iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** VI **Month of publication:** June 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

An Enhanced Methods of Preprocessing in Web Log Mining

Anu Panwar, Mr. Shiv Naresh Shivhare

M.Tech(CS&E)

Galgotias University, U.P, India

Abstract:This paper tries to evaluate the web logs by using the technology used in the domain of data mining. Through content mining, structure mining and usage mining, evaluates and studies the existing problems in current mining technology. Focusing at significant key preprocessing schedule such as Frame page filtration, time-out threshold value setting and long-time setting, etc. This paper puts forward a modified data preprocessing technical method and analyze the mining result through tests before and after the alteration. The experiment proves that, the modified preprocessing technology is feasible and it can solve problems existing in relevant preprocessing effectively.

Keywords-Data mining; web log mining; ID3 algorithm; preprocessing;

1 INTRODUCTION

Data mining, especially web mining is the process of information extraction from WWW resources, which is a process of extracting mode with unknown, implicative and potential application value. Because of the disorder and the hugeness of web data volume, and the dynamic property of information source, it is extremely difficult to find out the useful information from the web. Web mining consists structure mining, usage mining and content mining, etc. Data preprocessing is the primary and principle step, and also one of the most important steps of web log mining. Preprocessing procedure is the key to assure the quality of web log mining, and the result of preprocessing has a candid impact on the choosing of mining algorithm and mode discovery. General preprocessing of web log data includes: User identification, session identification, data purification, path supplement and transaction identification.

1.1 Data purification

Purification of data refers to deletion of data irrelevant to mining algorithm in Web logs, e.g. the information updated or amended by the supervisor. Deleting this information is more in favor of improving our mining efficiency. In addition, different web logs have different formats, so cleaning means shall be chosen according to the actual demand.

1.2 User identification

The so-called user identification is the process of identifying every user visiting the website and connecting it with the Requested page. It used mainly for identifying the Characteristics of the user's visit and preparing for the

Following User mode analysis. However, due to the usage of proxy server and firewall, the user identification becomes complex. There could be so many ways like cookies, embedded user's ID, client side software agent, etc., which can help us to identify potential users, but all these are based on client side, related to users' privacy and need the cooperation of users, and thus the usage has greatly been limited. The main work of user identification is to deal with these problems.

1.3 Session identification

Session is a continuous and effective visit of the user, and it is obvious that different users' visits belong to different sessions. At the same time, if the interval of two consecutive visits of one user is quite large, the user is considered to start two different sessions. General timestamp Timeout is set as 30min.

1.4 Path supplement

Due to buffer memory of client-side, the user may use the back function of the browser when browsing, therefore, ratiocination shall be conducted according to the backwards and forwards pages visited by the user making supplement to user's access path. In this way, website structure can be adjusted and optimized, so that users can visit pages more simply and fast. It also can be used in intelligent recommendation and customized electronic commerce activities according to user's typical browse mode.

IJRASET: ISSN: 2321-9653

ISSN: 2321-9653

1.5 Transaction Identification

A user session is the only element with natural transaction characteristic of web log mining, but for some mining algorithms, the data granularity of the user session may be too large and needs to be converted into smaller transactions by using segmentation algorithm to identify. The procedure of data preprocessing is shown in Figure 1.

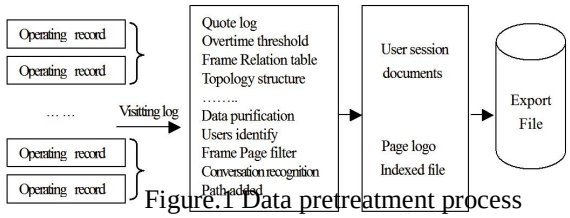


Figure.1 Data pretreatment process

The session identification of preprocessing technology has great influence on the follow-up mining. Domestic and foreign scholars carry out extensive research [2] on this. Literature [4] puts forward to add Frame page filtration between the two steps of the session identification and path completion. Although SubFrame page has been filtrated, it also leads to the loss of the page information. Literature [3] puts forward page filtering algorithm based on ID3 algorithm, which overcomes the problem of information loss in Literature [4]. In the time-out threshold value setting, Literature [5-6, 10] adopts fixed time threshold value setting based on time, structure and quote, respectively, but there is a problem that session record has wrongly been divided up. Literature [7] puts forward that each page generates a threshold value in the time-out threshold value setting, which effectively solves the identified problem of long time session. Based on the references, this paper combines ID3 algorithm with time-out threshold value dynamic correction. Based on overcoming the problem of information, loss of page filtration, it optimizes the time-out threshold value setting and increases the accuracy of the preprocessing result.

2 ANALYSIS OF PREPROCESSING METHOD

2.1 Frame Page Filter

In the actual process of web log mining, as the SubFrame page of session file is deleted, therefore, the hyperlink information contained in the SubFrame page will be lost. So it must rely on the site structure to extract useful Frame-SubFrame relational table. We call the searching for Frame page and its SubFrame page as Frame page filtration. The web log mining procedure of using Frame page filtration method is shown in Figure 2.

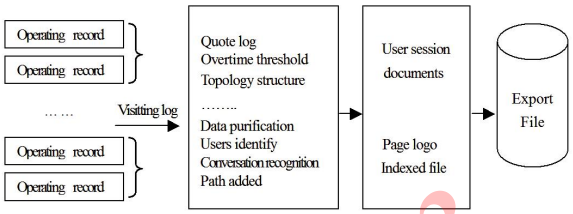


Figure 2. The web log mining procedure of using Frame page filtration

2.2 Time-out Threshold Value Setting

In the process of session identification, generally the time-out method is adopted to identify the user session. If the difference value of the requested time between two pages exceeds a certain limit (time -out threshold value), it can be identified that the user starts a new session. Adjusts this threshold value through setting threshold value of access time for pages and according to the degree of importance determined by page contents and site structure.

3. ENHANCED METHODS OF PREPROCESSING

3.1 Improve the filtration method of Frame page

As mentioned above, applying the filtration method of Frame page can effectively eliminate the influence of Frame page on the log mining. However, the records of web log mining are very large. The filtering algorithm of Frame page in literature is judging whether each page of each user's session is Frame or SubFrame and deleting the sub -frame which has been judged one by one. Moreover, because of the deleting of SubFrame page the upgraded site structure must be used in the following. Although interest degree has been added compared with the general preprocessing technology, the efficiency is still comparatively low and the spending also has been increased. Besides, SubFrame is deleted during the filtration, and it is also debatable that whether SubFrame can get full recovery in later path supplement. With consideration of decision tree sorting algorithm has the property of allowing quick sort of multi-granularity layer, therefore, it can solve this problem better. This paper adopts ID3 algorithm in decision tree algorithm, and improves the filtration method of Frame page, in this way, the filtering efficiency can be improved.

3.2 Algorithm description of ID3

The basic idea of ID3 algorithm is greedy algorithm which adopts super incumbent and divide and rule method to construct a decision tree. First, check all characteristics of the training data set and select feature A with the largest information gain to establish the decision tree root node. Establish branches according to the different values of this characteristic and conduct recursion on an example subset of each branch. Use this method to establish the node and branch

IJRASET: ISSN: 2321-9653

ISSN: 2321-9653

of a tree, till the data of a certain subset belongs to the same sort, or without characteristic to be used in data splitting. The algorithm is described as follows:

Algorithm: Generate decision tree generates one decision tree with the given training data set.

Input: Training samples are expressed by the attribute of the discrete value; the attribute list of alternate attribute.

Output: one decision tree y.

The computing method is shown in the formula (1) to formula (3).

In the formula, it represents the access time. To set threshold value D by access time in each page, firstly should obtain the access time t of the page after statistics and adjust D combining with RLCR effector B of the page. The set of t is recorded as St= $\{t_1, t_2, \ldots, t_n\}$ and the set of affector B of the page is recorded as SB=$\{B_1, B_2, \ldots, B_n\}$.

4 EXPERIMENTAL ANALYSIS

This paper proves the improvement effects of the above algorithm in page filtering through experiments. The experiment is implemented with 12 MB logs, including 130 thousand records and 712 different HTML pages, and identified 2122 user sessions from that. Comparing the earlier Page filtering preprocessing technology with the filtering technology of ID3 by mining frequent accessed page group. The method comparison is shown in Table 1-3.

TABLE 1 FILTERING DATA OF GENERAL TECHNIQUES PAGE

Sen	f1	f2	f3	f4	f5	f6
58	25	75	96	67	20	7*
72	29	54	83	61	14	4*

TABLE 2 FILTERING DATA OF FRAME IMPROVED TECHNOLOGY PAGE

Sen	f1	f2	f3	f4	f5	f6
30	25	34	18	3	0	0
15	55	98	69	18	2	0

TABLE 3 FILTERING DATA OF FRAME IMPROVED TECHNOLOGY PAGE BASED ON ID3

Sen	f1	f2	f3	f4	f5	f6
20	23	34	9	1	0	0
10	65	77	54	9	0	0

In Table 1-3, Sen refers to the minimal user session number containing frequent visit page group; fee refers to the frequent visit page group number with my length; * refers to the detected frequent visit page group which users are interested in; not referred to the detected frequent visit page group which users are relatively interested in; refers to the detected frequent visit page group which users are interested in. Through Table 1-3, it is observed that, after the application of ID3 in Frame filtration, the quality of data preprocessing results is improved, meanwhile, it increased the interest degree of the mining results in large degree. Due to reducing of the step of upgrading site structure, this improved method increased the preprocessing efficiency, thereby, increased the mining efficiency of the whole web log.

5. CONCLUSION

This paper aims at the research on the preprocessing procedure of web log mining, uses ID3 algorithm to improve Frame page filtration method, adds pages with largest information gain to Frame page, and adds the unsuitable ones to the SubFrame page, And the SubFrame has not been deleted, which reduced the step of upgrading the site structure in path supplement. Meanwhile, adopting dynamic correction method through time-out threshold value in session identification can strengthen the recognizing ability of comparatively long time session in preprocessing procedure. Applying the improved methods of this paper to the web log mining preprocessing procedure can increase the interest of mining results.

REFERENCES

[1] Ming Z. Data mining. Hefei: University of Science and Technology of China Press, 2008:13-56.

[2] Han J W, Meng X F, Ang J. Research on web miming. Journal of Computer Research & Development, 2001, 38 (4): 405-414.

[3] Jing Z, Fang L. Filtering Technology of Frame Page Based on Id3 Arithmetic. Computer and Information Technology, 2007, 15 (6):7-9.

[4]Yiling Y,Xudong G,Lina L. Frame Page Arithmetic of Web log Mining Preprocessing. The Computer Engineering, 2001, 27 (2):76-77.

[5] Cooley R, Mobasher B, Srivastava J. Data preparation for mining world wide web browsing patterns. Knowledge and Information System, 1999, 1 (1):5-32.

[6] Spiliopoulou M, Mobasher B, Berendt B, et al. A framework for the evaluation of session reconstruction heuristics in web usage analysis. Informs Journal of Computing, 2004, 15(2):171-179.

[7]Yuankang F,Xuegang H,Qishou X. An Improved Recognition Methods Web log Conversation. Computer Technology and Development, 2008, 18 (11): 214-216.

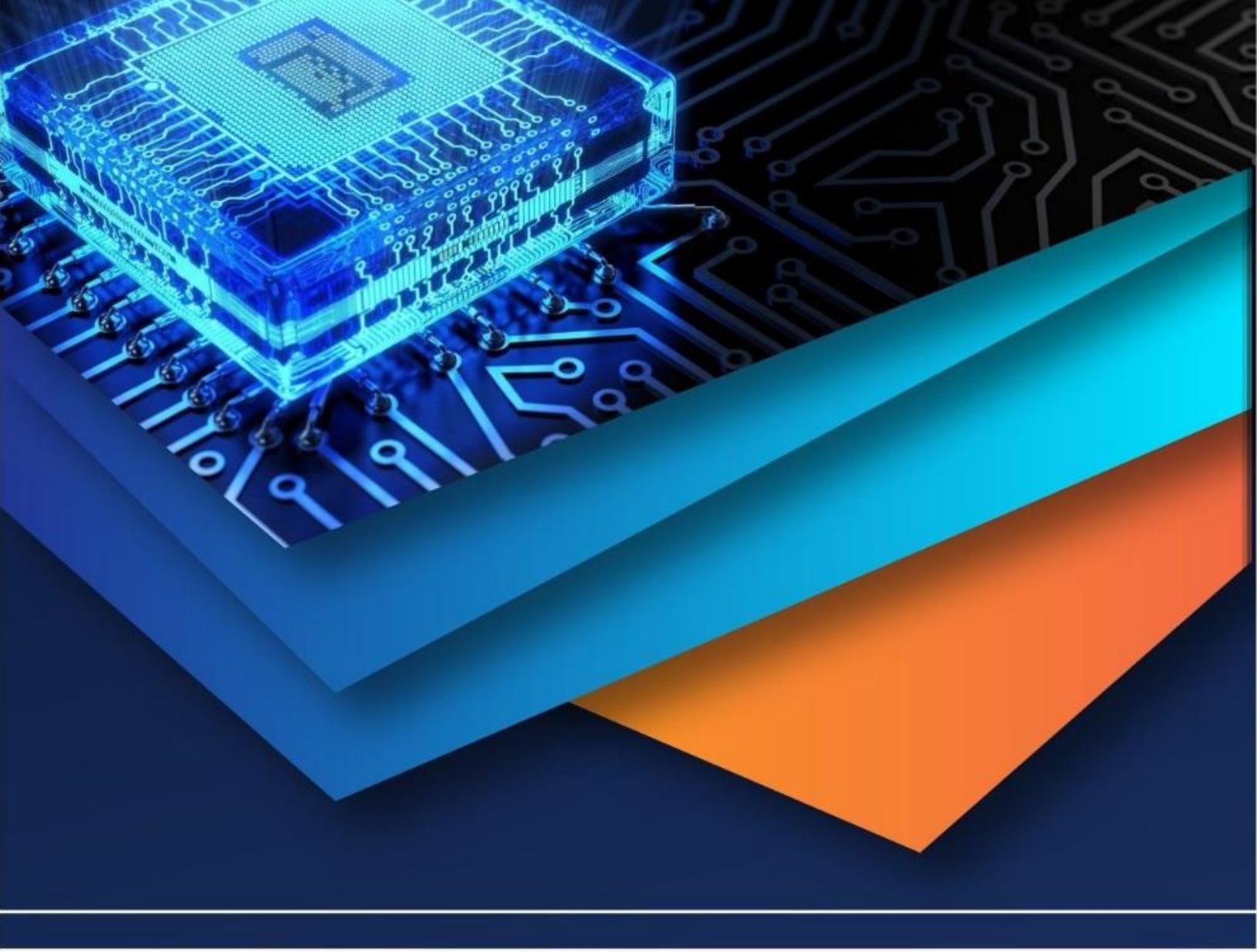
[8] Xianling Y,Wei Z. An Improved Recognition Methods of Web Mining. Huazhong University of Science and Technology (Natural Science), 2006, 34(7):33-35.

[9] Chen M S, Park J S, Yu P S. Data mining for path

IJRASET: ISSN: 2321-9653

traversal patterns in a
web environment[C] IEEEProc. Of 16th Internation
Conference Distributed Computing System, 1996:385-392.
[10] Fu Y, Sandhu K, Shih M. A generalization-based approach to clustering of web usage session[C] Proc. Of KDD Workshop Web Mining, LNCS1863, Springer-Verlag, 2000: 21-28.
[11] Facca M, Lanzi P L. Mining interstiong knowledge from web log.Data and Knowledge Engineering, 2005, 53(3):225-241.
[12]Yuquan Z,Hebiao Y,Lei S. Data Mining Technology. Nanjing: Southeast China University Press, 2006:108-111.

IJRASET: ISSN: 2321-9653



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)