doctoral Thesis, Virginia Polytechnic Institute and State University.
- Capra, R. G. and Pérez-Quñones, M. A. (2005). Mobile refinding of web information using a voice interface: an exploratory study. In: *Proceedings of the Latin American Conference on Human-computer Interaction, Cuernavaca, Mexico*. ACM, New York, NY, USA, pp. 88-99.
- Card, S. K., Robertson, G. G., et al. (1996). The WebBook and the Web Forager: an information workspace for the World-Wide Web. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada*. ACM, New York, NY, USA, pp. 111-117.
- Carter, S. and Mankoff, J. (2005). When participants do the capturing: the role of media in diary studies. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, Oregon, USA*. ACM, New York, NY, USA, pp. 899-908.
- Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*. 27(6), pp. 1065-1073.
- Ceri, S., Daniel, F., et al. (2006). Extended memory (xMem) of web interactions. In: *Proceedings of the International Conference on Web Engineering, Palo Alto, California, USA*. ACM, New York, NY, USA, pp. 177 - 184.
- Cl, E. (2006). Graphic JavaScript tree with layout. Retrieved 11/11, 2011, from <http://www.codeproject.com/Articles/16192/Graphic-JavaScript-Tree-with-Layout>.
- Cockburn, A. and Jones, S. (1996). Which way now? Analysing and easing inadequacies in WWW navigation. *International Journal of Human-Computer Studies*. 45(1), pp. 105-129.
- Cockburn, A. and McKenzie, B. (2001). What do web users do? An empirical analysis of web use. *International Journal of Human-Computer Studies*. 54(6), pp. 903-922.

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, pp. 37- 46.
- Cohen, J. (2004). *Memory in the Real World*. Psychology Press.
- Cole, J. I., Suman, M., et al. (2003). *The ucla internet report surveying the digital future year three*. UCLA Center for Communication Policy.
- Costa, R. P. and Seco, N. (2008). Hyponymy Extraction and Web Search Behavior Analysis Based on Query Reformulation. *Proceedings of the Ibero-American conference on Artificial Intelligence*. Lisbon, Portugal, Springer-Verlag: 332-341.
- Cramer, P. and Eagle, M. (1972). Relationship between conditions of crs presentation and the category of false recognition errors. *Journal of Experimental Psychology*. 94(1), pp. 1–5.
- Craswell, N., Hawking, D., et al. (2001). Effective site finding using link anchor information. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA. ACM, New York, NY, USA, pp. 250-257.
- Cselle, G., Albrecht, K., et al. (2007). BuzzTrack: topic detection and tracking in email. In: *Proceedings of the International Conference on Intelligent User Interfaces*, Honolulu, Hawaii, USA. ACM New York, NY, USA, pp. 190-197.
- Czerwinski, M. and Horvitz, E. (2002). An investigation of memory for daily computing events. In: *Proceedings of Human-Computer Interaction*, pp. 230–245.
- Dai, N. and Davison, B. D. (2010). Mining anchor text trends for retrieval. In: *Proceedings of the European Conference on Information Retrieval*, Milton Keynes, UK. Springer, Berlin, Heidelberg, pp. 127-139.
- Darken, R. P. and Sibert, J. L. (1996). Navigating large virtual spaces. *International Journal of Human-Computer Interaction*. 8(1), pp. 49-71.
- Dix, A., Finlay, J., et al. (2003). *Human-Computer Interaction*. Pearson, Prentice Hall.
- Dömel, P. (1995). WebMap: a graphical hypertext navigation tool. *Computer Networks and ISDN Systems*. 28(1-2), pp. 85-97.

- Dziadosz, S. and Chandrasekar, R. (2002). Do thumbnail previews help users make better relevance decisions about web search results? In: Proceedings of the International Conference on World Wide Web, Tampere, Finland. ACM, New York, NY, USA, pp. 365-366.
- Eiron, N. and McCurley, K. S. (2003). Analysis of anchor text for web search. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada. ACM, New York, NY, USA, pp. 459-460.
- Eklund, J., Sawers, J., et al. (1999). NESTOR navigator: a tool for the collaborative construction of knowledge through constructive navigation. In: Proceedings of the Australian World Wide Web Conference, Ballina, NSW, Australia, pp. 396–408.
- Elsweiler, D. (2007). Supporting human memory in personal information management. Doctoral thesis, University of Strathclyde.
- Elsweiler, D., Baillie, M., et al. (2011). What makes re-finding information difficult? A study of email re-finding. In: Proceedings of the European Conference on Information Retrieval, Dublin, Ireland. Springer-Verlag, Heidelberg, pp. 568-579.
- Elsweiler, D. and Ruthven, I. (2007). Towards task-based personal information management evaluations. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands. ACM New York, NY, USA, pp. 23-30.
- Erlandson, D. A., Harris, E. L., et al. (1993). Doing Naturalistic Inquiry: A Guide to Methods. Sage.
- Eysenck, M. W. (2001). Principles of Cognitive Psychology. Psychology Press.
- Farah, J. W. T. a. M. J. (1993). Parts and wholes in face recognition. Quarterly Journal of Experimental Psychology. 46A(2), pp. 225–245.
- Feldman, S. (2004). The high cost of not finding information. KMWorld Retrieved 20/11/2012, from <http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=9534>.
- Fisher, D., Brush, A. J., et al. (2006). Revisiting Whittaker & Sidner's "email overload" ten years later. In: Proceedings of the International Conference on Computer Supported Cooperative Work, Banff, Alberta, Canada. ACM, New York, NY, USA, pp. 309-312.

- Fleeson, W. and Kihlstrom, J. F. (1988). Memory for episodic context. In: Annual meeting of the Psychonomic Society.
- Fox, S. (2002). Search engines. The Pew Internet & American Life Project.
- Fox, S., Karnawat, K., et al. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*. 23(2), pp. 147-168.
- Frécon, E. and Smith, G. (1998). WebPath - A Three-Dimensional Web History. *Proceedings of the IEEE Symposium on Information Visualization*. North Carolina, IEEE Computer Society: 3-10.
- Friedman, W. J. (2004). Time in autobiographical memory. Special issue *Autobiographical Memory: Theoretical Applications*. 22(5), pp. 591–605.
- Frost, N. (1972). Encoding and retrieval in visual memory tasks. *Journal of Experimental Psychology*. 95, pp. 317–326.
- Fujii, A. (2008). Modeling anchor text and classifying queries to enhance web document retrieval. In: *Proceedings of the International Conference on World Wide Web*, Beijing, China. ACM, New York, NY, USA, pp. 337-346.
- Gandhi, R., Kumar, G., et al. (2000). Domain name based visualization of web histories in a zoomable user interface. In: *Proceedings of the International Workshop on Database and Expert Systems Applications*, Greenwich, London, United Kingdom. IEEE Computer Society, Washington, DC, USA, pp. 591-598.
- Gillund, G. and Schiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*. 91, pp. 1–67.
- Godden, D. R. and Baddeley, A. D. (1975). Contextdependent memory in two natural environments: On land and underwater. *British Journal of Psychology*. 66(3), pp. 325–331.
- Grossman, L. and Eagle, M. (1970). Synonymity, antonymity, and association in false recognition responses. *Journal of Experimental Psychology*. 83, pp. 244–248.
- Haber, R. N. (1969). Eidetic images. *Scientific American*. 220, pp. 36–44.

- Haney, W., Russell, M., et al. (1998). Drawing on education: Using student drawings to promote middle school improvement. *Schools in the Middle*. 7(3), pp. 38- 43.
- He, D. and Göker, A. (2000). Detecting session boundaries from Web user logs. In: *Proceedings of the Annual Colloquium of IR Research*, Cambridge, UK pp. 57–66.
- He, D., Göker, A., et al. (2002). Combining evidence for automatic web session identification. *Information Processing and Management*. 38(5), pp. 727-742.
- Hearst, M. A. (2009). *Search user interfaces*. Cambridge University Press.
- Hightower, R. R., Ring, L. T., et al. (1998). Graphical multiscale Web histories: a study of padprints. In: *Proceedings of the ACM Conference on Hypertext and Hypermedia*, Pittsburgh, Pennsylvania, United States. ACM, New York, NY, USA, pp. 58 - 65.
- Hoeber, O. and Gorner, J. (2009). BrowseLine: 2D Timeline Visualization of Web Browsing Histories. In: *International Conference Information Visualisation*, pp. 156-161.
- Hoelscher, C. (1998). How Internet experts search for information on the Web. In: *Proceedings of the International Conference on World Wide Web*, Orlando, FL.
- Hong, J. I. and Landay, J. A. (2001). WebQuilt: a framework for capturing and visualizing the web experience. *Proceedings of the International Conference on World Wide Web*. Hong Kong, Hong Kong, ACM, New York, NY, USA: 717-724.
- Huang, J. and Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *Proceedings of the ACM conference on Information and knowledge management*. Hong Kong, China, ACM, New York, NY, USA: 77-86.
- Huttenlocher, J. and Prohaska, V. (1997). *Memory for Everyday and Emotional Events*. Lawrence Erlbaum Associates: 165–179.
- Hyldegård, J. (2006). Using diaries in group based information behavior research: a methodological study. In: *Proceedings of the International Conference on Information Interaction in Context*, Copenhagen, Denmark. ACM, New York, NY, USA, pp. 153-161.

- Jaimés, A., Omura, K., et al. (2004). Memory cues for meeting video retrieval. In: Proceedings of the workshop on Continuous archival and retrieval of personal experiences. ACM Press, New York, NY, USA, pp. 74-85.
- Jansen, B. J., Booth, D. L., et al. (2009). Patterns of query reformulation during Web searching. *American Society for Information Science and Technology*. 60(7), pp. 1358-1371.
- Jansen, B. J. and Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*. 52(3), pp. 235-246.
- Jansen, B. J. and Spink, A. (2003). An analysis of web information seeking and use: documents retrieved versus documents viewed. In: Proceedings of the international conference on Internet computing, pp. 65–69.
- Jansen, B. J. and Spink, A. (2006). How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Information Processing Management*. 42(1), pp. 248-263.
- Jansen, B. J., Spink, A., et al. (1998). Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*. 32(1), pp. 5-17.
- Jansen, B. J., Spink, A., et al. (2007). Defining a session on Web search engines: Research Articles. *Journal of the American Society for Information Science and Technology*. 58(6), pp. 862-871.
- Jiang, J., He, D., et al. (2012). Contextual evaluation of query reformulations in a search session by user simulation. In: Proceedings of the International Conference on Information and Knowledge Management, Maui, Hawaii, USA. ACM, New York, NY, USA, pp. 2635-2638.
- Jiao, B., Yang, L., et al. (2010). Visual summarization of web pages. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland. ACM, New York, NY, USA, pp. 499-506.
- Jin, R., Hauptmann, A. G., et al. (2002). Title language model for information retrieval. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland. ACM, New York, NY, USA, pp. 42-48.
- Jones, R. and Fain, D. C. (2003). Query word deletion prediction. Proceedings of the international ACM SIGIR conference on Research

and development in information retrieval. Toronto, Canada, ACM: 435-436.

Jones, W. (2004). Finders, keepers? The present and future perfect in support of personal information management. *First Monday*. 9(3).

Jones, W. (2008). Personal information management. *Annual Review of Information Science & Technology*. 41(1), pp. 453-504.

Jones, W., Bruce, H., et al. (2001). Keeping found things found on the web. In: *Proceedings of the International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA. ACM, New York, NY, USA, pp. 119 - 126.

Jones, W. and Dumais, S. (1986). The spatial metaphor for user interfaces: experimental tests of reference by location versus name. *ACM Transactions on Office Information Systems*. 4(1), pp. 42-63.

Kaasten, S., Greenberg, S., et al. (2002). How people recognize previously seen Web pages from titles, URLs and thumbnails. In: *Proceedings of BCS Human Computer Interaction*, London, UK. British Computer Society, Swinton, UK, pp. 247-265.

Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, United Kingdom. ACM, New York, NY, USA, pp. 377 - 384.

Kerr, N. H. (1983). The role of vision in "visual imagery" experiments: Evidence from the congenitally blind. *Journal of Experimental Psychology*. 112, pp. 265-277.

Kokosis, P., Krikos, V., et al. (2005). HiBO: a system for automatically organizing bookmarks. In: *Proceedings of the ACM/IEEE-CS joint Conference on Digital Libraries*, Denver, Colorado, USA. ACM, New York, NY, USA, pp. 155-156.

Koolen, M. and Kamps, J. (2010). The importance of anchor text for ad hoc search revisited. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland. ACM, New York, NY, USA, pp. 122-129.

Kosslyn, S. M. (1973). Canning visual images: Some structural implications. *Perception and Psychophysics*. 14, pp. 90-94.

- Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology*. 7, pp. 341–370.
- Kosslyn, S. M. (1976). Can imagery be distinguished from other forms of internal representation? Evidence from studies of information retrieval time. *Memory and Cognition*. 4, pp. 291–297.
- Kosslyn, S. M. (1981). The medium and the message in mental imagery: a theory. *Psychological Review*. 88, pp. 46–66.
- Kosslyn, S. M., Ball, T. M., et al. (1978). Visual images preserve matrix spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*. 4, pp. 47–60.
- Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage.
- Lansdale, M. (1998). The psychology of personal information management. *Applied Ergonomics*. 19(1), pp. 55-66.
- Larsen, S. F., Thompson, C. P., et al. (1996). *Remembering our past: Studies in autobiographical memory*. Cambridge University Press.
- Lazar, J., Feng, J. H., et al. (2010). *Research methods in human-computer interaction*. John Wiley & Sons.
- Li, J. and Zhao, Y. (2009). PathRank: Web page retrieval with navigation path. In: *Proceedings of the European Conference on Information Retrieval, Toulouse, France*. Springer, Berlin, Heidelberg, pp. 350-361.
- Liebscher, P. and Marchionini, G. M. (1988). Browse and analytical search strategies in a full text CD-ROM encyclopedia. *School Library Media Quarterly*. 7, pp. 223-233.
- Liu, C., White, R. W., et al. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland*. ACM, New York, NY, USA, pp. 379-386.
- Loumakis, F., Stumpf, S., et al. (2011). This image smells good: effects of image information scent in search engine results pages. *Proceedings of the International Conference on Information and Knowledge*

Management. Glasgow, Scotland, UK, ACM, New York, NY, USA: 475-484.

Maglio, P. P. and Barrett, R. (1997). On the trail of information searchers. In: The Annual Conference of the Cognitive Science Society, Stanford University, USA. Psychology Press, New York, NY, USA, pp. 466–471.

Malone, T. W. (1983). How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems*. 1(1), pp. 99-112.

Marchionini, G. (1997). *Information seeking in electronic environments*. Cambridge University Press.

Marchionini, G. and Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *Computer*. 21(1), pp. 70-80.

Mayer, M. (2007). *Visualizing web sessions: improving web browser history by a better understanding of web page revisitation and a new session- and task-based visual web history approach*. Hamburg.

Mayer, M. (2009). Web history tools and revisitation support: a survey of existing approaches and directions. *Foundations and Trends in Human-Computer Interaction*. 2(3), pp. 173-278.

Mayer, M. and Bederson, B. B. (2001). *Browsing icons: a task-based approach for a visual Web history*. HCIL Technical Report. MD, USA, University of Maryland.

McBryan, O. A. (1994). GENVL and WWW: Tools for taming the Web. In: *Proceedings of the International Conference on World Wide Web*, Geneva, Switzerland.

Microsoft. (2012). *How to choose between SVG and Canvas (Windows)*. Retrieved 05/12, 2012, from [http://msdn.microsoft.com/en-gb/library/ie/gg193983\(v=vs.85\).aspx](http://msdn.microsoft.com/en-gb/library/ie/gg193983(v=vs.85).aspx).

Morgan, R. and Wilson, M. L. (2010). The Revisit Rack: grouping web search thumbnails for optimal visual recognition. In: *The Annual Meeting of the American Society for Information Science and Technology*, Pittsburgh, Pennsylvania. American Society for Information Science, pp. 1-4.

Morris, D., Morris, M. R., et al. (2008). SearchBar: a search-centric web history for task resumption and information re-finding. In: *Proceedings*

of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy. ACM, New York, NY, USA, pp. 1207-1216.

Nelson, T. O. and Rothbart, B. (1972). Acoustic savings for items forgotten from long-term memory. *Journal of Experimental Psychology*. 93, pp. 357–360.

Nielsen, J. (1995). *Multimedia and hypertext - the internet and beyond*. Boston. Academic Press.

Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.

Norman, D. A. and Draper, S. W. (1986). *User centered system design: New perspectives on Human-Computer Interaction*. Lawrence Erlbaum Associates.

Obendorf, H., Weinreich, H., et al. (2007). Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA. ACM, New York, NY, USA, pp. 597 - 606.

Panasiti, G. (2009). Display your browsing history as a tree: History Tree. Retrieved 20/02, 2010, from <http://www.browserland.com/add-ons/display-your-browsing-history-as-a-tree-history-tree/>.

Park, S., Ho Lee, J., et al. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*. 27(2), pp. 203-221.

PewInternet (2012). What internet users do online. Pew Internet & American Life Project tracking surveys.

Pitkow, J. E. and Kehoe, C. M. (1996). Emerging trends in the WWW user population. *Communications of the ACM*. 39(6), pp. 106-108.

Preece, J., Rogers, Y., et al. (2002). *Interaction design: beyond human-computer interaction*. Wiley.

Ravasio, P., Schär, S. G., et al. (2004). In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction*. 11(2), pp. 156-180.

Robertson, G., Czerwinski, M., et al. (1998). Data mountain: using spatial memory for document management. In: *Proceedings of the ACM*

Symposium on User Interface Software and Technology, San Francisco, California, United States. ACM, New York, NY, USA, pp. 153 - 162.

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior*. 21, pp. 21–38.

Ruddle, R. A. (2009). How do people find information on a familiar website? In: *Proceedings of BCS Human Computer Interaction Cambridge*, United Kingdom. British Computer Society, Swinton, UK, pp. 262-268.

Sellen, A. J., Murphy, R., et al. (2002). How knowledge workers use the web. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Minneapolis, Minnesota, USA. ACM, New York, NY, USA, pp. 227-234.

Silverstein, C., Marais, H., et al. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*. 33(1), pp. 6-12.

Smith, S. M., Glenberg, A. M., et al. (1978). Environmental context and human memory. *Memory & Cognition*. 6, pp. 342–353.

Spence, R. (2007). *Information visualization: design for interaction*. Prentice Hall.

Spink, A., Jansen, B. J., et al. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy*. 10(4).

Spink, A., Jansen, B. J., et al. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*. 35(3), pp. 133–135.

Spink, A., Wolfram, D., et al. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*. 52(3), pp. 226-234.

Staff, C. and Bugeja, I. (2007). Automatic classification of web pages into bookmark categories. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands. ACM, New York, NY, USA, pp. 731-732.

Szóstek, A. M. (2011). 'Dealing with My Emails': Latent user needs in email management. *Computers in Human Behavior*. 27(2), pp. 723-729.

- Tauscher, L. and Greenberg, S. (1997). How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*. 47(1), pp. 97-137.
- Tauscher, L. M. (1996). Evaluating history mechanisms: an empirical study of reuse patterns in World Wide Web navigation. Master thesis, University of Calgary.
- Teevan, J. (2007a). The re:search engine: simultaneous support for finding and re-finding. In: *Proceedings of the ACM Symposium on User Interface Software and Technology*, Newport, Rhode Island, USA. ACM, New York, NY, USA, pp. 23-32.
- Teevan, J. (2007b). Supporting finding and re-finding through personalization. Doctoral thesis, Massachusetts Institute of Technology.
- Teevan, J., Adar, E., et al. (2007). Information re-retrieval: repeat queries in Yahoo's logs. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands. ACM, New York, NY, USA, pp. 151 - 158.
- Teevan, J., Alvarado, C., et al. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria. ACM, New York, NY, USA, pp. 415 - 422.
- Teevan, J., Cutrell, E., et al. (2009). Visual snippets: summarizing web pages for search and revisitation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA, USA. ACM, New York, NY, pp. 2023-2032.
- Thorndyke, P. W. and Stasz, C. (1980). Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology*. 12, pp. 137-175.
- Tulving, E. (1974). Cue-dependent forgetting. *American Scientist*. 62, pp. 74-82.
- Tulving, E. and Psotka, J. (1971). Retroactive Inhibition in Free Recall: Inaccessibility of Information Available in the Memory Store. *Journal of Experimental Psychology*. 87(1), pp. 1-8.
- Tversky, B. (1991). Spatial mental models. *The Psychology of Learning and Motivation*. 27, pp. 109-145.

- Tyler, S. K. and Teevan, J. (2010). Large scale query log analysis of re-finding. In: Proceedings of the ACM International Conference on Web Search and Data Mining, New York, New York, USA. ACM, New York, NY, USA, pp. 191-200.
- Underwood, B. J. and Schulz, R. W. (1960). Meaningfulness and verbal learning. Philadelphia: Lippincott.
- Voorbraak, F. (1991). On the justification of Dempster's rule of combination. *Artificial Intelligence*. 48(1), pp. 171-197.
- Walker, J. Q. (1990). A node-positioning algorithm for general trees. *Software - Practice & Experience*. 20(7), pp. 685-705.
- Weber, R. P. (1990). *Basic Content Analysis*. Sage.
- Weinreich, H., Obendorf, H., et al. (2006). Off the beaten tracks: exploring three aspects of web navigation. In: Proceedings of the International Conference on World Wide Web, Edinburgh, Scotland. ACM, New York, NY, USA, pp. 133 - 142.
- Westerveld, T., Kraaij, W., et al. (2002). Retrieving web pages using content, links, URLs and anchors. In: Proceedings of the Text REtrieval Conference, pp. 663–672.
- Wexelblat, A. and Maes, P. (1999). Footprints: history-rich tools for information foraging. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, Pennsylvania, United States. ACM, New York, NY, USA, pp. 270 - 277.
- White, R. W. and Drucker, S. M. (2007). Investigating behavioral variability in web search. In: Proceedings of the International Conference on World Wide Web, Banff, Alberta, Canada. ACM, New York, NY, USA, pp. 21 - 30.
- White, R. W. and Huang, J. (2010). Assessing the scenic route: measuring the value of search trails in web logs. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland. ACM, New York, NY, USA, pp. 587-594.
- Whittaker, S. (2013). Personal information management: From information consumption to curation. *Annual Review of Information Science and Technology*. 45(1), pp. 1-62.

- Whittaker, S., Bellotti, V., et al. (2006). Email in personal information management. *Communications of the ACM - Personal information management*. 49(1), pp. 68-73.
- Whittaker, S., Bergman, O., et al. (2010). Easy on that trigger dad: a study of long term family photo retrieval. *Personal and Ubiquitous Computing*. 14(1), pp. 31-43.
- Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction*. 8(2), pp. 150-170.
- Whittaker, S. and Sidner, C. (1996). Email overload: exploring personal information management of email. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada*. ACM, New York, NY, USA, pp. 276-283.
- Wiederkehr, B. (2009). 16 JavaScript libraries for visualizations. from <http://datavisualization.ch/tools/13-javascript-libraries-for-visualizations/>.
- Won, S. S., Jin, J., et al. (2009). Contextual web history: using visual and contextual cues to improve web browser history. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA*. ACM, New York, NY, USA, pp. 1457-1466.
- Woodruff, A., Faulring, A., et al. (2001). Using thumbnails to search the Web. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, Washington, USA*. ACM, New York, NY, USA, pp. 198-205.
- Woodruff, A., Rosenholtz, R., et al. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of the American Society for Information Science and Technology*. 53(2), pp. 172-185.
- Zhang, J. (2007). *Visualization for information retrieval*. Springer.