Behavior-Based Web Page Evaluation

Ganesan Velayathan
The Graduate University for Advanced Studies
National Institute of Informatics (NII)
2-1-2 Hitotsubashi
Chiyoda, 101-8430 Tokyo, Japan
ganesan@grad.nii.ac.jp

Seiji Yamada
National Institute of Informatics (NII)
2-1-2 Hitotsubashi
Chiyoda, 101-8430 Tokyo, Japan
seiji@nii.ac.jp

Abstract

This paper describes our efforts to investigate factors in a user's browsing behavior to help automatically evaluate web pages that the user shows interest in. To evaluate a web page automatically, we have developed a client-side logging/analyzing tool: the GINIS Framework. We do not focus on clicking, scrolling, navigation, or duration of visit alone, but we propose integrating these patterns of interaction to recognize and evaluate a user's response to a given web page. Unlike most previous web studies that have analyzed access seen at proxies or server, this work focuses primarily on client site user behavior using a customized web browser. First, GINIS unobtrusively gathers logs of user behavior through the user's natural interaction with the web browser. Then it analyses the logs and extracts effective rules to evaluate web pages using a machine-learning method. Eventually, GINIS will be able to automatically evaluate web pages using these learned rules, after which the evaluation can be utilized in a variety of user profiling.

1. Introduction

Due to the overload of information on the web, the need to automatically build a user profile to assist the user in everyday browsing and searching tasks is steadily increasing, mainly for web personalization purposes. Building a user profile that adapts to a user's daily interests is a challenging task. This is because it is hard to predict which web sites interest the user most without asking the user to interact explicitly with the system. A less intrusive approach to the construction of user profiles is required. How can this be best achieved?

When users use a web browser to browse the Internet, they (intentionally or not) interact with the browser's interface; by clicking, scrolling, bookmarking, printing or selecting text, etc. If we can detect and learn the user's pattern of behavior, we can use these individual browsing habits to perform automatic page evaluation, and hence preferential page selection and prediction.

We aim to build a behavioral interaction database which matches a user's actions with pre-established behavioral patterns mined from that user, thus giving some idea of the user's interest (or lack of interest) in a particular webpage. Using this method, we believe we can achieve a detailed, granular, self-constructing user profile. We believe that careful study of these patterns at the client-side will lead us to a better understanding of which pages interest the user most. However, logging the user behavior at the client-side is a challenging task, for which we needed to build a high-level user client and logging tools. In this paper, we introduce the GINIS Framework, a framework developed to automatically log user browsing behavior; furthermore, we carried out an open environment user test using this framework.

2. Related Work
2.1 Web Navigation

Interest in web navigation has been an issue for commercial companies and researchers all around the world since the beginning of the WWW. One of the latest studies is the work presented by Weinreich et al., [5] observing 25 users for a mean usage duration of 105 days. Prior to this was the work presented by Catledge and Pitkow [1], (the first long-term client-side usage study ever made), observing 107 users over a duration of 21 days. However, most of these past works tended to focus on tracking user behavior itself rather than looking for connections between user behavior and user interest.

Past work on user interest and web usage behavior tended to focus more on navigation history, explicitly clicked links and the time spend on a web site. Morita et al., [2] found that there is a strong tendency for users to spend a greater length of time reading articles of interest to that user, and to spend less time on pages of less interest. However, [5] found that users spend less than 12 seconds on nearly 50% of the web pages shown to them. This means that users make nearly 50% of their decisions to navigate to the next page before reading substantial part of the contents.

2.2 Personalized Web Search

Personalization is the process of presenting the right information to the right user at the right moment [6]. In most cases, personalization begins with building a user profile based on user interest. Interest data can be collected either explicitly by asking for feedback from the user regarding preferences or implicitly by observing user behavior, such as time spent on a web page.

Much research on personalizing web searches tends to focus on building user profiles based on requiring users either to fill out questionnaires or to specify web page categories of interest. These conventional pre-defined user profiles must be manually constructed when a tool is initially used, and can be refined only through manual adjustment. However, since user interests change with time and task-orientation, every time a user's interest changes, the user has to manually re-create the user profile. Users may provide inconsistent or incorrect information, and the profile built up is static, whereas user interest may change over time, meaning that the construction of such a profile may place extra burden on users.[6]. Researches like [8] propose to automatically build user preferences by studying user browsing history. Studies such as [6] focus on automatically learning from user search history. Many other studies research have been focusing on building accurate user profiles based on implicit user behavior. To achieve effective personalization, profiles should distinguish between long-term and short-term interests, and should include a model of the user's context, i.e., the task in which the user is currently engaged and the environment in which they are situated [6].

3. Behavior-Based Web Page Evaluation

When browsing or searching the web, we tend (intentionally or not) to interact with the browser's interface. For example, we print, save, bookmark, move the mouse, click a link etc. This kind of navigation action is known as *user behavior*. Behavior such as printing, bookmarking, saving a page is considered to be *direct behavior*.[7] found that 95% of users are interested in the contents when they perform these kinds of direct actions. However, [5] finds that only 2% (which does not cover all direct behavior but most of it) of all user activity constitutes this direct behavior. From this we can conclude that, even though we can predict user interest from direct behavior, not many users are willing to perform these kinds of actions when browsing the Web.

When it comes to automatically building user profiles, the use of implicit user behavior (*indirect user behavior*) will play a major role. Careful studying indirect user behavior can help us to better understand user interest, and thus enhance personalization techniques.

We define navigation action as the "atom" of behavior; it can itself be a navigation behavior (e.g.: printing, clicking on a link), or a few navigation actions can combine to make up a new navigation behavior (e.g. click, move, click → Text Highlights).

3.1 GINIS Architecture

A framework called GINIS was developed during this research for the purpose of logging user behavior at the client side. The GINIS Framework was built based on .NET Framework 2.0, using MS Internet Explorer Component. It is an extendable framework intended for client-side logging.

The experiment for collecting user behavior is divided into 2 separate parts. The first part is the learning part, during which the users is prompted to answer every time before navigating to a new page whether they like the page they just viewed. This favorable information (along with behavior data) is collected and stored in a database.

The learning engine refines the logs from the raw data database and stores it in the User Behavioral Database (UDB). Classification learning (building a decision tree using C4.5) is performed based on the UDB. During the testing stage, the prediction engine generates and compares the user's new behavior with the behavior patterns stored in the UDB, and the user's implicit interest in a particular page is predicted based on this comparison.
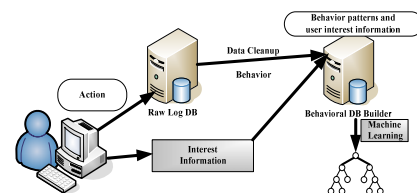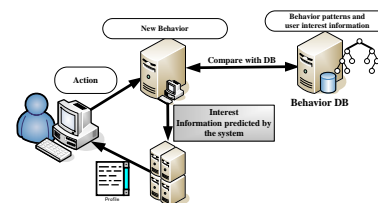

Figure1. GINIS Learning Stage

Figure2. GINIS Testing Stage

The GINIS Framework consists of 4 main modules: a client interface to detect and log user behavior (the browser), a database to store the user log information (the raw logger), a learning engine and a prediction engine.

The browser: Used to log user navigation actions, this mimics almost all of IE6.0, with some extension of the right-click menu. It is a tab browser; because of the popularity of tab browsers, and since IE7 will be tab-browsable, we hope that the result of this study will be more useful in the future.

The logger: A real-time user navigation logger, this logs user navigation behavior, including actions like mouse movement, scrolls, clicks, bookmarking, etc. More than 70

actions and 40 behaviors are logged. The fine details of its attributes are beyond the scope of this paper, so these are not covered here.

The analyzer: A behavioral database builder, this builds a decision tree based on user interest information. C4.5 decision tree algorithm[3] was chosen as the classification algorithm. The survey[4]shows that C4.5 provides good classification accuracy, and it is the fastest among the main-memory classification algorithms compared.

The predictor: Based on the classification learning built at the learning stage using the analyzer, the predictor module will automatically classify and assign user behavior to "interested" and "not-interested" classes.

4. Experiments

We conducted systematic experiments in order to investigate the ability of GINIS to gather logs of user behavior, learn rules to evaluate web pages, and predict user interest. We pre-tested the GINIS browser with 10 users, and asked the users to evaluate the browser. 3 simple questions were asked: How is the usability? Does GINIS contain all the features of a browser for normal daily use? Were users able to browse as if using their normal favorite browser? We obtained the score of 4.2, 4.3, & 4.3 for these questions respectively (on a 5-point evaluation scale).This demonstrates that the browser is easy to use and that users could navigate through their daily browsing without additional stress. We choose IE over other browsers because it is easy to log high-granularity data using the provided .NET framework.

4.1 Data Gathering

10 unpaid volunteers (6 male and 4 female) were recruited to participate in this study. They ranged in age from 21 years old to 38 years old (mean age: 29.1 years). All of them have between 4 and 12 years of web browsing history (mean history: 8.3 years). Almost all the participants used the internet for more than 6 hours per day (mean usage time: 10 hours per day), mainly either during work hours or for pleasure and entertainment. The users used the GINIS browser for a period of 16 to 31 days (mean usage duration: 22 days).

When using the GINIS browser, a questionnaire pops up every time the user undertakes a new navigation action (navigating to the next page). On that questionnaire, the user can choose "interested", "not-interested" or "unknown" in regard to the page they just viewed. The default answer was set to "unknown" - this was because sometimes when the users are very busy with deadlines etc, they tend to click without really considering the questions. The browser also allows the user to stop the popup, and instead use the question button at the side of the menu-bar.

The log file generated by GINIS contains data on date, time, Window ID, Tab ID in which the event occurred, the events that took place, how the event took place (toolbar, menu, right click, mouse, keyboard etc.), details of the event (document complete and browser functions: mouse movement, view source etc.), the URL of the page where the event occurred (once at Document Complete) and user interest information, the web page, detail query keywords, and character strings inputted via keyboard are collected for future use.

4.2 Data Processing and Cleanup

The original log file comprises over 460,000 lines, including 195,000 action data. Since GINIS was built with a very high logging capability (for future use of the data), we performed a log clean-up up to 65% (around 65,000 lines were left). Most of the data cleaned up were mouse movement logging (mouse locus and axis). The original log file also contains the text copied/pasted to/from the clipboard by the user. At the moment, behavior like text highlighting and Japanese-language keyboard input are logged at the action level but are not used. These actions had to be limited because they are far beyond the scope of this paper at the moment.

4.3 Data Summary

Table 1 shows the 5 most and 5 least frequently performed behaviors. Behaviors that did not take place are omitted. The total number of behaviors logged from the users was 52,161, and average per-page behavior was 4.7.

Table 1: Most and least frequent behaviors

Behavior	Most Frequent 5 (times)	Behavior	Least Frequent 5(times)
Scroll	19091	Go Forward	126
Key Input	14188	Stop Loading	88
Form Input	9329	Add to Favorite	79
Copy Text	1685	Print	64
Search Text	1284	Save As	2

5. Results

Classification learning was performed using C4.5, 2856 cases was used as training data, with 29 behavior attributes. 13 rules were generated based on this data (with default C4.5 settings) with 8 rules for "not-interested", and 5 for "interested". Out of 2856 training data, 1897 pages were classified as "interested" and 648 pages was classified as "not-interested" by the user. Here we show the top 5 "interested" rules and top 5 "not-interested" rules generated by C4.5.

5.1 Interested Rules

Rules generated and classified as "interested"

Rule 1: *Scroll > 14 & Reload <= 0* **[86.1%]**
Rule 2: *Stay Time <= 14 & Search Text > 0 & Form Input <=0 & Key Input <=0* **[85.4%]**
Rule 3: *Reload <=0 & Form Input > 2 & Navigate <=0 & Go Back > 0 & Go Back < 1***[83.5%]**
Rule 4: *Scroll < 3 & Reload <= 0 & Search Text < = 0 & Go Back <= 0* **[73.8%]**
Rule 5: *Search Text >0 & Key Input > 0* **[73.0%]**

5.2 Non-Interested Rules

Rules generated and classified as "not-interested".

Rule 6: *Form Input > 2 & Go Back > 1* **[82.0%]**
Rule 7: *Stay Time > 5 & Scroll > 3 & Scroll <= 7 & Search Text <= 0 & Form Input > 1 & Go Back <= 0* **[72.2%]**
Rule 8: *Stay Time > 11 & Scroll <= 0 & Search Text <= 0 & Form Input > 0 & Form Input <= 1 & Go Back <= 0 & Copy <= 0 & Paste Text <= 0* **[72.0%]**
Rule 9: *Stay Time > 4 & Scroll > 3 & Form Input > 1 & Key Input <= 0 & Go Back <= 0* **[67.5%]**
Rule 10: *Reload > 0 & Key Input <= 1 & Copy <= 0* **[59.8%]**

6. Discussion

6.1 Data Cleaning

Two things that affected our clean-up result were Google's result list page and online dictionaries. Most of the form input takes part here over at these pages. All our participants used a few freely available online dictionaries, which gave us considerable difficulty in cleaning up the results. The participants in this study tended to search for a particular word and then evaluate the answer of the query. To overcome this problem, a few restrictions have already been applied to the GINIS browser.

6.2 Classification Rules

As for the rules, generated from the classification learning, at present the combination of all 10 users were classified. One early focus of this work is to locate the common feature of web browsing of all users. The mean accuracy of rules classified as "interested" is 80.36% (SD=11.24%) and 70.7% (SD=7.22%) for rules classified as "not-interested".

At present, the common finding from all users, when the page is scrollable and when the user scroll few times or the user search for text (Ctrl – F etc.) the user show interest on that particular page. This can be seen from rule 1, 2 or 5. *Going back to previous page* can be concluded as one of the strongest behavior of non-interest. This is similar to previous findings and intuitively we do know about this. However, as we can learn from rule 7, 8 and 9, even if the user stays for a certain period of time (in seconds) on a page, that does not always means that it interests the user. This result shows that using longer stay time on a page as a positive example for machine learning is not always correct, and this result opposes most of the findings by previous researchers.

7. On-Going and Future Work

The scope of this paper consisted of collecting data using the GINIS Framework and performing classification learning to find new rules between user interest and behavior. The scroll action is the most frequently performed action on a page (after mouse action was removed from the log during our cleanup); we assume that more detailed study of this action will result in a better understanding of action/behavior conversions. A great number of mouse locus were logged during this experiment, and these still have to be studied carefully, as there is a strong possibility of finding new behaviors as well as valuable research ideas. At present, for the sake of simplicity, we have only taken into consideration the frequency of occurrence of a behavior on a single page during classification. There is a great need to carefully study the order of appearances, which we think would give us a better understanding of the relationship between user interest and user behavior.

8. Conclusion

We have presented here a method to automatically detect and log user behavior at the client-side by creating a client-side browser (the GINIS Framework). We used this framework to conduct an open environment (non-laborious) field study. While past researches tend to focus on user visit history and the time spent on a web page, our experimental result shows that other user behavior can also be used to find the user's interest. A higher granularity of behavior still needs to be taken into consideration, of which this study is limited at present.

9. References

[1] L. Catledge and J. Pitkow. Characterizing browsing behaviors on the World Wide Web. Computer Networks and ISDN Systems, 27(6), 1995.

[2] M. Morita and Y. Shinoda. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94), pp 272-281, 1994.

[3] R. Quilan, C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

[4] T. Lim, W. Loh, and Y. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine Learning Journal,40, pp203-228, 2000.

[5] H. Weinreich, H. Obendort, E. Herder, and M. Mayer, Off the Beaten Tracks: Exploring Three Aspects of Web Navigation, WWW Conference2006, ACM Press, 2006.

[6] M. Speretta, S. Gauch, Personalizing Search Based on User Search Histories, In: Thirteenth International Conference on Information and Knowledge Management, CIKM 2004.

[7] H. Kim, K. Chan, Implicit Indicators for Interesting Web Pages, Web Information System and Technologies(WEBIST2005), pp. 270-277,2005.

[8] J. Chaffee, S. Gauch. Personal Ontologies for Web Navigation. In Proceedings of the 9th International Conference on Information and Knowledge Management.(CIKM2000), pp. 227-234, 2000

[9] C. Chen, M.C. Chen, Y. Sun. PVA: a self-adaptive personal view agent. In proceedings of the seventh ACMSIGKDD international conference on Knowledge discovery and data mining 2001, San Francisco, California, pp. 257-262, 2001