October 2012

AN INDISCERNIBILITY APPROACH FOR PRE PROCESSING OF WEB LOG FILES

JEEVA JOSE

1Department of Computer Applications, BPC College, Piravom, Kerala, India, vijojeeva@yahoo.co.in

P. SOJAN LAL

School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India, padikkakudy@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcct>

Recommended Citation

JOSE, JEEVA and LAL, P. SOJAN (2012) "AN INDISCERNIBILITY APPROACH FOR PRE PROCESSING OF WEB LOG FILES," *International Journal of Computer and Communication Technology*. Vol. 3 : Iss. 4 , Article 2.

DOI: 10.47893/IJCCT.2012.1147

Available at: <https://www.interscience.in/ijcct/vol3/iss4/2>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

AN INDISCERNIBILITY APPROACH FOR PRE PROCESSING OF WEB LOG FILES

JEEVA JOSE¹ & P. SOJAN LAL²

¹Department of Computer Applications, BPC College, Piravom, Kerala, India

²School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India

E-mail : vijojeva@yahoo.co.in¹, padikkakudy@gmail.com²

Abstract - World Wide Web has a spectacular growth not only in terms of the number of websites and volume of information, but also in terms of the number of visitors. Web log files contain tremendous information about the user traffic and behavior. A large amount of pre processing is required for eliminating the noise and is one of the challenging tasks in web usage mining. This paper proposes an indiscernibility approach in rough set theory for pre processing of web log files.

Keywords - *Web Logs, Web Usage Mining, Rough set, Preprocessing.*

I. INTRODUCTION

Web usage mining is the extraction of information from web log files generated when a user visits the web site [1][2][3]. Web usage analysis includes straightforward statistics such as page access frequency as well as more sophisticated forms of analysis like clustering, classification and association rule mining. Access log files contain large amount of HTTP information. Every time a web browser downloads an HTML document from the internet, the images are also downloaded and stored in the log file [4]. A *hit* is any file from a web site that a user downloads and *accesses* are an entire page downloaded by users regardless of the number of images, sounds or movies [5]. A hit can be a text document, image, movie or a sound file. If a user downloads a web page that has 6 images on it, then the number of hits is 7(1 text page + 6 images) whereas the user has accessed just one page of the web site. Most web analysis software counts the number of *hits* a server receives rather than the number of *accesses*. In order to assure that only the data contain useful information for the users, it is absolutely necessary for the log files to be cleaned and filtered.

II. WEB LOG FILES

Web logs are maintained by web servers and contain information about users accessing the site. Logs are mostly stored as simple text files, each line corresponding to one access. The most widely used log file formats are Common Log File Format and Extended Log File Format [6] [7] [8]. Traditionally there are four types of server logs [17].

A. Transfer log

The Transfer log or Access log provides the greatest amount of server data, including the date, time, IP address and user action. The information like percentage of users accessing the site from a specific domain type (.com, .edu etc), the number of unique IP addresses accessing the site, number of accesses

the server receives during specific hours and days of the week and the path a user take through a site. Such analysis enables server administrators to characterize the server load and usage pattern.

B. Agent Log

The Agent log provides data on a user's browser, browser version and operating system. This is a significant information, as the type of browser and operating system determine what a user is able to access on a web site. These data are essential for the design and development of web sites.

C. Error Log

The average web user will receive an error "Error 404 File Not Found" message several times a day. When a user encounters this message, an entry is made in the Error Log. It also contains information about stopped transmission. This occurs when a user click on the "Stop" button while downloading a file. This information can indicate patterns with large files (eg, movies and images) that users consistently stop downloading. The analysis of error log can also provide important server information such as missing files, erroneous links and aborted downloads.

D. Referrer Log

The Referrer Log indicates what other sites on the web link to a particular server. Each link made to a site generates a Referral log entry. Such referral data is critical to alleviate missing link data, path reconstruction etc.

The referrer and agent logs may or may not be turned on at the server or may be added to the transfer log file to create an extended log file format. There are three main sources to get the raw web log file [18]. They are the a) client log file, b) Proxy log file and c) Server log file. Client log files are most accurate and authentic to depict the user behavior but it is a difficult task to modify the browser for each client. In proxy servers, same IP address is used by many users and hence to identify users is difficult. Most researchers consider the web server log file as

most reliable and accurate for web usage mining process. The Web log contains the following information: a) user's IP address b) user's authentication name c) the date-time stamp of the access d) the HTTP request e) the URL requested f) the response status g) the size of the requested file and optionally also h) the referrer URL i) the browser and its version and j) the operating system.

Usually there are three ways of HTTP requests namely GET, POST and HEAD. Most HTML files are served via GET method while most CGI functionality is served via POST or HEAD. The status code 200 is the successful status code. Table 1 shows various status codes of Hyper Text Transfer Protocol [9].

III. DATA CLEANING AND PREPROCESSING

The first pre processing task is data cleaning. The process of data cleaning is to remove noise or irrelevant data. Web server access logs represent the raw data source.

101	Switching Protocols	404	Not Found
200	OK	405	Method Not Allowed
201	Created	406	Not Acceptable
202	Accepted	407	Proxy Authentication Required
203	Non Authoritative Information	408	Request Time Out
204	No Content	409	Conflict
205	Reset Content	410	Gone
206	Partial Content	411	Length Required
300	Multiple Choices	412	Precondition Failed
301	Moved Permanently	413	Request Entity Too Large
302	Moved Temporarily	414	Request URL Too Large
303	See Other	415	Unsupported Media Type
304	Not Modified	500	Server Error
305	Use Proxy	501	Not Implemented
400	Bad Request	502	Bad Gateway
401	Unauthorized	503	Out of Resources
402	Payment Required	504	Gateway Time Out
403	Forbidden	505	HTTP Version not Supported

Table 1 : Status Codes of Hypertext Transfer Protocol

It is important to identify and discard the data recorded by web robots or web crawlers, the images, sounds, java scripts etc that is often redundant and irrelevant [10]. The problems identified in pre processing of log files are

- The need for a large storage space due to considerable volume of data saved on disk.
- The existence of large amount of data that is irrelevant for the web mining process containing images, sounds etc.
- The storage of requests performed by search engines and various automated scripts.
- The storage of data containing error messages such as 400(bad request).

The advantages of pre processing are

- The storage space is reduced as only the data relevant to web mining is stored.
- The search engine requests and image files are removed so that the precision of web mining is improved.

IV. INDISCERNIBILITY RELATIONS AND ROUGH SET THEORY

Indiscernibility relations in rough set theory [11] can be used for the data cleaning of web log files. Rough set is based on the assumption that with every object of the universe of discourse, some information is associated. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. Any set of all indiscernible (similar) objects is called an elementary set and forms a basic granule of knowledge about the universe. Any union of some elementary sets is referred to as crisp (precise) set otherwise the set is rough (imprecise, vague).

Let a given pair $S = (U, A)$ of non-empty finite sets U and A , where U is the Universe of objects and A is the set consisting of attributes. The function $a: U \rightarrow V_a$, where V_a is the set of values of attribute a , called the domain of a . The pair $S = (U, A)$ is called an information system. Any information system can be represented by a data table with rows labeled by objects and columns labeled by attributes. Any pair (x, a) where $x \in U$ and $a \in A$ defines the table entry consisting of the value $a(x)$.

Any subset B of A determines a Binary relation $I(B)$ on U , called an indiscernibility relation defined by $xI(B)y$ if and only if $a(x)=a(y)$ for every $a \in B$, where $a(x)$ denotes the value of attribute a for object x . $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$ will be denoted by $U/I(B)$ or simply U/B . Equivalence classes of the relation $I(B)$ of the partition U/B are referred to as B -elementary sets or B -elementary granules. Let U

represents the set of the set of all user sessions with the path traversed. Let A be the subset of U which represents the HTTP request, URL requested and status code. An indiscernibility relation $I(B)$ is defined for every $a(x)=a(y)$ [12].

A rough set based feature selection for web usage mining is used in [13]. Feature selection is a pre processing step in data mining and is very effective in reducing dimensions

The indiscernibility relations in rough set theory is used for clustering in [14].

V. CONSTRUCTION OF EQUIVALENCE CLASSES FROM WEB LOGS

We consider the subset of web log file entry which includes <HTTP request, URL requested, status code> for indiscernibility relation and is denoted by B. Based on the values of these attributes, the family of all equivalence classes is generated and it is denoted as $U(B)$. Out of the equivalence classes generated, the indiscernibility relation having the attribute values <GET, *.html, 200> and <GET, *, 200> are the equivalence classes which is useful for further data mining to follow. The advantage of this method is that all other equivalence classes generated can be further used to perform other statistical analysis like the percentage of bad requests, mean number of redirects, percentage of reduction in the size of log file etc.

A. Sample Code

1. Extract the fields from a raw log file to a spreadsheet or a data base.
2. Eliminate the request by bots, spiders and crawlers.
3. Generate Equivalence classes considering the attributes <HTTP Requested, URL Requested, Status Code>
4. Filter the equivalence classes with values <GET, *.html, 200> and <GET, *, 200>
5. Eliminate all other fields excluding IP, date, time and URL requested.
6. All other equivalence classes can be saved in another file for statistical analysis if required.

We have used raw log file of a business organization called NeST ranging from January 1, 2011 to March 31, 2011. A sample output after cleaning and pre processing is shown below in Table 2.

IP	DATE	TIME	URL REQUESTED
117.196.136.242	01-Jan-11	00:45:59	/
117.196.136.242	01-Jan-11	00:46:19	/index.php?option=com_content&task=view&id=70&Itemid=72
219.64.67.45.hyd-cdma.dialup.vsnl.net.in	01-Jan-11	00:55:02	/nestit/

219.64.67.45.hyd-cdma.dialup.vsnl.net.in	01-Jan-11	01:05:05	/nestit/services.html
host-196.218.41.130-static.tedata.net	01-Jan-11	01:16:34	/
115.113.10.228.static-bangalore.vsnl.net.in	01-Jan-11	02:33:21	/index.php?option=com_content&task=view&id=6&Itemid=9
91.201.66.28	01-Jan-11	02:34:58	/
91.201.66.135	01-Jan-11	03:33:18	/
hosted-by.altushost.com	01-Jan-11	07:17:28	/index.php

Table 2 : Cleaned Web Log in Common Log File Format

B. User and Session Identification

A time out of 30 minutes is used to identify a user session [15]. Each request from the same IP before 30 minutes is considered as the same user. If the log file is an extended log file, the browser and its version, operating system etc are also considered for identifying users and sessions. After user session identification, if the task to be performed is association rule mining or frequent itemset mining, the log file is further pre processed to obtain records in the form $\langle T_{id}, p_1, p_2, \dots, p_n \rangle$, where T_{id} represents the transaction id and $p_1, p_2, \dots, p_n$ represents the pages requested. Table 3 shows the statistics of the weblog data before and after preprocessing.

Total number of records	8,68,598
Number of records after removing search engine visits	7,75,738
Number of records after pre processing	60,606
Percentage in reduction	93.02%
Total number of users	25,015

Table 3 Statistics of web log data before and after pre processing

C. Prefetching and Caching

If the mining of web logs is performed for prefetching and caching of web pages in proxy servers, then the pre processing is done in a different way. The web objects like images, sounds, movies can not be eliminated. References to embedded objects are usually preceded by their HTML container [16]. If the status code is successful, then the number of bytes transferred is important for caching and prefetching purposes.

D. Sample Code

1. Extract the fields from a raw log file to a spreadsheet or a data base.
2. Eliminate the request of bots, spiders and crawlers.
3. Generate Equivalence classes considering the attributes <HTTP Requested, URL Requested, Status Code>

4. *Filter the equivalence classes with values <GET, * 200>*
5. *Compute the total bytes transferred including the embedded objects for each container page.*
6. *Eliminate all other fields excluding container page and total bytes transferred.*
7. *Calculate the frequency of each container page.*

VI. CONCLUSION

Rough set theory has found applications in many domains like engineering, environment, banking, medicine and others. This paper is an attempt to highlight another application of rough set theory, which is the pre processing of web log files. Pre processing of web log files is very crucial as it may bias the data mining tasks to follow. The advantage of this method is that we will get different equivalence classes which can be further used for statistical analysis including prediction. The pre processed file can be used for various web usage mining tasks.

REFERENCES

- [1] M. F. Facca & P. L. Lanzi, "Mining interesting knowledge from web logs: a survey", Data & Knowledge Engineering, 2004, Vol. 53, pp.225-241.
- [2] R.Cooley, B. Mobasher and J. Srivastava, "Data Preparation for Mining World Wide Web Patterns", Knowledge and Information Systems, 1999, Vol 1(1), pp.5-32.
- [3] J.Pitkow, "In Search of Reliable Usage Data on WWW", Sixth International WWW Conference, 1997.
- [4] R.M. Suresh & R. Padmajavall, "An Overview of Data Preprocessing in Data and Web Usage Mining", First International Conference on Digital Management, Bangalore, India, IEEE, 2006, pp.193-198.
- [5] M.C.Burton & B. J.Walther, "A Survey of Web Log Data and Their Application in Use-Based Design", 34th Hawaii International Conference on System Sciences, Maui, Hawaii, IEEE, 2001, pp.1-10.
- [6] M.Grcar, "User Profiling: Web Usage Mining", 7th International Multiconference Information Society, Ljubljana, Slovenia, 2004.
- [7] A. H. M.Wahab, H.N.M.Mohd, F.H.Hanaf & M.F.M.Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology, 2008, pp.190-197.
- [8] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation", Communications of the ACM, 2000, vol. 43(8), pp.127-134.
- [9] Internet: Hypertext Transfer Protocol Overview, <http://www.w3.org/protocols>. Last retrieved October 2011.
- [10] D. Mican & D. Sitar-Taut, "Preprocessing and Content/ Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development", Informatica Economica, 2009, vol. 13(4), pp.168-179.
- [11] Z. Pawlak, "Rough Set Theory and its Applications", Journal of Telecommunications and Information Technology, 2002, pp.7-10.
- [12] Z. Pawlak & A. Skowron, "Rudiments of Rough Sets", Information Sciences, 2007, vol. 177(1), pp.3-27.
- [13] H.H. Inbarani, K. Thangavel & A. Pethalakshmi, "Rough set based Feature Selection for Web Usage Mining", International Conference on Computational Intelligence and Multimedia Applications, Sivakasi, India, IEEE, 2007, 33-38.
- [14] S. Hirano & S. Tsumoto, "An Indiscernibility Based Clustering Method", IEEE International Conference on Granular Computing, Izumo, Japan, IEEE, 2005, pp. 468-473.
- [15] Z. Pabarskaite, & A. Raudys, "A process of knowledge discovery from web log data: Systematization and critical review", Journal of Intelligent Information Systems, 2007, Vol.28, pp. 79-104.
- [16] Qiang Yang, H. H. Zang & Tianyi Li, "Mining Web Logs for Prediction Models in WWW Caching and Prefetching", Knowledge Discovery in Databases, San Fransisco, 2001, pp. 473-478.
- [17] J. C. Bertot, C.R. McClure, W. E. Moen & Jeffrey Rubin, "Web Usage Statistics: Measurement Issues and Analytical Techniques", Government Information Quarterly, Jai Press Inc., 1997, Vol. 14 (4), pp. 373-395.
- [18] T. Hussain, S. Asghar and N. Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File", International Conference on Information and Emerging Technologies, IEEE, 2010, pp. 1-6.