JOURNAL OF KOREA MULTIMEDIA SOCIETY, VOL. 6, NO. 4, JULY 2003

Design and Implementation of an Interestingness Analysis System for Web Personalizatoion & Customization

Youn-Hong Jung, Il Kim and Kyoo-seok Park

ABSTRACT

Convenience and promptness of the internet have been not only making the electronic commerce grow rapidly, in case of website, analyzing a navigation pattern of the users has been also making personalization and customization techniques develop rapidly for providing service accordant to individual interestingness. Web personalization and customization skill has been utilizing various methods, such as web log mining to use web log data and web mining to use the transaction of users etc, especially e-CRM analyzing a navigation pattern of the users. In this paper, We measure exact duration time of the users in web page and web site, compute weight about duration time each page, and propose a way to comprehend e-loyalty through the computed weight.

Key words: *duration time, interestingness, personalization, customization, e-loyalty, log file*

1. INTRODUCTION

Convenience and promptness of the internet have been making the electronic commerce grow rapidly.

Especially, web log mining to use web data and web mining to use the transaction of users etc, have been utilizing various methods on e-CRM analyzing a navigation pattern of the users.

Web log mining system processes mining of navigation pattern, relation analysis etc. through the preprocessing steps-such as data cleansing, User identification, Session Identification, path completion, transaction Identification etc.[2-4].

In chapter 2 of this paper, we describe duration time measurement and navigation pattern analysis to use existing log file analysis[2,4]. In chapter 3, we measure the exact duration time using web browser event, and then compute e-loyalty of duration time. In chapter 4, we analyze the duration time to measure using the proposed system, and then come to a conclusion in chapter 5.

This research has been funded by the Kyungnam University, Masan, Korea.

The authors are with the Dept. of Computer Engineering, Kyungnam Univ. Masan, Korea.
E-mail: vh5919@chollian.net, clinicagent@korea.com and kspark@kyungnam.ac.kr

2. RELATED RESEARCH

2.1 Duration Time Measurement

Most web servers save both request of web service and response in the log file. So we can know by using log file who requested or received which service when, and can also analyze the information of how many people visited the web server, came from where, what is liked most, or what is hated most, what page is seen longest and most etc.[1].

2.2 Duration Time Analysis

Computation of the duration time through analyzing web server log file of existing system includes the following problems.

i. Duration time can't be measured in external pages.
ii. Time to download web page is included in duration time. That is, As removing time to

download webpage in the computed duration time we can comprehend the exact duration-time. Finally, computing the duration time by analyzing the existing log file, a considerable error happens in case of web page that has a short duration time.

iii. In case of using proxy server or local cache, to connect the visiting page again can't be computed for the exact duration time because a record of the page don't be saved in the log file.

Therefore, To compute exact page & site duration time need the exact measurement of web page & site access speed.

3. PROPOSED SYSTEM

In this paper, we propose the following measurement method because of a number of errors in the measurement of duration time in the existing log file analysis system.

3.1 Duration time measurement

To analyze visitors' duration time exactly, we don't use the server standard time that makes use of at the existing log file analysis, we measure the duration time on the basis of event that happens in the web browser installed in the client, and use the following event.

- **onLoad**: The happening event when web page's download is completed.
- **onUnload**: The happening event when web page is externalized.

In Fig. 1, when visitiors navigate '***P1→P2→P3***

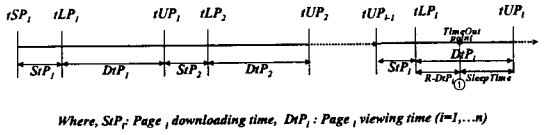


Fig. 1. Page Navigation Time.

→' in order, among data measured by web browser event, time measured by onLoad event is tLP_i, and time measured by onUnload event is tUP_i, if navigation time was measured like Table 1, navigating page duration time can be computed like this.

Table 1. Events' occurred time

event	occurred time
tLP_1	14:12:02
tUP_1	14:12:23
tLP_2	14:12:28
tUP_2	14:12:38
tLP_3	14:12:43
tUP_3	14:12:55

- Duration time of the Page1:
  tUP_1-tLP_1 = 14:12:23 - 14:12:02 = 21 sec.
- Duration time of the Page2:
  tUP_2-tLP_2 = 14:12:38 - 14:12:28 = 10 sec.

Therefore, Page duration time DtP_i can be computed by formula 1.

$$DtP_i = tUP_i - tLP_i \quad (1)$$

If there are no the user's activation after '①', in DtP_i of Fig. 1, the $R\text{-}DtP_i$ (: real DtP_i) should be computed as follow.

$$R-DtP_{oi} = DtP_i - SleepTime \quad (2)$$

Like this each page's access speed(download time) can be computed by measuring navigation time that uses web browser event.

When the download time for each page's access is StP_i, StP_i can be computed by formula 3.

$$StP_i = tLP_i - tUP_{(i-1)} \quad (3)$$

In case of the first accessing page tSP_1, it is impossible to measure by the download time StP_i that uses event of the web browser.

Therefore, download time StP_i of the first accessing page like '***P1***' can be computed as follow.

Compute download time StP_i of the navigating page after visiting web site and the average down-

load time $Average(StP_i)$, then compute average download time $Average(StP_i)$, StP_1 can be computed by formula 4.

$$StP_1 = Average(StP_1) \times \frac{\sum_{i=2}^{n} StP_i}{\sum_{i=2}^{n} Average(StP_i)} \quad (4)$$

Site duration time Dt_S, networking time $TStP$, and page duration time $TDtP$ can be computed by each formula 5, 6, 7.

$$TStP = \sum_{i=1}^{n} StP_i \quad (5)$$

$$TDtP = \sum_{i=1}^{n} Dt P_i \quad (6)$$

$$Dt\_S = tUP_n - tSP_1 = TStP + TDtP \quad (7)$$

3.2 Interestingness Analysis Process

When we compute e-loyalty to analyze users' interestingness each page, in case of applying the bulk same weight about the duration time of all pages like the existing method, we can't compute the exact e-loyalty.

So, we have to consider the size of the page and produce the users' interestingness and e-loyalty for the duration-time on each page. The algorithm for this is as follows.

step 1: To compute and save the users' exact duration-time on each page.

step 2: To classify and group the duration-time for each page.

step 3: If the duration-time is 0 or too excessive for the grouped duration-time, to replace '1' or remove such cases.

step 4: To decide standards to estimate the interestingness and to give the value of 0 or 1 to the 'class' variable.

step 5: From the result of step 4, to make the duration-time the independent variable and the class, the dependent variable and estimate the parameter.

step 6: For all pages the processes of *step 2~5* are iterated.

step 7: The estimated weight from the prior steps are applied to the new connectors and used for the e-loyalty after scoring.

This(the users and the duration-time by pages) can be summarized as in the form of Table 2.

In Table 2, to estimate the characteristics of the duration-time by page, we group for each page, remove the cases in which the duration-time is 0 or excessive where we can't determine the exact duration-time, and decide proper standards and give the value 0 or 1 to the class variables as in Fig. 2.

In Fig. 2, The cases where the value of accumulated relative frequency is over 90% are removed. The case where the value of accumulated relative frequency is 50% becomes the standard. If the value is less than that, 0 is assigned and over that, 1 is assigned to each class variable.

In Table 3, If we suppose the duration time x

Table 2. Users' duration time by pages

user	page	duration-time
u1	*P1*	21
u1	*P2*	10
u1	*P3*	12
u2	*P1*	18
u2	*P2*	11
:	:	:

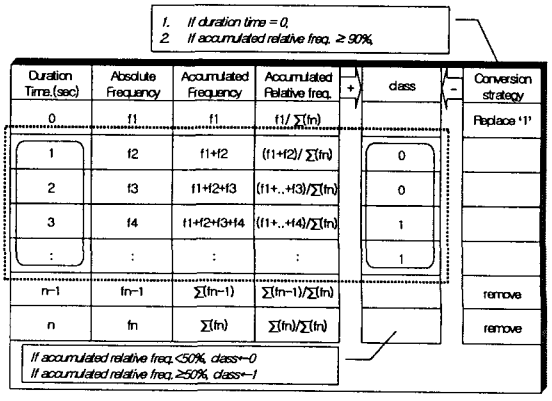


Fig. 2. Class generation for Pi's duration time.

Table 3. Duration time by page and class

sequence-no(i)	duration-time(x)	class(Y)
1	1	0
2	2	0
3	3	1
4	5	1
5	6	1
:	:	:

is the independent variable and class Y is the dependent variable, the dependent variable is a categorical variable that has only two values and the independent variable becomes one continuous variable.

3.3 The computation of a weight about th time

In Table 3, *Logit* Model is used for the judgement and prediction on the strength between the dependent variables and the independent variables. This dependent Y_i on independent x_i follows Bernoulli's trial of binomial distribution, $y_i \sim B(n, \pi_i)$, $i=1,\ldots,n$ and probability function is like formula 8. *Likelihood* function is regarded as combination probability function about n number's data[5]

$$f(Y_i|x_i) = \pi(x_i)^{Y}[1-\pi(x_i)]^{1-Y_i},\ i=1,\ldots,n \quad (8)$$

In formula 8, when we arrange to get the natural logarithm in the combined probability function which is the product of the each probability functions, that is like formula 9.

$$\log_e f(Y_1,\ldots,Y_n|X) = \log_e L(\beta_0,\beta_1) = \sum_{i=1}^{n} Y_i(\beta_0,\beta_1 x_i) - \sum_{i=1}^{n} \log_e[1+\exp(\beta_0,\beta_1 x_i)] \quad (9)$$

In formula 9, to obtain the *maximum likelihood estimated value* of the β_0 and β_1, when we do partial differencing $\log_e L(\beta_0, \beta_1)$ for the coefficient β_0 and β_1, $\log_e L(\beta_0, \beta_1)$ is the natural logarithm likelihood function, that is like formula 10.

$$\frac{\partial \log_e L(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$$

$$\frac{\partial \log_e L(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \quad (10)$$

Here, the *maximum likelihood estimated value* of the β_0 and β_1 is to get the solution of the normal equation after putting formula 10 into 0, and because the normal equation is nonlinear about parameter, getting solution to use *Newton Raphson iterative method* to handle in numerical analysis is saved in weight table each page like Table 4.

Table 4. Format of the weight table

page	coefficient0(β_0)	coefficient1(β_1)
p1	*-0.384*	0.274
p2	β_{20}	β_{21}
p3	β_{30}	β_{31}
p4	β_{40}	β_{41}
:	:	:
pi	β_{i0}	β_{i1}

3.4 Design of interestingness Analysis System

Fig. 3 is to represent a diagram of the process for the interestingness analysis system. In *Process A*, when users' connect homepage, *Pager, webdata* collector, is inserted in web page to detect the event of web browser and measures the exact duration time, and then transfer to '*Pager processor*' which saves to collect visitors' additional information. *Pager processor* analyzes information and pre-

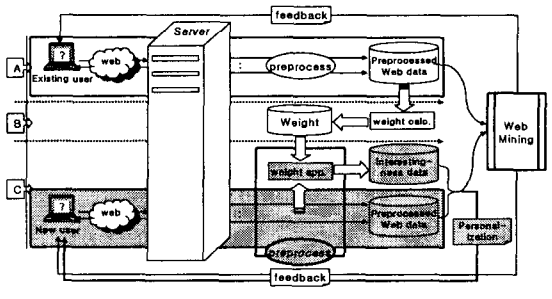


Fig. 3. Interestingness Analysis System of Duration time.

process the web-data and stores it.

In Fig. 3, *Process B* brings information of duration time each page in the preprocessed web data which is collected out of *Process A* and computes the coefficient of all pages by formula 10, and then saves in the weight Table.

Here the important facts are that we don't apply the same weight about duration time of all pages, and must award a different weight to consider each page's features, and then we can comprehend the exact interestingness of the individual duration time.

The form of the data to save in weight table is like Table 4.

Process C in Fig. 3, when new user accesses, after computing by formula 11 duration time to measure in *Pager* and weight to save in weight table, and save in L_{ij} of interestingness table to score each page with user's information.

$$P(Y_i=1|x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} \quad (11)$$

Here, $P(Y_i = 1|x_i)$ appears into form of probability value about a variable x_i, this is the meaning of the e-loyalty that can comprehend the interestingness of degree of duration time for user to stay in certain page, information to save here is to feedback to users through appropriate personalization and customization process.

According to information to save in the interestingness DB, used for giving individualized and differentiated service to each user, the structure of database to save is like Fig. 4.

In Fig. 4, the schema's structure of a interestingness DB is like Table 5.

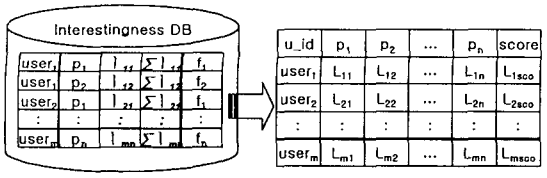


Fig. 4. The structure of the Interestingness DB.

Table 5. Interestingness table

field	memo
ID	Sequence No.
user	User ID
page	Page ID
loyalty	e-loyalty
aloyalty	accumulated e-loyalty
afreq	visits by pages

In Fig. 4, 'L_{ij}' is to indicate scoring e-loyalty for each page of each user, this is the meaning of interestingness information that a certain user is most interested in a certain page. And in score item, 'L_{isco}' which sums each page's e-loyalty is expressed like formula 12.

$$L_{isco} = \sum_{j=1}^{n} L_{ij}, \; i=1,\ldots,m \quad (12)$$

Because e-loyalty saved in interestingness table of Fig. 4 expresses to absolute value, it is appropriate for all pages' size to similar. But when page's size is irregular, e-loyalty's value saved must be applied to convert into relative value like formula 13.

$$L_{cj} = \frac{L_{cj}}{\sum_{i=1}^{m} L_{ij}}, \; j=1,\ldots,n, \; c=1,\ldots,m \quad (13)$$

4. IMPLEMENTATION

In this paper, the proposed system is implemented to use Windows 2000, IIS, MS-SQL 2000, ASP, JavaScript.

Fig. 5 is a diagram of the whole system that this paper proposes. When users access to the homepage where the proposed system is installed, the *Pager* inserted in web pages detects the browser's events, measures the exact duration time and sends it to the *Pager processor* that collects, analyzes and saves the information about the visitors.

About the *webdata* collected for a certain period, the weight on the duration time is produced and saved. This can be applied to the weight produced by new visitors and then interestingness DB is

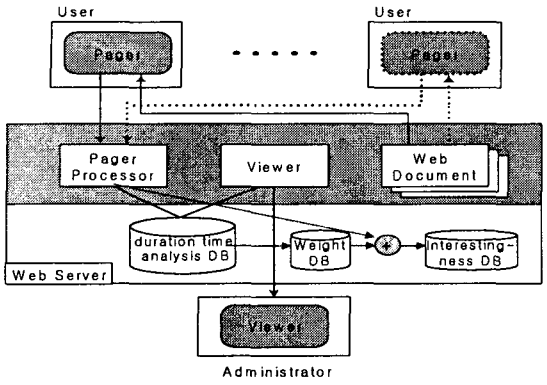


Fig. 5. The proposed system structure.

generated.

The administrator can examine the analyzed information with a special viewer.

Fig. 6. is the web interface about the duration time analyzed by the proposed system.

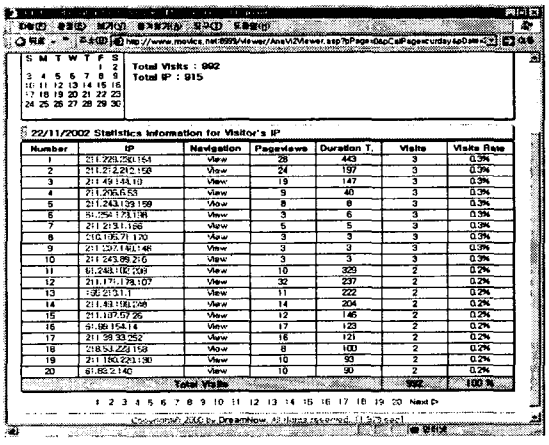


Fig. 6. Analysis of Duration time by user.

5. CONCLUSION

In this paper, we have designed and implemented a new system that can reduce the errors of the measurement on the existing the log file analysis about web users' pages and sites duration time, and we have proposed a interestingness analysis system using this duration time.

The interestingness analysis system proposed in this paper executes a lot of preprocessing work in a *webdata* collection phase, a *'preprocessed web-data'* based on this doesn't apply the same weight for duration time of all pages like the existing system, but awards different weights considering each page's features, thus we can comprehend more exactly individual interestingness than the existing systems.

Our future work includes the integrated weight with the *page view* and *visits* to comprehend the user's interestingness more exactly, e-CRM and recommendation system for effective customer management.

6. REFERENCES

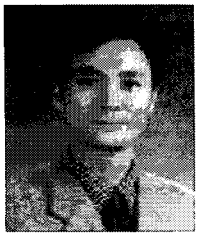
[1] http://www.dreamnow.co.kr

[2] Robert Walker Cooley, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", Ph.D.of Minnesota university, May 2000.

[3] Jose Luis Cabral de Moura Borges, "A Data Mining Model to Capture User Web Navigation Patterns", Ph.D. of London University, July 2000.

[4] Sun-Hee Yoon, Hae-Seok Oh, "Page Logging System for Web Mining Systems", Journal of the Korea Information Society, Vol. C-3, No. 6, pp.847-854, January 2001.

[5] Woong-Hyun Sung, "Applied Logistic Regression Analysis", TamJin, Korea, November 2001.



Youn-Hong Jung

He is a PhD course graduate student at Kyungnam University studying e-CRM, Internet Computing and Multimedia

Il Kim

He is currently working toward the PhD degree at Kyungnam University studying e-CRM, Internet Computing and Multimedia

Kyoo-seok Park

He is a professor of Computer engineering at Kyungnam University, Korea. He is the president of Korea Multimedia society now. His research focuses on Distributed system, specifically applied to Internet computing, Security system and Multimedia system. MS and PhD in Computer science from the Chung-Ang University, Korea

For information of this article, please send e-mail to: kspark@kyungnam.ac.kr(Kyoo-seok Park)