User Session Identification Using Enhanced Href Method

Jozef Kapusta, Peter Svec, Michal Munk, Jan Skalka

Department of Computer Science, Constantine the Philosopher University in Nitra, Slovakia
jkapusta@ukf.sk, psvec@ukf.sk, mmunk@ukf.sk, jskalka@ukf.sk

Abstract

One part of web log mining covers the process of discovering web users' behaviour patterns. This process employs user session identification techniques based on the web structure. There are some common heuristic methods which use referral URL from the web server log file. Using the referral URL is an alternative technique for the user session identification. We identified possible deficiencies of the common used h-ref methods and propose its enhancements. We applied the enhanced h-ref method on the web server log file and thus identified the user session in a different way. Next we compared basic characteristics of extracted user behaviour rules using the descriptive statistic methods among different h-ref methods. Results of the experiment show that the enhanced h-ref method does not affect session identification, only affect the inclusion of page visits into existing sessions. The new h-ref method is as effective as the generic one.

Keywords

Web Log Mining. Session Identification. Href method.

INTRODUCTION

Main source of information for web log mining is web server log file. A common web server keeps accesses of user in the log file and logs basic information about user's computer (e.g. IP address, date and time of referring page, browser version). Logs provide basic data as they record page accesses, not interaction with the page, and cannot make relevant distinctions such as that between the time a user spends reading and the time they spend away from the screen (Thomas, 2012). Data of outstanding quality requires rigorous data gathering as the data pre-processing. Data preparation is probably the longest and most time-consuming phase in the process of web usage mining because of incompleteness of accessible data as well as irrelevant information present in the collected data (Balogh & Koprda, 2012; Petr, Krupka, & Provaznikova, 2010a; Petr, Krupka, & Provaznikova, 2010b; Škorpil & Šťastný, 2009; Turcani & Kuna, 2012; Turčínek, Šťastný, & Motyčka, 2012) .

SESSION IDENTIFICATION METHODS

In the following section, we describe methodologies how to reconstruct the activity of every user, how to detach his activities from activities of other users while preserving his anonymity. This is demanding procedure from the perspective of the theory, time and technical realisation. Different aspects of data preparation can be found, for example, in work by Chitraa and Davamani (Chitraa & Davamani, 2010), Liu, Mobasher and Nasraoui (Nasraoui & Saka, 2007) or Liu (Liu, 2007). We also try to improve methods of session time threshold as this is key variable in session identification.

The separation of the user session on the basis of IP addresses is the simplest solution. But we must note the fact that IP addresses are not suitable in general for mapping and identification of individual site visitors. Currently it is not rare that several users share a common IP address, whether they are situated under a certain NAT (Network Address Translation), or proxy equipment. Another problem raise the situation where one user access the content using more than one computer (multiple IP address regarding one session) or using more than one browser at the same time.

User session identification using time threshold

By using the user session identification, we can differ users sharing one computer (classroom, library, etc.) and we can eliminate NAT and proxy devices. The option for the user session identification using time threshold (STT) are as follows:

- We can consider the session to be a set of user's clicks during the selected period, for example, during 30 minutes, 10 minutes etc. (Berendt & Spiliopoulou, 2000). It follows the duration of the session cannot be greater than $\boldsymbol{\theta}$. Define $\boldsymbol{Date_1}$ to be the access time (recorded in the log file) of the first page of session *S.* Next page with the access time $\boldsymbol{Date_k}$ can be added to session *S* only if $\boldsymbol{Date_k} - \boldsymbol{Date_1} \leq \boldsymbol{\theta}$. All other records of the log file with timestamp greater than $\boldsymbol{Date_1} + \boldsymbol{\theta}$ belong to the next user session.
- The second, more effective, method expects that the session is identified on the basis of sufficiently long interval of time among two recorded visits of the web page. Define σ to be selected time interval and $\boldsymbol{Date_i}$ to be the access time of the page added to session *S*. Next access to page with the access time $\boldsymbol{Date_{i+1}}$ can be added to session *S* only if $\boldsymbol{Date_{i+1}} - \boldsymbol{Date_i} \leq \boldsymbol{\sigma}$. If this condition is not true for two consecutive records of log file, these records belong to two different user sessions.

Heuristic method using the Referer and website map

We can get another view for the session identification if we take into account the website structure. Regular user browse the website using the hyperlinked structure among website pages. Each access to webpage is in the log file identified by the URL of accessed page and the prior webpage, so called *Referer*. Using the *Referer,* we can identify the session using the heuristic methods. In addition to *Referer* field, heuristic method uses also the website map. If we find two consecutive records with the same IP field, which are not directly connected with a hyperlink, we can say that we found access of two unique visitors sharing same IP address.

ISBN: 978-80-7478-497-2

Heuristic method h-ref

The h-ref method, as another heuristic method for user session identification based on the structure, takes into account the duration of the session and the referral page. (Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003).

Define *i* and *i+1* to be the consecutive request, $Date$ to be time of the request, URI to be requested page and $ReferURI$ to be referral page. For the defined time threshold σ, two consecutive requests fit one session if it is true that

$$ReferURI_i = URI_{i-1}, \quad (1)$$

Or two consecutive requests fit one session if (1) is not true or $ReferURI$ is not defined and it is true that

$$Date_i - Date_{i-1} \leq \sigma, \quad (2)$$

We can describe the problem of the page assignment into the session on the example in the Figure 1. When we reach request for the page *E*, two separate sessions have to be created. The first one is represented by the sequence *A-B-C-D,* and the page *A* represents the other one. The page *E* fits the first session as the referral page *D* was accessed in the first session either. The request for the *B* page (13:09) fits both sessions as the referral page *A* belongs to both sessions. According to (Liu, 2007) the page *A* fits the second session as it was created later.

Log file:

Request time	IP address	URI	RefererURI
12:00	194.160.10.10	A	-
12:01	194.160.10.10	B	A
12:04	194.160.10.10	C	B
12:08	194.160.10.10	D	C
13:00	194.160.10.10	A	-
13:04	**194.160.10.10**	**E**	**D**
13:09	194.160.10.10	B	A
13:12	194.160.10.10	C	B

Session 1

Request time	IP address	URI	RefererURI
12:00	194.160.10.10	A	-
12:01	194.160.10.10	B	A
12:04	194.160.10.10	C	B
12:08	194.160.10.10	D	C
13:04	**194.160.10.10**	**E**	**D**

Session 2

Request time	IP address	URI	RefererURI
13:00	194.160.10.10	A	-
13:09	194.160.10.10	B	A
13:12	194.160.10.10	C	B

Figure 1: Example of two sessions creation

ISBN: 978-80-7478-497-2

The value of time threshold is also important. We use the 60 minutes threshold time as suggested in (Seco & Cardoso, 2006). They assigned page into the session if at least one of the following condition were true. The page wasn't referred in previous sessions or the time interval between adjacent records was less than 60 minutes.

MODIFICATION OF THE H-REF METHOD

Let us assume a situation in the Figure 2. The last page added into session *S3* was the page *J*. The forthcoming page *Y* can be added into all three sessions, as the page *X* referenced it. According to (Cooley, Mobasher, & Srivastava, 1999; Liu, 2007) the page *Y* should be added to the session S3, as the lastly opened session. We propose to add the page *Y* into the session *S2* as the referral page *X* is "closer" to page *Y* in this session.

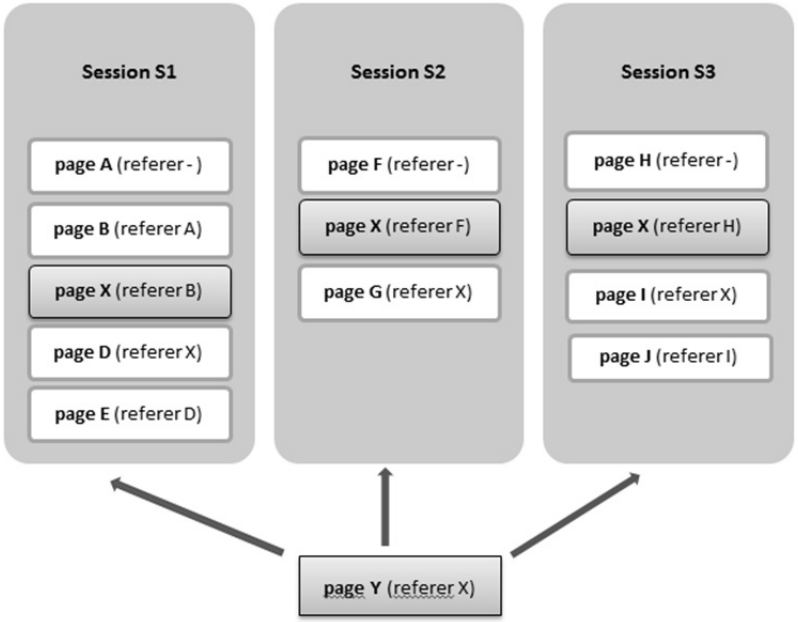

Figure 2: Sessions created using H-ref method

We add the page Y into that session which contains least requests for pages in the interval starting with the Y page and ending with the X page.

To verify the success of this method, we created an algorithm that identify session using the h-ref method. As the web server log file was in plain text format, we imported it into SQL database. In the first step, the algorithm creates sessions based on the IP address field. In the next step, it splits those session using the h-ref method.

Define row_i to be exactly one record of the log file, $S = \{s_1, \dots, s_n\}$ to be analyzed sessions and s_{actual} to be the session with the last analyzed page ($row_{i-1} \in s_{actual}$). If there is no *Referer* for the row_i or the referer is outside the analyzed domain, we have to create new session $s_j (s_j \in S)$. The row_i is the first record of the s_j session. If there is a referer for the row_i we have to create set of session S_T, where $S_T \subseteq S$ and S_T contains sessions, which are able to hold row_i.

ISBN: 978-80-7478-497-2

We can assign the row_i to actual session s_{actual} if following conditions are met:

- The URL of the last record row_{i-1} assigned to s_{actual} equals to *Referer* of row_i
- $S_T = \{s_{actual}\}$; The actual session is the only one session to which we can assign the row_i
- $S_T = \emptyset$; We cannot assign row_i to any session based on the *Referer*.

Above mentioned cases are identical for all modifications of h-ref method. If we cannot assign row_i to S_{actual} using above cases, we can try to consider following options, we tried in our experiment:

- If $s_{actual} \in S_T$ we assign the row_i into s_{actual}, otherwise we assign it into the first session s_1 ($s_1 \in S_T$). This step follows the origin method.
- We assign the row_i based on the closest page access.
- We can randomly assign the row_i into the session from the set S_T.

As the amount of data is quite big and in later experiments can be bigger, the performance of the algorithm can be the bottleneck. As the sessions are created based on the IP address, we can use this parameter as the unique pool ID for the parallel computation. The parallel algorithm is as follow:

```
Group_IP_addresses();
Get_list_of_grouped_addresses();
Calculate_the_portion_of IP_pool_based_on_the_number_of_threads();
For I = 1 to number_of_threads {
    Fork_child();
    Call_href_method(i-portion of ip pool)
}
```

Each thread takes records from log file based on the IP field. Consider we want to have eight threads as computing power. We assign the first eighth of IP pool to the first thread, second eighth to the second thread etc. It doesn't matter that in the web server log file records are not sorted based on the IP address, because we transformed the log file into the SQL database.

EXPERIMENTAL VERIFICATION OF ENHANCED H-REF METHOD

Appropriateness and effectiveness of enhanced method h-ref has been verified by experiments. Data source for the experiment is the commercial bank web server log file of the 2010 year acquired as part of our research VEGA project. We used standard methods for cleaning data from unnecessary data (requests for images, style sheets, etc.) and crawlers' accesses.

We have to take into account extreme cases, so we define following heuristic: If there is a time difference between two consecutive records in the log file higher than 60 minutes, the second record is considered as a new session (Seco & Cardoso, 2006). We

ISBN: 978-80-7478-497-2

consider that 60 minutes pause doesn't mean that the user is reading one page for one hour. This consideration is of course disputable, but due to the calculation of other real values for the time threshold for the identification of sessions, it is very important. Log file with identified session is the starting point for an application of h-ref method. We identified session more precisely using following methods:

- Href-ACTIVE: Session identification using the origin h-ref method.
- Href-SEARCH: Session identification using enhanced h-ref method.
- Href-RANDOM: Session identification using random assignment into multiple possible sessions.

Experiment Results

If we use the path-competition method, the numbers of records in the log file increase. The number of identified sessions depends on the method of identification. In case of h-ref method, a new sessions are created only in case that the *Referer* is empty or the *Referer* contain URL outside domain.

Table 1: The count of visits and count of sessions

	Href-SEARCH	**Href-ACTIVE**	**Href-RANDOM**
Number of records	938497	938497	938497
Number of visits (identified sessions)	378435	378435	378435
Average length of identified sessions	2	2	2
Frequent sessions (s = 0.5 %, c = 0.5 %)	110	109	110
Frequent sessions (s = 1 %, c = 1 %)	49	49	49

The session identification based on the enhanced h-ref method does not have a significant impact on the quantity of extracted rules (Table 1). There is also no difference when we consider results (support, confidence) of sequence rule analysis.

Next step we have to consider evaluating the quality of identified session is the categorization of extracted rules based on the level of usefulness into three categories (*useful, trivial* and inexplicable rules) and intercomparison among all three dataset. As the rating of the rule is the same in all three methods, it doesn't matter the rating is subjective. We weight the trivial rules at zero, as trivial rules represented with association rules do not bring a new view on users' behaviour. If we consider the amount of useful rules, the method, which discovers more useful rules, is better. On the other hand, if we consider the amount of inexplicable rules, the method, which doesn't generate inexplicable rules, is better.

Using the Href-Search and Href-Random we discover one extra rule missing in the Href-Active. The extra rule had similar values of support and confidence (Table 2).

ISBN: 978-80-7478-497-2

Table 2: Extracted rule missing in the Href-Active

Body	=>	Head	Support(%)	Confidence (%)
(/about/contacts/write-to-us.html)	=>	(/about/branches-and-atms.html)	0.4989	7.5712

We can interpret the rule as follows: From all visits at the portal, the portion of 0.5% view the */about/contacts/write-to-us.html* page. From all these visitors, just 7.57% visitors continued to page */about/branches-and-atms.html*. Even if we can consider this rule as useful, the values of support and confidence are negligible.

The difference between all three methods can be the length of identified sessions. We detected most differences between Href-Active and Href-search method (1190 sessions, but this represents only 0.31% of all sessions). Comparison of the two methods against Href-Random method was almost the same (889 in case of Href-Active and 878 in case of H-ref Search).

CONCLUSION

The experiment did not show a significant difference in results between the examined methods. Improved efficiency of enhanced h-ref method has not been proved. The main causes of this finding might include the fact that we made just a small improvement which does not affect the creation of new sessions. It only affects the assignment of visits to existing sessions.

We analysed 938,497 records and just in case of 2,611 records the algorithm had to determine the correct open session. This represents only about 0.29% of all cases.

The original idea for enhancement of h-ref algorithm was focused on the optimization for the path completion method. This is applied as the last method of preparing data for sequence analysis. In the next experiment, we plan to use this method on the same dataset and compare generated sessions and extracted rule based on their quantity and quality.

REFERENCES

Balogh, Z., Koprda, S. 2012. Modeling of Control in Educational Process by LMS. In. Divai 2012: 9th International Scientific Conference on Distance Learning in Applied Informatics: Conference Proceedings, p. 43-51

Berendt, B., Spiliopoulou, M. 2000. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, 9(1), p. 56-75.

Cooley, R., Mobasher, B., Srivastava, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information System, 1.

Chitraa, V., Davamani, A. S. 2010. An Efficient Path Completion Technique for web log mining. In: IEEE International Conference on Computational Intelligence and Computing Research.

Liu, B. 2007. Web data mining. New York: Springer.

ISBN: 978-80-7478-497-2

Nasraoui, O., Saka, E. 2007. Web Usage Mining in Noisy and Ambiguous Environments: Exploring the Role of Concept Hierarchies, Compression, and Robust User Profiles. In B. Berendt, A. Hotho, D. Mladenic & G. Semeraro (Eds.), From Web to Social Web: Discovering and Deploying User and Content Profiles (Vol. 4737, p. 82-101): Springer Berlin Heidelberg.

Petr, P., Krupka, J., Provaznikova, R. 2010a. Mathematics Model Design Based on Genetic Programming. In. Proceedings of 13th International Symposium on Mechatronics. Trencianske Teplice, Slovakia. p. 115-117

Petr, P., Krupka, J., Provaznikova, R. 2010b. Statistical Approach to Analysis of the Regions. In. 10th WSEAS International Conference on Applied Computer Science. Iwate, Japan. p. 280-285

Seco, N., Cardoso, N. 2006. Detecting User Sessions in the Tumba! Query Log. Tech. rep. Faculdade de Ciências da Universidade de Lisboa.

Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M. 2003. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. INFORMS Journal on Computing, 15(2), 171-190.

Škorpil, V., Šťastný, J. 2009. Comparison Methods for Object Recognition. Paper presented at the Proceedings of the 13th WSEAS International Conference on Systems, Rhodos, Greece.

Thomas, P. 2012. Explaining difficulty navigating a website using page view data. Paper presented at the Proceedings of the Seventeenth Australasian Document Computing Symposium, Dunedin, New Zealand.

Turcani, M., Kuna, P. 2012. Modelling the Student's Transition Through the E-course "Discrete Math" Using Petri Nets. In: Divai 2012: 9th International Scientific Conference on Distance Learning in Applied Informatics: Conference Proceedings, p. 319-328

Turčínek, P., Šťastný, J., Motyčka, A. 2012. Usage of Data Mining Techniques on Marketing Research Data. Paper presented at the Proceedings of the 11th WSEAS International Conference on Applied Computer and Computational Science (ACACOS '12), Rovaniemi, Finland.

ISBN: 978-80-7478-497-2