EXTRACTION OF INTERESTING PATTERNS THROUGH ASSOCIATION RULE MINING FOR IMPROVEMENT OF WEBSITE USABILITY

Resul DAŞ [1], İbrahim TÜRKOĞLU [2]

[1] Firat University, Department of Informatics, 23119, Elazig/TURKEY
[2] Firat University, Department of Electronics and Computer Science, 23119, Elazig/TURKEY

[1] E-mail: rdas@firat.edu.tr   [2] E-mail: iturkoglu@firat.edu.tr

ABSTRACT

In this paper, a new approach for pre-processing stage of the web usage mining is proposed. The proposed method provides more successful results when dealing with raw log files. Moreover the complexity of the proposed method is considerably low. The related areas in the log files are determined for mining stage. Then, by using the association rules method the interesting patterns are extracted for consequent goals. In the result of this study, knowledge extraction was performed by proposed method. The significant and pure knowledge is contributed to structural organization of website. Besides, web site development and about the availability of the site to give advice to managers and designers are available.

Keywords: Web usage mining, Web user access logs, Association rule mining, Knowledge extraction.

1.   INTRODUCTION

Websites are one of the most important advertisement tools in international area for universities and other foundation. Therefore, content and design of web pages are very significant for web designers. Analyzing of the web user access log files can help understand the user behaviours and web structure, there by improving the design of web components and web applications.

In recent years, with the rapid growth of the amount of information on the web, most research in web mining has concern on web usage mining. Web usage mining techniques have been widely used for discovery of interesting usage patterns from access log files [1]. Soft computing methods (neural networks, fuzzy logic, genetic algorithms, and rough sets, etc.) have been intensively used in web usage mining studies. Some of them are described in [2-5] and related work in section 2.1.

Araya et al. [2] provides such a methodology that is based on suggestions from literature and own experience from various web mining projects. Its application in a Chilean Bank has shown how a combined use of data from a data warehouse and web data can contribute to improve marketing activities. Liu et al. [3] presented an automatic classification of web user navigation patterns and propose a novel approach to classifying user navigation patterns and predicting users' future requests. Srivasta et al. [5] described data can be collected at the server-side, client-side, proxy servers or a consolidated web database in web mining.

This paper focuses on using association rule analysis method to investigate a database of URL information regarding access to electronic sources. The specific objective is to determine the different rules that describe associations between sets of items in the web server user access logs at Firat University by using SAS Enterprise Miner software. Association rules analysis is used to deal with the URL links. The purpose of this study performs association rules analysis between web links and allows the user to extract sequential association rules among large sets of URL links.

This paper is organized as follows: Section 2 describes several background objects related to web usage mining and the association rules method that is used in this study. The implementation of proposed methodology for Firat University is extensively described in Section 3. Obtained useful results are presented in Section 4. Finally, the conclusion is represented in Section 5.

2. BACKGROUND

2.1 Related works

One of the basic methods of web usage mining is association rules that indicate relationships among common use of web pages. Up to now, many application of web usage mining have been studied based on association rules analysis. Mobasher et al. [6] use mined association rules to realize effective web personalization. Shen et al. [4] propose a three-step approach to mine the most interesting web access associations. Tan and Kumar [7] propose a mining technique called indirect association that is capable of combining the various association patterns into a more compact structure. This new technique can also capture both positive and negative correlations that exist in the data. Yao et al. [8] use mined association rules to build intelligent web agents. Xue et al.

propose a novel re-ranking method based on site logs. In addition, they obtained from web logs each page's access frequency and the traversal patterns of information finding [9].

Cooley et al. did in-depth research to all the procedure of web usage mining. They discuss methods to pre-process the user log data and to separate web page references into those made for navigational purposes and those made for content purposes. They proposed a user browsing behaviour model which assumes that a given user's treatment of each page is either for the purpose of 'navigation' or 'actual content,' and this is determined by the page references and associated time obtained in web server logs [10-11]. Drott [12] explains the various web server logs mining methods to improve site design. Zhu et al. [13] extend this by introducing the maximal forward reference to eliminate the effect of backward references by the user. They also predict user behaviour within the 'n' future steps, using an N-Step Markov chain as opposed to the one step approach by Sarukkai. Information foraging concepts have also been used recently by Chi et al [14] to incorporate user behaviour into the existing content and link structure. They have modelled user needs and user actions using the notion of Information Scent as described earlier.

2.2 Web usage mining

Web usage mining is to analyze web log files to discover user accessing patterns of web pages. In order to effectively manage and report on a website, it is necessary to get feedback about activity on the web servers [15]. Web usage mining performs mining on web usage data or web logs. A *web log* is a listing of page reference data. It is at times referred to as *clickstream* data as each entry corresponds to a mouse click [16].
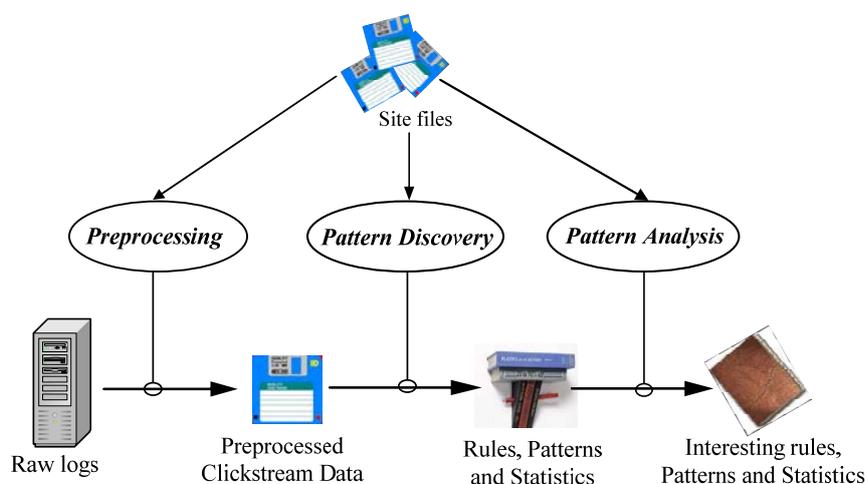

Figure 1. High level web usage mining process [10]

R. DAŞ, İ. TÜRKOĞLU

Fig. 1 illustrates the overall data flow of the high level web usage mining system, which consist of three main tasks for performing web usage mining: *pre-processing, pattern discovery* and *pattern analysis*. Srivasta et al. [5] have given an excellent discussion on the entire web usage mining process.

2.3   Association rules

The purpose of mining association rules is to find out which web pages are usually visited together in a session. Association rules are find interesting associations and correlation relationships among large set of data items. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. For instance, association rule discovery using the Apriori algorithm [17] may reveal a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment [5].

Association rules discovery is based on the sessions for the web usage mining. Each session is interpreted as a transaction and occurrence of the web pages in the session is relevant to the existence of the items in one basket. Additionally, there is possibility of keeping sequence in which web pages were viewed in web usage mining. These web pages sequences are called paths, as the user was following pages in specific order until session ended. When association mining is applied to web usage mining, one can discover correlations among web pages and interesting access patterns. For instance, the problem of finding pages visited together is similar to finding associations among itemsets in transaction databases. Once the user sessions or transactions are identified, one can immediately apply the Apriori algorithm or its variations to discover page association [18].

An association rule is a statement of the form (*item set A* $\Rightarrow$ *item set B*). The rule is defined by *A* and *B* where *A* is the set of antecedent items and *B* is the consequent item. An association rule has two numbers that express the degree of uncertainty about the rule: the *Support* and *Confidence*. *The support* is the number of transactions that include all items in the antecedent *A* and consequent B divided by the total number of transactions. It is the probability that a transaction contains both *A* and *B*. The support for the rule $A \Rightarrow B$ is the probability that the two item sets occur together. The support of the rule $A \Rightarrow B$ is estimated by Eq. (1) [4].

$$\frac{\text{(transactions that contain every item in A and B)}}{\text{(all transactions)}} \quad (1)$$

The support is reflexive. That is, the support of the rule A=>B is the same as the support of the rule B=>A. *The confidence* is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent. It is the conditional probability that if a transaction contains A, it also contains B. The confidence of an association rule $A \Rightarrow B$ is the conditional probability of a transaction containing item set *B* given that is contains item set *A*. The confidence is estimated by Eq. (2) [4].

$$\frac{\text{(transactions that contain every item in A and B)}}{\text{(transactions that contain the items in A)}} \quad (2)$$

In the results of association mining, we can find pairs of groups of web pages "visited" together. Association rules help organizations to develop effective marketing strategies and arrange their websites for better use.

3.   METHODOLOGY AND IMPLEMENTATION

In this study, we present a new approach for pre-processing stage of the web usage mining is proposed. In addition, we present a methodology for rule extraction from the web user access logs. The contributions of the paper include the specification and associated tool that combines the operations of association to derive semantically interesting patterns. These patterns can be used in decision making because the patterns were mined from the user's original hypothesis. In this approach, SAS Base Software 9.1.3 was used (Licence number: 291468). SAS Base Software 9.1.3 includes two different programs [18-19]. These programs are called SAS Enterprise Guide 4.3 and SAS Enterprise Miner 5.2. While SAS Enterprise Guide program 4.3 was used for data pre-processing from the raw logs, SAS Enterprise Miner 5.2 program was used to explore and analyze the association rules from the original database.
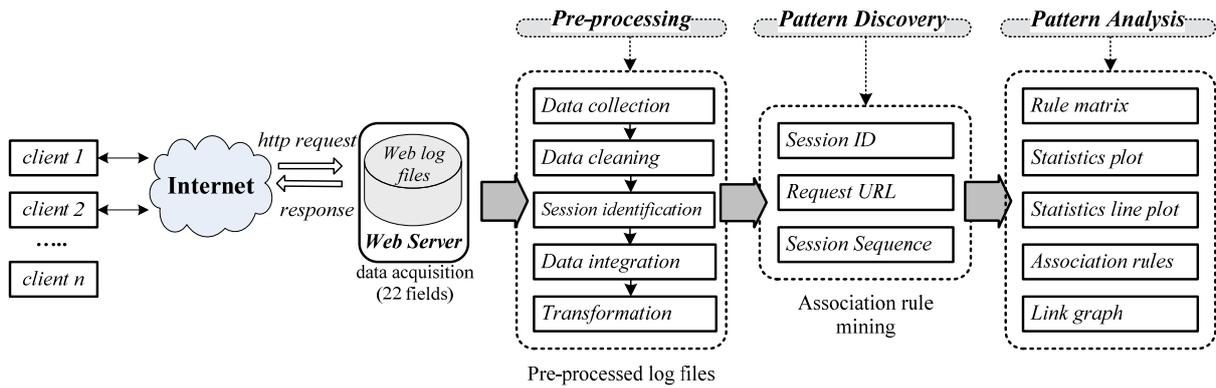
Figure 2. The algorithm scheme of the proposed methodology

Fig. 2 shows the framework of our performed scheme. The brief description of each step is given in the following;

3.1 Pre-processing

Data sources used in this study are the web user access log files which use a well-defined standard format, such as the extended log file format proposed by W3C [15]. In an IIS web server, using Windows 2003 advanced server, a new log file is automatically generated every day. The name of the file is "ex<yymmdd>.log", where *yymmdd* is the date. For example, a log file for January 14, 2008 would be called ex080114.log. Every entry in that file will be a record of successful and unsuccessful web queries on the web server. A three digit code determines the status of the query; a code value of 200 means that the query was successful. Those successful entries are the ones we want to check. Any other code value was an unsuccessful query, be it legitimate or not. We are not concerned with them because if they were hacker attacks, they did not succeed.

An important task in web usage mining application is the creation of a suitable pre-processed usage data set. This process is usually complex and critical to the extract useful information from the log files in web usage mining. Purpose of the pre-processing is to offer a structural, reliable and integrated data source for pattern discovery. Usually, several pre-processing tasks need to be done before performing web mining algorithms on the web server logs, which includes:

Step 1: We removed all the images and style sheet pages as they are a routine part of the initial access. In order to obtain an effective and interesting result, these URLs were filtered from the raw dataset.

Step 2: After the data cleaning process, we applied the *referrer-basic heuristic (h-ref)* method for session identification. The web log entries were partitioned into logical clusters using one or a series of session identification modules.

Step 3: And then, the data integration step looks up the entry's request URL in the mapping table. If the URL is present, the corresponding type label is attached to the entry. Otherwise, the extension of the URL is checked for a valid document type, other than HTML. If the extension was an HTML page, it should be deleted.

Step 4: In this step, the data transformation is realized to obtain desired fields data. We are used to import the data and create three variables: *session id, request file* and *session sequence*. The first lines of formatted data from our dataset are given in Table 1.

Table 1. An example of formatted web data in SAS database

session_id	request_file	session_sequence
67e2a4da826149c5 2008-01-14 08:38:17	/default.asp?id=7	1
67e2a4da826149c5 2008-01-14 08:38:17	/?git=fakulteler	2
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	3
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/dersler_detay.asp	4
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	5
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/foto/igallery.asp	6
67eb39ebce6c73c3 2008-01-14 23:43:23	/default.asp	1
67eb39ebce6c73c3 2008-01-14 23:43:23	/?git=enstituler	2
67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/index.htm	3
67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/dersler.htm	4

67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/computer.htm	5
67cb8704bb370ee2 2008-01-14 13:44:04	/default.asp	1

3.2 Pattern discovery

Pattern discovery is the key process of the web mining, which includes the algorithms and techniques from several research areas, such as data mining, machine learning, statistics and pattern recognition. The techniques such as statistical analysis, association rules, clustering, classification, sequential pattern, dependency modelling are used to discover rules and patterns [5, 20]. Association rules analysis allows the user to determine the paths that visitors take as they navigate through a website. Also, it performs association analysis between web links and allows the user to extract sequential association rules among large sets of web links.

In order to extract patterns of usage from web log files are used data mining techniques for web usage mining. In this phase, association analysis method is applied to analyze pre-processed web log data files. We created and used three variables: *session id, request file* and *session sequence*. The format of these variables is given in Table 1.

3.3 Pattern analysis

The last stage of the overall web usage mining is pattern analysis, as described in Fig. 2. The goal of this process is to extract the interesting rules, unknown patterns or statistics from the pre-processed data stored in web server logs. The pattern analysis phase of web usage mining is one of providing tools to facilitate the transformation of information into

knowledge [15]. The results of our application are given in section 4.

4. EXTRACTED INTERESTING PATTERNS AND ASSOCIATION RULES

In this study, web user access log files acquired from the computer centre of Firat University were used. The dataset was very large so we had to choose only one file out of these data. The pre-processed data set has 146178 rows. The association rule mining contains the underlying data that is used to generate the rule matrix plot, statistic plot, statistics line plot, rules table and link graph. The results of obtained through association rules analysis are given following:

4.1 Rule matrix

Association rules are made up of a left-hand item and one or more right-hand items. The left-hand item is called the left hand of rule. The right-hand items are called the right hand of rule. The rule matrix displays a grid plot of the items in the left hand of rule on the X-axis and the items in the right hand of rule on the Y-axis. A data point in the plot corresponds to a transaction rule.

The rule matrix plots the rules based on the items on the left side of the rule and the items on the right side of the rule. The points are coloured based on the confidence of the rules. The related figure can be seen in Fig. 3.
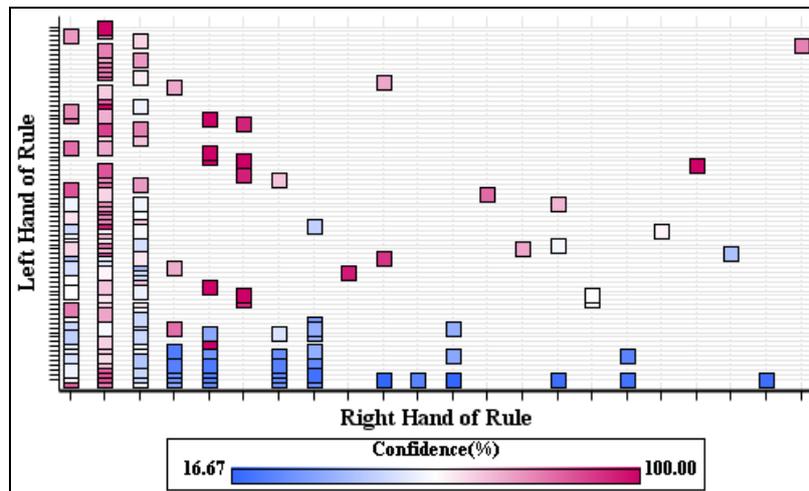

Figure 3. Rule matrix

As shown in fig 3, if confidence values are increase, the number of rules is increase. It is seen that the most of the extracted rules confidence values were below 50%. This situation indicates the unreliability due to the untidiness between web pages of a web site.

4.2 Statistic plot

The statistics plot displays a scatter plot of support (%) on the Y-axis versus Confidence (%) on the X-axis. The points are coloured either by the number of items (association) or the chain size (sequence). The X-axis and Y-axis represent the confidence and the

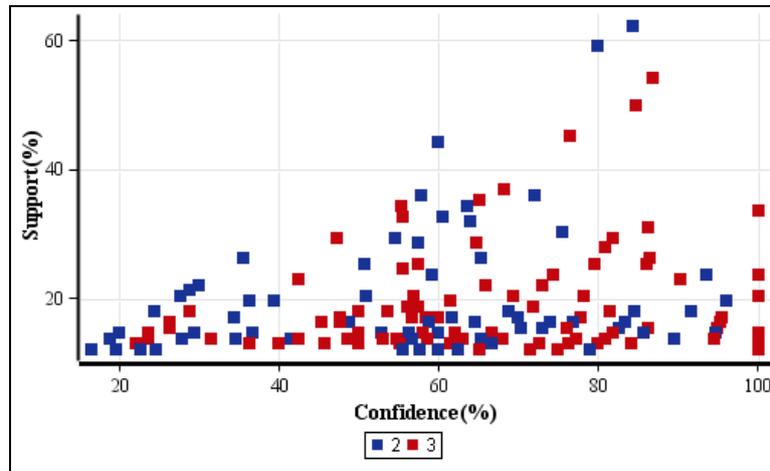support, respectively. Fig. 4 is an example of the statistics plot.

Figure 4. Statistic plot

As shown in Fig. 4, the most of the extracted rules support values were below %40. Consequently, due to low support value of extracted rules for web pages in a web site, reliability of these rules is decrease.

4.3 Statistics line plot

The statistics line plot displays a line plot of various values for each rule on the Y-axis. The X-axis represents the values of Rule Index. The statistics line plot graphs the confidence and the support for each of the rules by rule index number. The percent support is the transaction count divided by the total number of users, which would be the maximum transaction count. The percent confidence is the transaction count divided by the transaction count for the left side of the sequence. For example, consider the rule $A \Rightarrow B$. Recall that the support of $A \Rightarrow B$ is the probability that a user has both A and B. The confidence of $A \Rightarrow B$ is the probability that a user has B given that the user has A. Fig. 5 is an example of the statistics line plot.
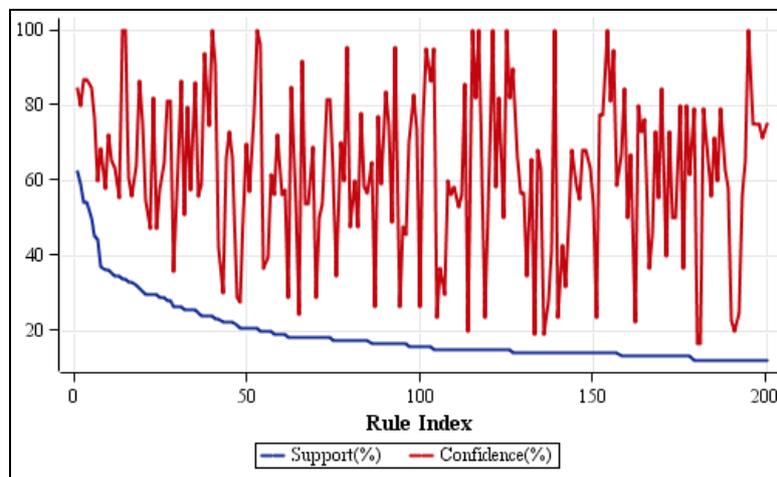
Figure 5. Statistics line plot

The confidence and support values for 200 rules are shown in Fig 5. The rules which have high support value and high confidence value indicate the productivity of the web pages.

4.4 Association rules table

The association rules table displays the information about the association rules that are generated. We used SAS Enterprise miner 5.2 software to mine association rules from a sampling data set. We set up parameters that minimize transaction frequency at 2% where the number of items in longest chain is 3. As a result, we generated 200 association rules from the sampling data set. Table 2 represents association rules table discovered. This table is composed of 12 columns and first 200 rows respectively. The columns of the table are called as the chain length, transaction count, support, confidence, generated association

In the results of association rules mining one can find pairs of groups of items "visited" together. Although very often the above situation is described as the association rules mining, it covers positive association rules discovery. In other words, if two or more pages tend to be viewed together they are connected by positive association rule. As shown in Table 2, the discovered first rule is */sosyalbil/sonsite/index.htm⇒ /sosyalbil/sonsite/ilan.htm.* The confidence of this rule is 84.44 % and the support value is 62.30 %. The chain size of this rule is 2.

4.5   Link graph

The link graph displays association results by using nodes and links. The size and colour of a node indicate the transaction counts in the rules data set. Larger nodes have greater counts than smaller nodes. The colour and thickness of a link indicate the confidence level of a rule. The thicker the links are, the higher confidence the rules have.

This link graph is created using the data residing on the web logs. Its nodes represent the web pages of the site, whereas the link between them the hyperlinks between the pages. This links carry weights, which represent the number of transitions from the "source" web page to the "target" web page. A link graph is a graphical representation of the data in the rules data set. The link graph looks very meaningful, interesting and easy to interpret by the required hyperlinks. Fig. 6 displays a link graph of the association mining analysis. The largest node is for the item */sosyalbil/sonsite/index.htm* because it is the most popular hyperlink that users used.
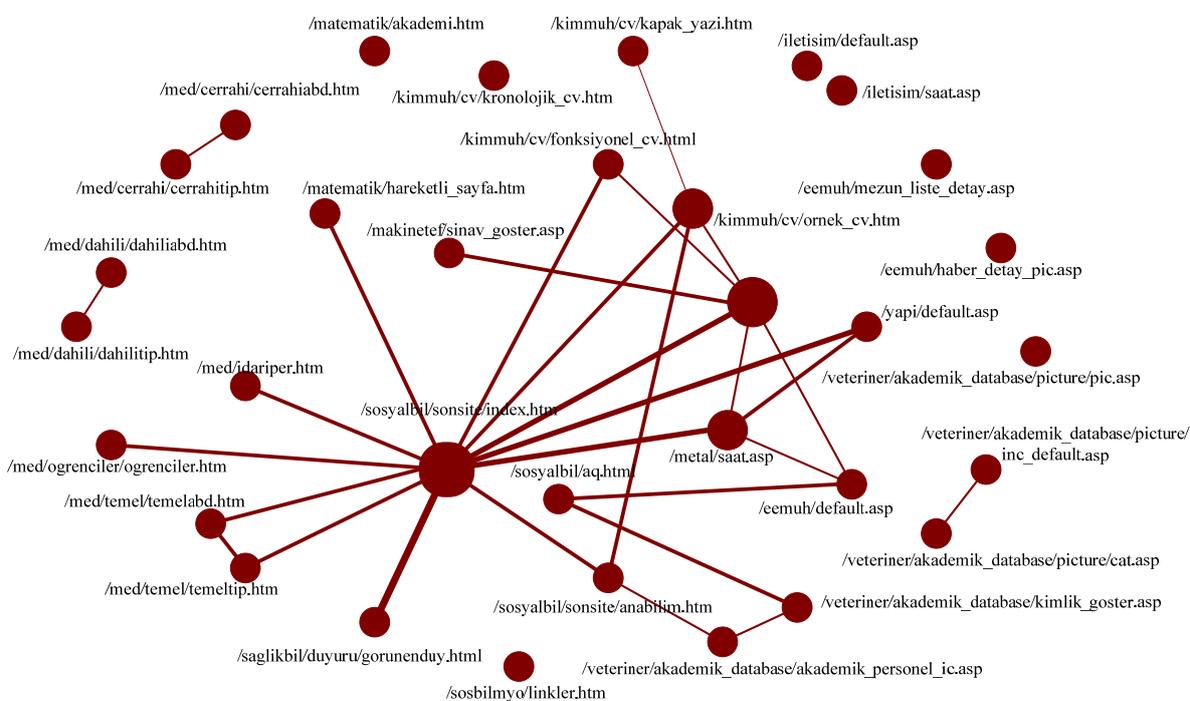

Figure 6. Link graph

5.   CONCLUSIONS

In this study, we have presented a new approach for pre-processing stage of web usage mining. The proposed method provides more successful results when dealing with raw log files moreover the complexity is considerably low.

The goal of this study is to help the web designers to improve their website by determining related link connections in the website. The results and findings of this experimental study can be used by the web designer in order to plan the upgrading and enhancement to the website. The outcome of the research shows that web mining gives decision support for the changes in the web site navigational structure. The followings are brief outcome of the proposed study;

- The support and the confident values of extracted rules are considered for obtaining

the interest of the web visitors. Consequently, the number of hit can be increased by analyzing the visitor attitude.

- The patterns which are extracted from the web links can be analyzed for proving the web site usability and the new interesting ideas can be used for increased the efficiency of the web site.

Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. More research needs to be done in e-commerce, bioinformatics, computer security, web intelligence, intelligent learning, and database systems, finance, marketing, healthcare and telecommunications by using web usage mining.

ACKNOWLEDGEMENTS

REFERENCES

[1] Das, R., Turkoglu, I., Poyraz, M., "Analyzing of system errors for increasing a web server performance by using web usage mining", Istanbul University–Journal of Electrical & Electronics Engineering (IU-JEEE), 7(2), 379-386, 2007, Istanbul.

[2] Araya, S., Silva, M., Weber, R. (2004). "A methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems 148, 139–152.

[3] Liu, H., Keselj, V. (2007). "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests" Data & Knowledge Engineering, 61(2), 304–330.

[4] Shen, L., Cheng, L., Ford, J., Makedon, F., Megalooikonomou, V., Steinberg, T. (1999). "Mining the most interesting web access associations", Proc. The 5th International Conference on Knowledge Discovery and Data Mining (KDD'99), pp.145-154.

[5] Srivasta, J., Cooley, R., Deshpande, M., and Tan, P., "Web usage mining: Discovery and applications of usage patterns from web data" SIGKDD Explorations. 1(2), 12-23, 2000.

[6] Mobasher, B., Dai, H., Luo, T., Sun, Y., Zhu, J. (2000). "Combining web usage and content mining for more effective personalization", Proc. The International conference on E-Commerce and Web Technologies.

[7] Tan, P., Kumar, V., "Mining indirect associations in web data", LNAI 2356 (Springer, 2002) pp. 145-166.

[8] Yao, Y.Y., Hamilton, H.J., Wang, X. (2000). "PagePrompter: an intelligent web agent created using data mining techniques", Technical Report, CS-2000-08, Department of Computer Science, University of Regina.

[9] Xue, G.R., Zeng, H.J., Chen, Z., Ma, W.Y., Lu, C.J. (2002). "Log mining to improve the performance of site search", Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02), p.238.

[10] Cooley, R., Mobasher, B., Srivastava, J. (1999). "Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information Systems, 1(1), 1-27.

[11] Cooley, R., Mobasher, B., Srivastava, J. (1999). "Grouping web page references into transactions for mining World Wide Web browsing patterns", Journal of Knowledge and Information Systems, 1(10), 1-13.

[12] Drott, M. (1998). "Using web server logs to improve site design", Proc. ACM Conference on Computer Documentation, 43-50.

[13] Zhu, J., Hong, J., Hughes, J.G. (2002). "Using Markov Chains for Link Prediction in Adaptive Websites", In Proc. of ACM SIGWEB Hypertext.

[14] Chi, E.H., Pirolli, P., Chen, K., Pitkow, J. (2001). "Using information scent to model user information needs and actions on the web", In Proc. of ACM CHI 2001 Conference on Human Factors in Computing Systems, pp. 490-497, ACM Press, Seattle, WA.

[15] Das, R., Turkoglu, I. (2009). "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method", Expert Systems with Applications, 36(3), 6635-6644.

[16] Dunham, M.H., (2003). Data mining: Introductory and advanced topics, ISBN-13: 9780130888921, published by Prentice Hall, 315 pp.

[17] Agrawal, R., Imielinski, T., & Swami, A.N. (1993). "Mining association rules between sets of items in large databases", ACM SIGMOD'93 International Conference on Management of Data, 207-216, Washington.

[18] Das, R., Turkoglu, I., Sengur, A. (2009). "Diagnosis of valvular heart disease through neural networks ensembles", Computer Methods and Programs in Biomedicine, 93(2), 185-191.

[19] Das, R., Turkoglu, I., Sengur, A. (2009). "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, 36(4), 7675-7680.

[20] Kosala, R., Blockeel, H. (2000). "Web mining research: a survey", SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1), 1–15.

Resul DAS was born in Elazig, Turkey, 1975. He received his BS and M.S. degrees in Department of Computer Science from Firat University, Turkey in 1999, 2002 respectively. He received Ph.D degree in Electrical-Electronics Engineering from Firat University, in 2008. Now, he is working as an instructor in Department of Informatics at Firat University. His research interests are Knowledge Discovery, Web Mining, Computer Networks and Network Technologies, Distance Education Technologies.

Ibrahim TURKOGLU was born in Elazig, Turkey, 1973. He received the B.S., M.S. and Ph.D. degrees in Electrical-Electronics Engineering from Firat University, Turkey in 1994, 1996 and 2002 respectively. He is working as an assistant professor in Electronics and Computer Science at Firat University. His research interests include artificial intelligent, pattern recognition, intelligent modeling, radar systems and biomedical signal processing.