A Web Browsing Behavior Recording System

Hayato Ohmura, Teruaki Kitasuka, and Masayoshi Aritsugi

Computer Science and Electrical Engineering
Kumamoto University, Kumamoto 860-8555, Japan
{kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

Abstract. In this paper, we introduce a Web browsing behavior recording system for research. Web browsing behavior data can help us to provide sophisticated services for human activities, because the data must indicate characteristics of Web users. We discuss the necessity of the data with potential benefits, and develop a system for collecting the data as an add-on for Firefox. We also report some results of preliminary experiments to test its usefulness in analyses on human activities in this paper.

Keywords: Web browsing, browsing behavior, browser

1 Introduction

There is no doubt that World Wide Web has given tremendous influence over human activities. Many technologies on Web services, e.g., Web search and recommendation, have been developed by both academias and industries. Every Web user usually accesses webpages for getting valuable information for their activities of every kind. Note, however, that needs of Web users can be different from each other.

In order to provide good services to users, many researchers have focused on Web users' behavior. For example, Fu et al. [6] collected user's navigation history and applied data mining techniques to discover hidden information from the history for assisting in surfing the Web. WebQuilt [11] is a Web logging and visualization system for usability analysis on webpages. Shahabi and Chen [16] proposed an adaptive recommendation system using many sources including human experts, web navigation patterns and clusters of user evaluations. Gauch et al. [7] did research for adapting information navigation based on a user profile constructed automatically using visited webpages extracted from user's Web cache. Sugiyama et al. [19] exploited Web browsing histories when constructing user profiles based on collaborative filtering for Web search. Teevan et al. [24] proposed personalized search algorithms using various information such as previously issued queries and previously visited webpages. Shen et al. [17, 18] exploited Web browsing data, namely query history and clickthrough history, as implicit feedback. They used TREC[1] data for evaluation of their proposals. Dou

[1] Text REtrieval Conference, http://trec.nist.gov/

et al. [4] studied personalized search strategies based on query logs of MSN[2]. White and Morris [25] investigated interaction logs of advanced search engine users to better understand how users search. Liu et al. [14] developed a personalized news recommendation system using click behavior on Google News articles. Holub and Bielikova [10] proposed a method for automatic estimation of user's interest in a visited webpage. Guo and Agichtein [8, 9] presented a search behavior model for effective detection of user's search goals using interaction data captured by instrumenting the LibX toolbar[3] for collecting GUI events such as mouse movements, scrolling, and key press events. Aula et al. [2, 1] studied Web search behavior for successful strategies in Web search. Druin et al. [5] experimentally studied children's roles as information seekers with using keyword search interfaces toward new interfaces. We can also find several studies recently using browsing history and behavior (e.g., [3, 15]). To summarize, many studies of Web browsing behavior have attempted to improve Web services. On the other hand, we attempted to exploit Web browsing behavior in spam filtering [22, 21] and tag recommendation to photos in Flickr [23, 20].

In this paper, we develop a system for recording Web browsing behavior for research. In our previous studies [22, 21, 23, 20], we used synthesis data of Web browsing behavior in evaluating our proposals instead of using real Web browsing behavior data because it was difficult for us to collect a large amount of real data. Mainly four schemes have been used for collecting Web browsing behavior. One is to collect and analyze Web logs stored at a Web server [12, 4, 14]. This scheme can only be carried out by an organization having a large amount of server logs. Another is to take videos of Web browsing and interviews to users [5]. It may need to take long time and to pay large cost for this scheme. Another is to analyze Web caches in the local machine [7]. The information collected by this scheme may be restricted because Web caches are designed not for collecting Web browsing behavior data but for performance. The other is to record Web browsing behavior by their own systems [25, 8, 9]. Our proposal described in this paper is categorized in this scheme. We describe the design of our system for collecting Web browsing behavior data in this paper.

The remainder of this paper is organized as follows. Section 2 describes how we designed our system and explains recorded data. Section 3 reports some preliminary experiments to test its usefulness in analyses on human activities, and Section 4 concludes this paper.

2 A Web Browsing Behavior Recording System as an Add-on for Firefox

2.1 Design Overview

As described in the previous section, there have been conventional methods for collecting Web browsing behavior data used in related studies. We observed them, and decided the following principles when developing our system.

[2] MSN Search, `http://search.msn.com/`

[3] Original LibX toolbar, `http://www.libx.org/`

1. Our system collects Web browsing behavior data from a machine on which the browser run instead of a Web server machine, thereby allowing a user of our system to collect the data.
2. Our system collects the data of many kinds, thereby allowing us to analyze Web browsing behavior in detail from a variety of aspects.
3. The data should be collected easily. For this purpose, we make the interaction between a user and our system be as little as possible.

For the first principle, we decided to build our system as an add-on[4] for Firefox[5]. In [11], three logging systems, namely server-side logging, client-side logging, and proxy-based logging systems were discussed and WebQuilt was developed as a proxy-based logging system. However, the proxy may become a bottleneck if the amount of user behavior data is large, and thus we decided to build our system as a client-side logging system. Discussions in Sections 2.2 and 2.3 correspond to the second and third principles, respectively.

2.2 Web Browsing Behavior Data

We chose data to be collected by mainly observing related studies, e.g., [25, 1, 13]. The data are described in the following in the three categories, namely System Information, Browser Situation, and Logged Inputs. The data in Browser Situation and Logged Inputs are recorded with their timestamps. In the current implementation, the data are recorded in XML files in the local machine of the user.

System Information The data in this category give us information about the browsing environment. Concrete data are as follows:

- Operating system
- Display resolution
- Color depth
- Browser version

These data are collected once at the beginning of recording.

Browser Situation

- Browser location
- Size
- Viewport size
- Scrollable capacity
- Scrolled amount
- URL
- Number of tabs

[4] Add-ons for Firefox, `https://addons.mozilla.org/firefox/`

[5] Firefox web browser, `http://www.mozilla.com/`

Browser location is the location of the browser on the display. Viewport size is the area where the loaded webpage is displayed. Browser location, Scrollable capacity, and Scrolled amount will tell us the focused areas of a webpage by the user.

Logged Inputs

- Event type
- Event location
- Mouse button
- Mouse trace
- Wheel
- Typed key
- Shift key
- Ctrl key
- Alt key
- Texts
- Highlighted characters

Event type records one of “init”, “click”, “dblclick”, “mousedown”, “mousup”, “keypress”, “keyup”, “mousescroll”, “mousemove”, “tabopen”, “tabclose”, and “locationchanged”.

2.3 User Settings

Figure 1 shows the interface to begin and end recording. To begin, a user just selects start of our system. There are two ways to start and stop recording, as shown in Fig. 1.

A user can select which data should be recorded by our system as shown in Fig. 2, which shows an example setting where all items are checked to be recorded. We assume our system is used for research only in the current implementation; in other words, a user must check the settings appropriately for the user’s privacy.

3 Preliminary Experiments

In this section, we report some results of preliminary experiments for discussing potential effectiveness of Web browsing behavior data described in the previous section. Since it takes long time to collect a large amount of data, evaluation of our system in terms of concrete services will be included in our future work.

3.1 Mouse Movement

As described in the previous section, our system can record mouse movement. Figure 3 shows mouse trace data on a webpage when a user browsed webpages

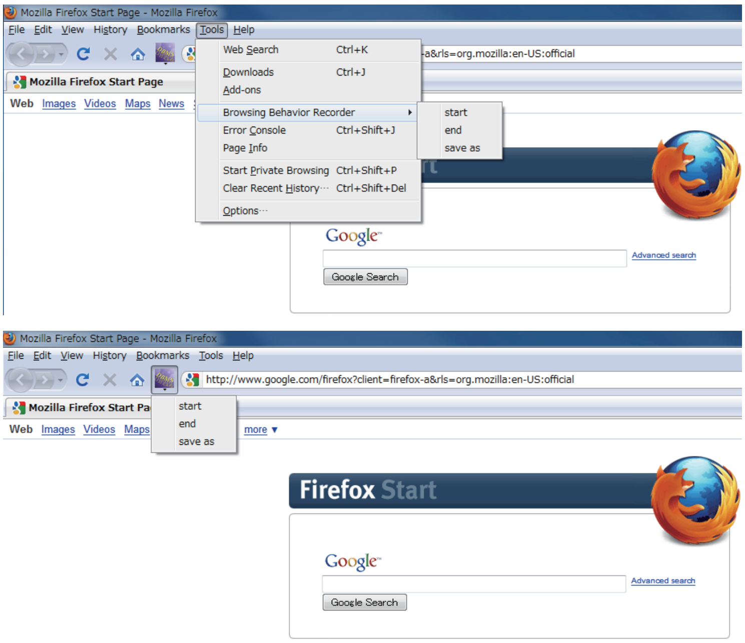
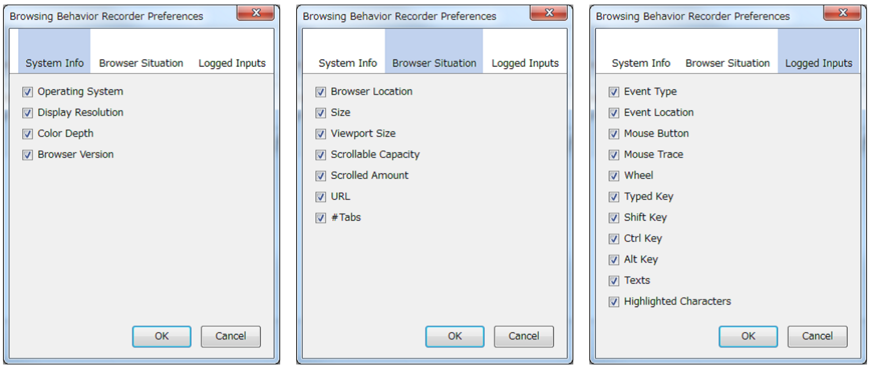

Mozilla Firefox Start Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Web Search Ctrl+K
Downloads Ctrl+J
Add-ons
Browsing Behavior Recorder
Error Console Ctrl+Shift+J
Page Info
Start Private Browsing Ctrl+Shift+P
Clear Recent History… Ctrl+Shift+Del
Options…
start
end
save as
-a&rls=org.mozilla:en-US:official
Mozilla Firefox Start Page
Web Images Videos Maps News
Google
Advanced search
Google Search
Mozilla Firefox Start Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/firefox?client=firefox-a&rls=org.mozilla:en-US:official
Mozilla Firefox Start Pa
Web Images Videos Maps
more ▾
start
end
save as
Firefox Start
Google
Advanced search
Google Search

Fig. 1. Interface to begin and end recording.



System Info

- ☑ Operating System
- ☑ Display Resolution
- ☑ Color Depth
- ☑ Browser Version

Browser Situation

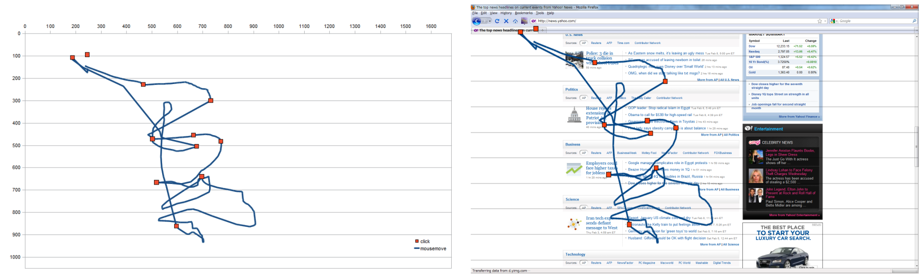
- ☑ Browser Location
- ☑ Size
- ☑ Viewport Size
- ☑ Scrollable Capacity
- ☑ Scrolled Amount
- ☑ URL
- ☑ #Tabs

Logged Inputs

- ☑ Event Type
- ☑ Event Location
- ☑ Mouse Button
- ☑ Mouse Trace
- ☑ Wheel
- ☑ Typed Key
- ☑ Shift Key
- ☑ Ctrl Key
- ☑ Alt Key
- ☑ Texts
- ☑ Highlighted Characters

Fig. 2. Recorded data selection.

of Yahoo!News. Trace data on a webpage can be extracted from mouse movement data recorded by our system easily. An example of trace data is shown in Fig. 3(a), and Fig. 3(b) shows it with the webpage obtained when analyzing the data. Although some parts of the trace data do not fit to the places the user saw at the webpage, it is almost possible to analyze the interesting places to the user from the data. Note, however, that the amount of the data shown in the figure became relatively large and the user may have to rid a record of mouse movement. Note also that we decided not to record a snapshot of the browsed webpage in the current implementation because this will make the performance of system bad.



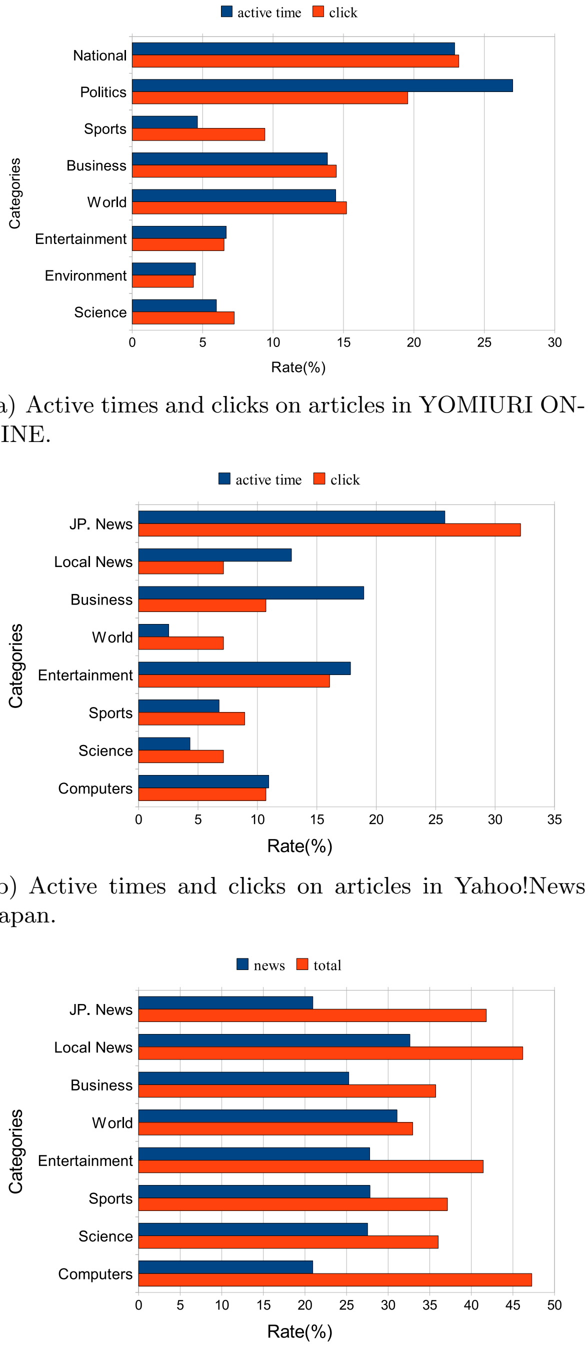
(a) Mouse movement trace data. (b) The data with the webpage.

Fig. 3. Mouse movement.

3.2 Spent Times and Browsed Areas

As described before, our system does not store snapshots of browsed webpages in the current implementation. In the following, we assumed that the webpages can be obtained when analyzing data recorded by our system.

Figures 4(a) and 4(b) show a user's active times and numbers of clicked articles in two news sites, namely YOMIURI ONLINE and Yahoo!News Japan, where active times are the total of times when the user stayed at articles. We can see from the figures that the tendencies of the active times and the numbers data in some categories were different from each other. For example, the user clicked many articles in sports category but spent little time to read them. This indicates that the user was interested in sports events, but the most interesting points might be match results. Figure 4(c) shows sizes of browsed areas in each news category. In the figure, "news" stands for the rate of read areas to total areas of body texts only and "total" stands for the rate of read areas to total areas of total articles, which included body texts, comments, and related descriptions. The fact that only about half area of articles in some categories were read can be indicated in the figure. From such situations, we may be able to improve layouts of contents in each category.



(a) Active times and clicks on articles in YOMIURI ONLINE.

(b) Active times and clicks on articles in Yahoo!News Japan.

(c) Browsed areas on articles in Yahoo!News Japan.

Fig. 4. Spent times, clicks, and browsed areas.

3.3 An Information Retrieval Case Study

We recorded a user's data of Web browsing behavior when he did an easy task of information retrieval. For the task we used one of the tasks in [1] as follows: Find an iphone app that tells us what song was playing when an iphone was held to the speaker playing the song. Google was used for the task in this study.

After finishing the task, the performed procedure the user remembered consisted of the following six steps. 1) He googled with "iphone app". 2) Then, he googled with "iphone app song title". 3) After observing the results, he accessed weblogs that introduced iphone apps. 4) He googled with "shazam". 5) Then, he accessed a webpage that explained shazam, and got a link to the app's developer's webpage. 6) Finally, he got correct information about the app.

The data recorded by our system told the procedure in more detail. 1) He googled with "iphone app". 2) Then, he googled with "iphone app song title". 3) After observing the results, he accessed weblogs that introduced iphone apps. When reading the weblogs, he highlighted "shazam". 4) He opened a new tab, and googled with "shazam" on the tab. 5) Then, he accessed a webpage that explained shazam, and opened another tab for browsing the app's developer's webpage linked from the webpage. 6) Finally, he got correct information about the app.

We found a couple of recorded points that may help analyzing information retrieval behavior. Also, the data may tell us many points if more difficult tasks were used; this will be included in our future work.

4 Conclusion

In this paper, we have introduced a Web browsing behavior recording system for research. Our system has been designed and developed for collecting Web browsing data to be used for research widely. The results of our preliminary experiments have shown that the data recorded by our system would tell us users' characteristics on browsing and thus be exploited in many applications for improving services. Collecting data from all over the world will be included in our future work. In addition, more detailed evaluation of our system in terms of concrete services will be performed in the future. We will also extend our system for protecting users' privacy more appropriately. Moreover, it would be interesting to develop a player of the data for visualizing them.

References

1. Aula, A., Khan, R.M., Guan, Z.: How does search behavior change as search becomes more difficult? In: Proceedings of the 28th international conference on Human factors in computing systems. pp. 35–44. CHI '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1753326.1753333
2. Aula, A., Nordhausen, K.: Modeling successful performance in web searching. Journal of the American Society for Information Science and Technology 57(12), 1678–1693 (2006), http://dx.doi.org/10.1002/asi.20340

3. Cheng, Z., Gao, B., Liu, T.Y.: Actively predicting diverse search intent from user browsing behaviors. In: Proceedings of the 19th international conference on World wide web. pp. 221–230. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772714
4. Dou, Z., Song, R., Wen, J.R.: A large-scale evaluation and analysis of personalized search strategies. In: Proceedings of the 16th international conference on World Wide Web. pp. 581–590. WWW '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1242572.1242651
5. Druin, A., Foss, E., Hutchinson, H., Golub, E., Hatley, L.: Children's roles using keyword search interfaces at home. In: Proceedings of the 28th international conference on Human factors in computing systems. pp. 413–422. CHI '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1753326.1753388
6. Fu, X., Budzik, J., Hammond, K.J.: Mining navigation history for recommendation. In: Proceedings of the 5th international conference on Intelligent user interfaces. pp. 106–112. IUI '00, ACM, New York, NY, USA (2000), http://doi.acm.org/10.1145/325737.325796
7. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1(3-4), 219–234 (2003), http://iospress.metapress.com/content/D68RMJ5V6C897X3C
8. Guo, Q., Agichtein, E.: Ready to buy or just browsing?: detecting web searcher goals from interaction data. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 130–137. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835473
9. Guo, Q., Agichtein, E.: Towards predicting web searcher gaze position from mouse movements. In: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems. pp. 3601–3606. CHI EA '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1753846.1754025
10. Holub, M., Bielikova, M.: Estimation of user interest in visited web page. In: Proceedings of the 19th international conference on World wide web. pp. 1111–1112. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772829
11. Hong, J.I., Landay, J.A.: Webquilt: a framework for capturing and visualizing the web experience. In: Proceedings of the 10th international conference on World Wide Web. pp. 717–724. WWW '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/371920.372188
12. Huntington, P., Nicholas, D., Jamali, H.R.: Employing log metrics to evaluate search behaviour and success: case study BBC search engine. Journal of Information Science 33(5), 584–597 (2007), http://jis.sagepub.com/content/33/5/584.abstract
13. Liu, C., White, R.W., Dumais, S.: Understanding web browsing behaviors through weibull analysis of dwell time. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 379–386. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835513
14. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 14th international conference on Intelligent user interfaces. pp. 31–40. IUI '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1719970.1719976

15. Matthijs, N., Radlinski, F.: Personalizing web search using long term browsing history. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 25–34. WSDM '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1935826.1935840
16. Shahabi, C., Chen, Y.S.: An adaptive recommendation system without explicit acquisition of user relevance feedback. Distributed and Parallel Databases 14(2), 173–192 (2003), http://dx.doi.org/10.1023/A:1024888710505
17. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 43–50. SIGIR '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1076034.1076045
18. Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: Proceedings of the 14th ACM international conference on Information and knowledge management. pp. 824–831. CIKM '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1099554.1099747
19. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive Web search based on user profile constructed without any effort from users. In: Proceedings of the 13th international conference on World Wide Web. pp. 675–684. WWW '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/988672.988764
20. Takashita, T., Abe, Y., Itokawa, T., Kitasuka, T., Aritsugi, M.: Design and implementation of a system for finding appropriate tags to photos in Flickr from Web browsing behaviour. Int. J. Web and Grid Services 7(1), 75–90 (2011), http://dx.doi.org/10.1504/IJWGS.2011.038385
21. Takashita, T., Itokawa, T., Kitasuka, T., Aritsugi, M.: Extracting user preference from Web browsing behaviour for spam filtering. Int. J. Advanced Intelligence Paradigms 1(2), 126–138 (2008), http://dx.doi.org/10.1504/IJAIP.2008.024769
22. Takashita, T., Itokawa, T., Kitasuka, T., Aritsugi, M.: A spam filtering method learning from Web browsing behavior. In: Lovrek, I., Howlett, R., Jain, L. (eds.) Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 5178, pp. 774–781. Springer Berlin / Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85565-1_96
23. Takashita, T., Itokawa, T., Kitasuka, T., Aritsugi, M.: Tag recommendation for Flickr using Web browsing behavior. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B. (eds.) Computational Science and Its Applications - ICCSA 2010, Lecture Notes in Computer Science, vol. 6017, pp. 412–421. Springer Berlin / Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-12165-4_33
24. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 449–456. SIGIR '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1076034.1076111
25. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 255–262. SIGIR '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1277741.1277787