Improving Browsing Time Estimation with Intentional Browsing Data

Yu-Hui Tao [†], **Tzung-Pei Hong and** [††] **Yu-Ming Su** [†††],

National University of Kaohsiung, Kaohsiung, Taiwan, R.O.C.

Summary
Web usage mining (WUM) utilizes mainly the Web logs as its data sources. A user's Web page browsing-time calculated from the Web logs usually involves some possible inaccuracies if the user was not navigating the Web page at the whole time. This paper proposes to use intentional browsing data (IDB), collected online from the user's interaction with the Web page, for adjusting the estimation of the browsing time for potential better application results. Two approaches on different phases of WUM, pre-processing adjustments versus mining algorithm, are proposed and illustrated for their rational benefits in addressing the estimation issue.

Key words:
Web Usage Mining, Web Log Files Browsing Behavior, Browsing Time.

1. Introduction

The majority of Web usage mining algorithms use Web log files as the main data sources in discovering useful information. Web log records that typically include host name or IP address, remote user name, login name, date stamp, retrieval method, HTTP completion code, and number of bytes in a file retrieved (http://www.w3.org/TR/WD-logfile). Therefore, browsing time or duration of stay on a Web page is a key item for Web mining algorithms. However, Web server logs only automatically record the time entering and leaving a certain Web page, without knowing whether the user is continuously working on that page? This precision of estimation presents an issue in many Web applications, such as online behavior analysis [1,5,8]. For example, in e-learning, browsing time within the user profile is used to measure the value of each Web page or performance of a user's learning. With inaccurate browsing time captured, decision makers become conservative to the reliability of the measurement. A very common practice is to automatically disconnect the user session by the Web server if no interactions over a certain period of time, such as 30 seconds, which is inconvenient and may cause customer retention problem.

Tao et al. [6] proposed a taxonomy of Web browsing behavior data and defined an intentional Behavior data (IBD), which was demonstrated for its potential benefits through the Web transaction mining algorithm by Yun and Chen [7]. Because an IBD is a Web browser component such as scroll bar, copy, save-as, and so on, the browsing idle time can be better adjusted with any event of IBD. To demonstrate the potential benefits of an IBD on browsing time estimate, the fuzzy Web mining algorithm (FWMA) proposed by Hong et al. [4] that mined fuzzy browsing path via time and Web Page is used for the illustration.

2. Web Usage Mining

In order to easily understand the core research content, the common background of the Web mining process is reviewed and the FWMA is briefed.

2.1 Web Mining Process

Applying data mining techniques to Internet data becomes a new area of Web mining [3]. Web mining has been developed into categories of Web structure mining that identifies authoritative Web pages, Web content mining that classifies Web documents automatically or constructs a multi-layered Web information base, and Web usage mining (WUM) that discovers users' access patterns of Web pages [2].

WUM can be further classified into three components, including pre-processing, mining algorithm and pattern analyzing [2] Pre-processing is proceeded before the actual mining procedure application for filtering and screening data in Web log files for appropriately feeding into the mining algorithm. The actual pre-processing consists of data cleaning, user identification, session identification, path complete and transaction identification. Mining algorithm is mainly the association rules or called basket analysis. When enough data is collected but not sure the appropriate statistical analysis, association can be applied. Pattern analyzing finds regularly occurred rules that are similar to association rules, except for the time sequential item is targeted in pattern analysis.

Manuscript received December 5, 2006.
Manuscript revised December 25, 2006.

2.2 Fuzzy Web Mining Algorithm

The FWMA [4] has total sixteen steps. Since the improvements in this research address only from step 4 to step 6, steps other than 4-6 were summarized in groups without unnecessary details:

Steps 1-3: Filter unnecessary data in Log Files and keep only the fields of date, time and client-imp. Transform the client-imps into contiguous integers as client ID, according to their first browsing time.

Step 4: Calculate the time durations of the Web pages by each client ID from the time interval between a Web page and its next page.

Step 5: For each client ID, sort its Web pages in order of browsing sequence.

Step 6: Transform the browsing time based on a given membership function as shown in Figure 1 into fuzzy values. For example, ID 1 browsed Web page B for 30 seconds, which was transformed into values of 0.8 in Low, 0.2 in Middle, and 0 in High according the membership function in Fig. 1. The fuzzy value representation is $(\frac{0.8}{Low}+\frac{0.2}{Middle}+\frac{0.0}{High})$.

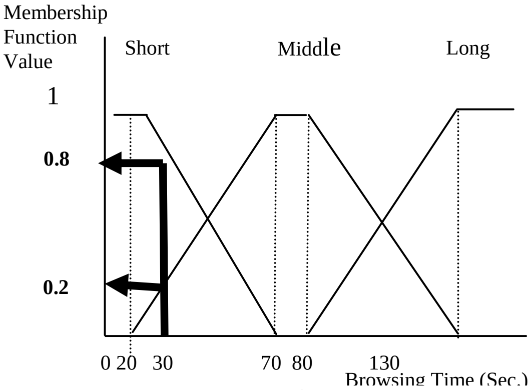

Fig. 1 Membership function diagram [4]

Steps 7-8: With the fuzzy value transformation, find the maximum value in each region in each browsing sequence and sum them up.

Steps 9-13: Find the short, middle and long paths in each fuzzy region, and record in L if higher than the pre-specified Sup value, where L is a 2-sequence candidate set if it is not empty.

Steps 14-16: Calculate the fuzzy membership values from each candidate sequence, and record in L if higher than the Sup value. Repeat steps 13-15 if L is not empty. When there is no sequence higher than the Sup value, stop the algorithm.

3. Benefits of IBD

The assumption is that when the browsing time is greater than the Middle range in the membership function, and no IBD appears, then the user is judged to be away from the computer screen. The rational is that when a Web page is longer than one screen page, the scroll-bar will be used to navigate down the Web page. Also, when there is any interested text or image, the user may select the interested content for further actions such as printing or cutting the selected area. Otherwise, when the browsing time captured is longer than a certain period while no IBD occurs, the most likely situation is that the user is not concentrating on the navigation.

3.1 Pre-Process Adjustment

If any IBD is captured while a user is browsing the Web pages, then the browsing times can be adjusted better before feeding into any formal algorithms for processing. Although this data pre-processing is not an exact solution to the inaccurate browsing-time problem, it helps to reduce the algorithmic error if any.

For example, if a total browsing time on a Web page is 120 seconds, and on the 40th second the user left the computer as shown in Fig. 2. In normal case, this particular browsing time estimate is either 80 seconds or three times larger. However, if a user's interaction with the Web page on the 15th second which is captured, and a pre-specified reasonable time for that action, say 30 seconds of buffer time (B), is assumed to be valid, then the time after 45 (15+30) seconds can be truncated. In other words, a pre-processing with IBD reduces the 120 seconds to 45 seconds of browsing time. Although there still exist 5 seconds of inaccuracy, 45 seconds is a much better estimate than 120 seconds compared to the actual browsing time of 40 seconds.

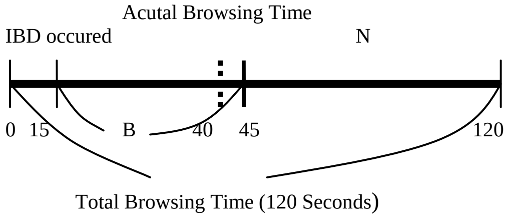

Fig. 2 Browsing time on a Web page

As illustrated, the pre-processing adjustment is not new and very intuitive. Its accuracy is realized only because of capturing an IBD as a new clue for aiding the truncation decision. The worse case is that if there is no IBD action captured during the browsing period, and the browsing time remains unchanged.

3.2 Illustration of Intention-Based FWMA

The browsing time can be adjusted before processing by the designated algorithm as a pre-processing approach or during the algorithmic process. This section adopts and modifies FWMA to an Intention-based FWMA (IFWMA) as an illustration for the second approach. The original example from Hong et al.[4] is used to illustrate how IBDs are added into the FWMA. In this paper, the Web pages browsed with their duration are represented by a tuple as (Web page, duration) as seen in Table 1.

Table 1 Web browsing data history

Client ID	(Web page, Duration)	Client ID	(Web page, Duration)
1	(B, 30)	4	(E, 118)
1	(E, 42)	4	(B, 11)
1	(D, 98)	4	(C, 42)
1	(C, 91)	5	(D, 64)
2	(D, 62)	5	(B, 29)
2	(B, 31)	5	(C, 74)
2	(D, 102)	6	(D, 80)
3	(A, 92)	6	(C, 61)
3	(D, 89)	6	(E, 122)
4	(B, 20)	6	(B, 17)
4	(C, 101)		

Based on the data prepared in Table 1, the FWMA starts from Step 5 as explained below.

Step 5：Add IBD occurrences corresponding to each page. The difference is on the judgment of browsing event by the IBDs, where $\mathbf{P}_{\mathbf{sub}}^{\mathbf{sup}}$ represents a **sub**-type of IBD **P** with **sup** number of occurrences as in Table 2.

Table 2 Browsing path Sequence with IBD elements

ID	Browsing Sequence
1	(B, 30, $\mathbf{P}_1^0$) (E, 42, $\mathbf{P}_4^1$) (D, 98, $\mathbf{P}_3^0$) (C, 91, $\mathbf{P}_2^2$)
2	(D, 62, $\mathbf{P}_3^3$) (B, 31, $\mathbf{P}_1^0$) (D, 102, $\mathbf{P}_3^2$)
3	(A, 92, $\mathbf{P}_0^1$) (D, 89, $\mathbf{P}_1^2$)
4	(B, 20, $\mathbf{P}_1^0$) (C, 101, $\mathbf{P}_2^2$) (E, 118, $\mathbf{P}_4^0$) (B, 11, $\mathbf{P}_1^2$) (C, 42, $\mathbf{P}_2^3$)
5	(D, 64, $\mathbf{P}_3^1$) (B, 29, $\mathbf{P}_1^0$) (C, 74, $\mathbf{P}_2^2$)
6	(D, 80, $\mathbf{P}_3^3$) (C, 61, $\mathbf{P}_2^1$) (E, 122, $\mathbf{P}_4^2$) (B, 17, $\mathbf{P}_1^1$)

Step 6：Transform the browsing times into fuzzy values according to the membership function as seen in Fig. 1. However, the IBDs remain unchanged with each fuzzy value as seen in Table 3.

Table 3 Fuzzy sets transformed from the browsing sequences

ClientID	Fuzzy Sets
1	$\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle},P_1^0\right)\left(\frac{0.6}{E.Short}+\frac{0.4}{E.Middle},P_4^1\right)\left(\frac{0.6}{D.Middle}+\frac{0.4}{D.Long},P_3^0\right)\left(\frac{0.8}{C.Middle}+\frac{0.2}{C.Long},P_2^2\right)$
2	$\left(\frac{0.2}{D.Short}+\frac{0.8}{D.Middle},P_3^3\right)\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle},P_1^0\right)\left(\frac{0.6}{D.Middle}+\frac{0.4}{D.High},P_3^2\right)$
3	$\left(\frac{0.8}{A.Middle}+\frac{0.2}{A.HIgh},P_0^1\right)\left(\frac{0.8}{D.Middle}+\frac{0.2}{D.High},P_1^2\right)$
4	$\left(\frac{1.0}{B.Short},P_1^0\right)\left(\frac{0.6}{C.Middle}+\frac{0.4}{C.High},P_2^2\right)\left(\frac{0.2}{E.Middle}+\frac{0.8}{E.High},P_4^0\right)\left(\frac{1.0}{B.Short},P_1^2\right)\left(\frac{0.6}{C.Short}+\frac{0.4}{C.Middle},P_2^3\right)$
5	$\left(\frac{1.0}{D.Middle},P_3^1\right)\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle},P_1^0\right)\left(\frac{1.0}{C.Middle},P_2^2\right)$
6	$\left(\frac{1.0}{D.Middle},P_3^3\right)\left(\frac{0.2}{C.Short}+\frac{0.8}{C.Middle},P_2^1\right)\left(\frac{0.2}{E.Middle}+\frac{0.8}{E.Long},P_4^2\right)\left(\frac{1.0}{B.Short},P_1^1\right)$

Table 4 IBDs judgment and Fuzzy values changes

Client ID	Fuzzy Sets
1	$\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle}\right)\left(\frac{0.6}{E.Short}+\frac{0.4}{E.Middle}\right)\left(\frac{0.6}{D.Middle}+\frac{0.4}{D.Long}\right)\left(\frac{0.8}{C.Middle}+\frac{0.2}{C.Long}\right)$
2	$\left(\frac{0.2}{D.Short}+\frac{0.8}{D.Middle}\right)\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle}\right)\left(\frac{0.6}{D.Middle}+\frac{0.4}{D.High}\right)$
3	$\left(\frac{0.8}{A.Middle}+\frac{0.2}{A.HIgh}\right)\left(\frac{0.8}{D.Middle}+\frac{0.2}{D.High}\right)$
4	$\left(\frac{1.0}{B.Short}\right)\left(\frac{0.6}{C.Middle}+\frac{0.4}{C.High}\right)\left(\frac{0.2}{E.Middle}+\frac{0.8}{E.High}\right)\left(\frac{1.0}{B.Short}\right)\left(\frac{0.6}{C.Short}+\frac{0.4}{C.Middle}\right)$
5	$\left(\frac{1.0}{D.Middle}\right)\left(\frac{0.8}{B.Short}+\frac{0.2}{B.Middle}\right)\left(\frac{1.0}{C.Middle}\right)$

6	$\left(\frac{1.0}{D.Middle}\right)\left(\frac{0.2}{C.Short}+\frac{0.8}{C.Middle}\right)\left(\frac{0.2}{E.Middle}+\frac{0.8}{E.Long}\right)\left(\frac{1.0}{B.Short}\right)$

Step 6-1：If a browsing time is greater than Middle and no IBD occurrence, change the fuzzy semantic value to Short and set the fuzzy value to 1 as illustrated in Table 4 .

The changes in Step 6 from FWMA to IFWMA can be seen from the records of ID=1 and ID=4. The Web page D of ID=1 changed from $\left(\frac{0.6}{D.Middle}+\frac{0.4}{D.Long}\right)$ to $\left(\frac{1}{D.Short}\right)$, while ID=4 changed from $\left(\frac{0.2}{E.Middle}+\frac{0.8}{E.High}\right)$ to $\left(\frac{1}{E.Short}\right)$. Both pages have browsing times greater than MIDDLE but without IBD occurrences and therefore the semantic value MIDDLE was changed to SHORT.

4. Rule Implications

In our own experiment for comparing IFWMA and FWMA, there are three rules derived by FWMA method, including (B.SHORT→ H.MIDDLE→ J.MIDDLE), (H.MIDDLE→ I.MIDDLE→ J.MIDDLE) and (K.SHORT→ L.LONG→ O.LONG), but only one rule derived by IFWMA method as (H.MIDDLE→ I.MIDDLE→ J.SHORT). The first and third rules generated by FWMA were purged in IFWMA due to possible inaccurate browsing time.

In another experiment for comparing our pre-processing adjustment to IFWMA, 37 browsing times out of 59 Web pages were adjusted with different amplitudes. Feeding the adjusted data set into the original FWMA, three rules derived as (B.SHORT → H.SHORT→ I.MIDDLE), (B.SHORT→ I.MIDDLE→ J.SHORT) and (H.SHORT→ I.MIDDLE → J.SHORT) while the original FWMA only derived one rule as (H.MIDDLE→ I.MIDDLE → J.SHORT). Contrary to the above IFWMA experiment, IBDs provided effective filtering on the raw data, which led to discovering the first two missing rules in the FWMA method.

The implications are two folds. First, the fuzzy semantic adjustment enhanced the predicative rules to be more accurate if any; second, both IFWMA and pre-processing adjustment reduce the possibility of misplacing limited resources such that the performance increases potentially. In practice, from a marketing viewpoint, allocating limited available resources to two possible faulty out of total three applicable rules greatly reduces the potential returns on the marketing investment. On the other hand, two additional quality rules provides more options in allocating resources to preferred marketing opportunities for better returns.

5. Conclusions and Future Work

As demonstrated, the major impact of IBD on the business implications is a more accurate WUM result if any, and consequently more appropriate usage of available resources in WUM applications. In addition to the applicable phases between pre-processing and mining algorithm in WUM as described by Cooley et al. (1999), another difference in applying IFWMA and pre-processing adjustments is that IFWMA targets on Web pages without IBDs while pre-processing adjustment on Web pages with IBDs. Accordingly, both methods can be applied together to the same data set without affecting each other, and may complement the effects of a more accurate predication.

One immediate future work is the decision of B value in the pre-processing adjustment. Even though B value is subjective to the Website management, IBD has also an influence on how B is calculated. For one end, browsing a professional technical Web page takes quite different amount of time from a news Web page. On the other end, for instance, a scroll-bar IBD and a printing IBD represent different meanings: a scrollbar indicates the B value should be higher since the navigation is not yet done, while printing implies a lower B value since the user may leave the page soon. Therefore, it takes some research to generate a few general rules of thumb and some domain-dependent rules for B-value setup.

Acknowledgments

This research is sponsored by the National Science Council of Taiwan, R.O.C. under grant number NSC 91-2416-H-214-002.

.

References

[1] Burklen, S., Marron, P. J., Fritsch, S. and Rothermel, K., “User centric walk: An integrated approach for modeling the browsing behavior of users on the Web,” Proceedings of the 38th annual Symposium on Simulation, IEEE Computer Society, pp. 149~159, 2005.

[2] Cooley, R., Mobasher, B. and Srivastava, J., ”Data preparation for mining World Wide Web browsing patterns”, *Journal of Knowledge and Information Systems*, Vol.1, Issue 1, pp.5~32, 1999.

[3] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, Academic Press, 2001.

[4] Hong, T. P., Lin, K. Y. and Wang, S. L., "Mining linguistic browsing patterns in the world wide web," *Soft Computing*, Vol. 6, No. 5, pp. 329-336, 2002.

[5] Li, H., C. Kuo, and M. G. Russell, "The impact of perceived channel utilities, shopping orientations, and demographics on the consumer's online buying behavior," Journal of Computer-Mediated Communication, 5(2), 15, 1999, available: http://jcmc.indiana.edu/vol5/issue2/hairong.html.

[6] Tao, Y., Su, Y. and Hong, T. P., "Web Transaction Mining Algorithm with Intentional Behaviour", *The Sixth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, Italy, September 2002.

[7] Yun, C.H. and Chen, M.S., "Using pattern-join and purchase-combination for mining transaction patterns in an electronic commerce environment", *The 24th Annual International Conference On Computer Software and Applications*, Taipei, Taiwan, pp.99~104, 25-27 Oct 2000.

[8] Zagorodnov, D., Brenna, L., Gurrin, C. and Johansen, D., " WAIFR: web-browsing attention recorder based on a state-transition model", in Proceedings of the 1st international workshop on Contextualized attention metadata: collecting, managing and exploiting of rich usage information, Arlington, Virginia, USA, pp. 21~26, 2006.

Yu-Hui Tao received his M.S. and Ph.D. degree in Industrial and Systems Engineering from the Ohio State University in 1989 and 1995, respectively. From 1994 to 1997, he was with Bank One Credit Card Company as a senior software consultant. Then, he started his teaching career in Taiwan, R.O.C. since 1997. His current research interests include e-business, management information systems, and Internet applications. He is now with National University of Kaohsiung.

Tzung-Pei Hong received his B.S. degree in chemical engineering from National Taiwan University in 1985, and his Ph.D. degree in computer science and information engineering from National Chiao-Tung University in 1992. He was a faculty at the Department of Computer Science in Chung-Hua Polytechnic Institute from 1992 to 1994, and at the Department of Information Management in I-Shou University from 1994 to 2001. He was in charge of the whole computerization and library planning for National University of Kaohsiung in Preparation from 1997 to 2000, and served as the first director of the library and computer center in National University of Kaohsiung from 2000 to 2001 and as the Dean of Academic Affairs from 2003 to 2006. He is currently a professor at the Department of Electrical Engineering and at the Department of Computer Science and Information Engineering. His current research interests include machine learning, data mining, soft computing, management information systems, and www applications.

Yu-Ming Su received his Master degree in Information Engineering from the I-Shou University in 2002. His current research interests include e-business, ERP systems and Web programming. Mr. Sun is involved in several steel-industry software development projects in both Taiwan and mainland China. He is currently a system analyst in InfoChamp System Corporation.