Research Trails: Getting Back Where You Left Off

Jiahui Liu, Peter Jin Hong, Elin Rønby Pedersen
Google, Inc.
Mountain View, CA 94043 USA
{jiahui, peterjinhong, elinp}@google.com

ABSTRACT

In this paper, we present a prototype system that helps users in early-stage web research to create and reestablish context across fragmented work process, without requiring them to explicitly collect and organize the material they visit. The system clusters a user's web history and shows it as research trails. We present two user interaction models with the research trails. The first interaction model is implemented as a standalone application, which presents a hierarchical view of research trails. The second interaction model is integrated with the web browser. It shows the user's research trails as selectable and manipulable visual streams when they open a new tab. Thereby, the NewTab page serves as a springboard in the browser for a user resuming an ongoing task.

Categories and Subject Descriptors: H.5.2 User Interfaces, Interaction Style and Evaluation.

General Terms: Experimentation, Human Factors.

Keywords: Early-stage research, research trail, browser.

1. INTRODUCTION

A recent ethnographic study [3] identified the behavior of early-stage research of normal web users. Different from one-off information gathering activities, which are now supported by many web tools, the early-stage research involves multiple related topics and expands through multiple time sessions. In this paper, we present a prototype of a browsing tool that supports this kind of user activities.

The user study identified a few important attributes of early-stage research. Users usually start with vague and open questions. The research topic may change radically, though gradually as they learn. For example, a user was looking for jobs. While reviewing open job postings she found that she was not immediately qualified for jobs that interested her most. Thus she started noting qualifications that she was missing and searching for classes that she could take, while she continued looking for jobs and submitting applications. The whole process took about a month and spread over short daily work sessions. In the work process, the user spent little effort collecting or organizing the materials.

As reported in the user study, a key challenge for early researchers is to maintain and reestablish the context of their work. Because the target is in flux in early research, people are unlikely to know what information to collect and organize. When they try to resume the research task, they are often at a loss, asking "what was it I looked at yesterday just before the meeting?" The study suggested that people would generally benefit from tools that allow them to browse through previous research sessions. These tools can provide a context for their work and enable users to easily pick up from where they last left off.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

In this paper, we present a prototype system that organizes user's web history into *research trails* [4]. A research trail is a series of work sessions represented as web pages the user visited in the past that belong to the same research task. The system automatically collects the user's web history and clusters them into research trails. It helps early researchers to create and reestablish context across fragmented work process, without requiring them to explicitly collect and organize their web history. A user can at any time look at recent research trails, unpack the relevant one and resume the task.

2. BUILDING RESEARCH TRAILS

The system applies two different perspectives on a user's web history. An *activity-based perspective* focuses on how a user interacts with the web page, i.e. when the page is visited and what other pages are visited together. A *semantic perspective* focuses on the content of the web page, i.e. the topics it might cover. First, the system organizes the web pages into *segments* according to the activity-based perspective. Each segment is a series of web pages visited by the user in a continuous work session. When more than M (e.g. M=5) minutes transpire between two web page visits, a segment boundary is produced. Second, the segments are clustered to form research trails according to semantics of the pages in each segment. A research trail consists of a series of segments that are topically related. The two perspectives complement each other in clustering and presentation. A segment is an intuitive representation of a work session, which helps users to find pages that are visited together. Clustering at the trail level helps users browse through fragmented work sessions related to the same research task.

The system is built upon Google Web History. In the back-end, the system builds an LDA model [1] for each user, thereby identifying the topics that the user is interested in. Using the model, the system generates topic vectors for the web pages and the segments in the user's web history. Then the segments are clustered using the algorithm described in [4]. It is worth noting two important features of the clustering algorithm. First, the algorithm was designed for the provision of topic-sliding, requiring only strong local semantic similarity between consecutive segments. Therefore, the research topic can change slightly overtime in a research trail. Second, the algorithm splits a segment into multiple virtual sub-segments when the topical similarity within the segment is low, and uses only the relevant virtual sub-segments in building a research trail. Therefore, a segment with multi-tasking can belong to multiple research trails.

3. HIERARCHICAL RESEARCH TRAILS

At first we developed a proof of concept prototype, aimed to assess whether people would find the clustering into trails useful. Figure 1 shows a screenshot of this prototype system. Users usually start with the upper left panel, which lists the most recent segments in reverse chronological order. Anchoring the representation in the "now" helps users quickly pick up from last

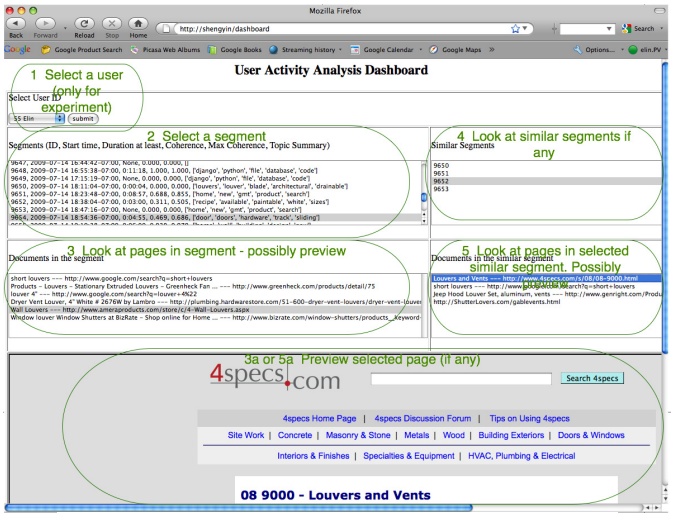

Figure 1. Hierarchical view of research trails

unfinished tasks. Each segment is denoted by the topic keywords extracted from the web pages and the time period the pages are accessed. When users are interested in a particular segment, they simply click on the segment, which loads (a list of) web pages belonging to that segments in the lower panel, At the same time, the research trail in the right panel is a zoom-out view of the segment, showing the overall research task with all the segments pertaining to the research trail. Clicking on a segment in the research trail loads its web pages in the lower right panel. Thus, the lower right panel serves as the zoom-in view of the specific segments, allowing the user to dig into the segment and revisit the pages she worked on before.

4. RESEARCH TRAILS IN BROWSER

The interface shown in Figure 1 provides a hierarchical view of the research trail, which is mostly useful for assessing the quality of the clustering mechanism. To support users in early research, the research trails are better situated in the web browser, the everyday tool they use for conducting their work. Therefore, we propose an interaction design to be triggered whenever the user requests a new tab in the browser. The NewTab page may also be conveniently set as the start up page for the browser. The research trails in NewTab page set up the context for the user to efficiently return and resume an ongoing research task, possibly left off the previous day or just a short while ago when a competing demand temporarily interrupted the user.

Figure 2 shows the layout of the NewTab page, which lists recent research trails of the user. We adopt the GLIDE streams [2] to represent each research trail. A GLIDE stream shows the thumbnails of the pages contained in a research trail, making it easy and efficient to get a sense of the materials in the trail. The research trails are initially shown with their most recent (right most) pages in frontal view and the rest of the trail going off to the left. To the right of the GLIDE stream is meta-information about the research trail, such as the start time, the accumulated duration, essential key words to describe the contents, the number of pages and segments, etc. User can move the research trail to the left to view older page visits. To inspect a particular page, the user can turn over the thumbnail, as shown in Figure 3, while clicking on it opens the page in another tab. Multiple thumbnails can be turned over, while the user scan through the stream of web pages, serving as a pin point for potentially interesting pages.

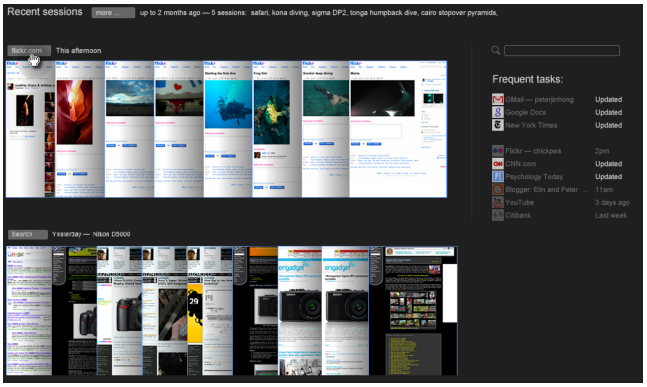

Figure 2. Research trails in NewTab page

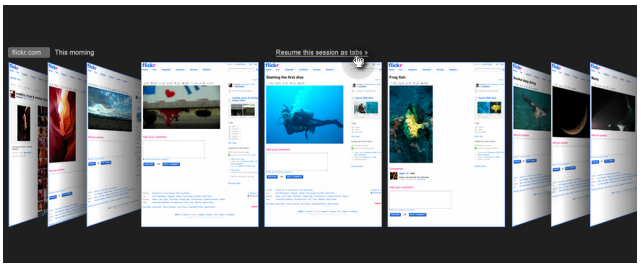

Figure 3. Inspect a research trail

The interaction model in Figure 2 and 3 creates a usable and enjoyable experience for users to browse their web history. It builds an environment that manages user's attention to focus on what matters and minimize distraction, with regard to their tasks of pattern recognition, recall, browsing and comparisons. The pages and segments are presented in a single stream, creating a smooth interaction flow. The thumbnails of web pages help users quickly recognize previously visited pages. In addition, showing multiple pages in the same stream allows users to compare the pages side by side.

5. CONCLUSION

We presented a prototype system that organizes user's web history into research trails and two interfaces to support early-stage research of everyday web users. Future work includes extending the clustering mechanism to make it scalable, experimenting with server and client based data capture, and assessing the user interaction design in real context.

6. REFERENCES

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. In Journal of Machine Learning Research 3.

[2] Hong, P, and Pedersen, E.R. (2009). G.L.I.D.E. Tab Navigation, Entry in the Mozilla design contest (http://design-challenge.mozillalabs.com/summer09/)

[3] Pedersen, E.R. (2010). Territorial marking – and other reasons why a web notebook might not be the killer app. Technical Report. Google Research, 2010.

[4] Pedersen, E.R., Gyllstrom, K., Gu, S., and Hong, P.J. (2010). Automatic Generation of Research Trails in Web History. In Proc. IUI 2010.