Experiences in Building the Public Web History Repository

Eelco Herder
L3S Research Center
Hannover, Germany
herder@L3s.de

Ricardo Kawase
L3S Research Center
Hannover, Germany
kawase@L3s.de

George Papadakis
L3S Research Center
Hannover, Germany
papadakis@L3s.de

EXTENDED ABSTRACT

Learning has become an integral part of many people's everyday working life. A more knowledge-based society and rapid changes in technology requires practically everyone, but in particular knowledge workers, to search for and read information in order to keep up-to-date. The character of learning at the workplace has shifted from a solitary, paper-based activity to a Web-based activity, making use of various resources, including discussion forums and social networking sites [2].

At the same time, we use the Web for our other daily activities. Search engines, travel planners, dictionaries and other online services have become essential for dealing with everyday tasks. News sites, portals, online games and streaming video are popular resources for information and entertainment. We communicate with our friends via email, social networking, forums, blogs and chat. As a result, one ends up with a large collection of scattered digital resources related to both our professional and personal lives.

The activities that people perform in order to acquire, organize, maintain, retrieve and use information items such as Web pages for everyday tasks are commonly refered to as Personal Information Management[1]. Search engines and social bookmarking sites play a major role in finding new information and services. For organizing and refinding these items, Web provide history mechanisms such as url autocompletion, the forward and back buttons, bookmarks and the history sidebar. However, this support is found to be suboptimal and skewed toward a small set of frequently visited resources [6].

For this reason, the analysis and prediction of online browsing behavior and revisitation patterns has received much attention from the research community as well as from industry [1, 8, 5, 3, 7]. Academic research delivered several alternative history mechanisms, including gesture navigation, 'smart' back buttons that recognize waypoints and many types of history visualizations. Browser add-ons that support users in revisiting pages and site include Delicious (social bookmarking), Infoaxe and Hooeey (full-text history search), WebMynd (history sidebar for search) and ThumbStrips (history visualization).

For the improvement and evaluation of personal history management concepts and tools, the availability of suitable data has become increasingly important. Companies like Google, Microsoft and Facebook constantly improve their products, based on the data that they collect from their users. Open-source software developers and researchers normally have no access to this data, which puts them into a disadvantage.

The Web History Repository Project (WHR) aims to leverage this disadvantage by building a public repository of web usage data, which researchers can use to gain new insights in online browsing behavior. Using a Mozilla Firefox add-on, users can upload their anonymized usage data to the server.

Anonymity was a central issue during the conceptualization of the WHR. The uproar around the AOL Query Log Dataset in 2006[2] showed that it is not sufficient to hide the identity of the users: by combining evidence from search queries, it is still possible to trace these queries back to one specific person. As the WHR is meant to be available to the public, all urls and hosts are represented by a GUID. Apart from this factual risk that user identities will be involuntarily become exposed, the rigorous removal of any potential privacy concerns was meant to minimize subjective worries about privacy loss.

The biggest challenge was (and still is) to convince regular users to contribute to the repository. It is unlikely that some user will go to the trouble of installing an add-on and uploading their highly personal data, just to make two researchers happy. For this reason, an explicit effort was made to communicate the importance of the WHR for researchers and open-source developers:

"*Did you ever spend minutes or even hours trying to re-find a specific page? Do you want your Web browser to be smarter than just recommending the last visited pages or just showing you a list that you have to dig through yourself? So do we. And it is easy for you to help if you are a regular user of Firefox: just send us your anonymized Web history*".

[1] http://en.wikipedia.org/wiki/Personal_information_management

[2] http://techcrunch.com/2006/08/07/aol-this-was-a-screw-up/

Extensive information was provided on what information will be sent and how this is encrypted: "*The add-on does not track data - your anonymized data will only be sent once you click on the 'Send now' button yourself*".

Before the launch of the Web History Repository project website[3], several rounds of user testing were held. An extensive review process by Mozilla volunteers provided a guarantee to the user that the add-on[4] is 'safe'. We promoted the Web History Repository through several targeted mailinglists, Facebook, Blogspot and Twitter. After the first disappointing two weeks, during which we continued promoting the project, a critical threshold of 'fans' and 'followers' (including a number of influential people) was reached. Currently, two months after the release of the add-on, more than 200 anonymous volunteers contributed over 1.3 million entries from their browser history; every day 5-10 new contributions are added to the repository.

The data in the Web History Repository includes the list of visited pages, including timestamp and browser session. For each visited page the (encrypted) url and host, the total number of visits, the frecency and the last visit is listed in a separate table[5]. The WHR does not allow for extensive qualitative analysis as a similar dataset that we gathered in 2006 [9], but the far larger quantity of data provides a good basis for analysis of *patterns* in online browsing behavior as well as for evaluation of machine learning algorithms. Such analyses can lead to improved models of online browsing behavior, better page (revisit) recommendations and mechanisms for organizing and refinding information and services found on the Web.

We used the WHR for the evaluation and further improvement of SUPRA[6], a generic library for real-time, contextual prediction of navigational activity that encompasses a set of methods aligned in two tiers [4]. The first tier ranks resources according to their likelihood of being used in the immediate future, as it is derived from their recency and frequency of use. The second tier complements the ranking methods with techniques that identify resources that are commonly visited within the current user context. The contextual prediction library is used as a basis for the PivotBar, a dynamic browser toolbar that recommends visited pages that are relevant to the page currently viewed. Results of a first controlled user study, that a significant amount of revisits has taken place via the PivotBar.

Both the WHR and the contextual prediction framework are available to the community for further experimentation.

REFERENCES

1. E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *CHI*, pages 1197–1206. ACM, 2008.
2. M. A. Chatti and M. Jarke. The future of e-learning: A shift to knowledge networking and social software. *Int. J. Knowledge and Learning*, 3 (4/5), 2007.
3. F. Chierichetti, R. Kumar, and A. Tomkins. Stochastic models for tabbed browsing. In *Proc. WWW 2010*, 2010.
4. R. Kawase, G. Papadakis, and E. Herder. How predictable are you? a comparison of prediction algorithms for web page revisitation. In *ABIS 2010*, 2010.
5. R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proc. WWW 2010*, 2010.
6. H. Obendorf, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *CHI*, pages 597–606. ACM, 2007.
7. A. G. Parameswaran, G. Koutrika, B. Bercovitz, and H. Garcia-Molina. Recsplorer: recommendation algorithms based on precedence mining. In *SIGMOD*, pages 87–98, 2010.
8. S. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *WSDM 2010*, pages 191–200, 2010.
9. H. Weinreich, H. Obendorf, E. Herder, and M. Mayer. Off the beaten tracks: exploring three aspects of web navigation. In *WWW*, pages 133–142. ACM, 2006.

[3] http://webhistoryproject.blogspot.com/
[4] https://addons.mozilla.org/en-US/firefox/addon/226419/
[5] see https://wiki.mozilla.org/Places:Design_Overview
[6] http://sourceforge.net/projects/supraproject/