Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 1, January 2017, pg.77 – 83

Mining Weblogs – A Survey

Abha Narwal[1], Dr. R. K. Chauhan[2]

[1]Research Scholar, Kurukshetra University, Kurukshetra, India

[2]Professor, Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India

[1] abha.narwal.kuk@gmail.com; [2] rkchauhankuk@gmail.com

Abstract— Web Usage Mining is the area of Web Mining that deals with the discovery of potentially useful usage patterns from web usage data produced by web servers. This paper presents the developments in this area discussing several aspects being worked on so far

Keywords— Web Mining, Web Usage Mining, Web Log Mining

I. Introduction

The World Wide Web is source of huge data. This data could be the Web content, incorporates billions of pages publically available on the web, along with the Web usage data, incorporates the user access patterns collected and stored as web logs daily by all the web servers. Web Mining [47] initially evolved from data mining as it used data mining techniques to automatically discover and extract knowledge from web documents and services. More precisely, web Mining can be classified into three categories: web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

Web Content Mining focuses on discovering useful information from the web page contents; basically the source data consist of hypertext documents mainly containing textual information. It can be thought of extending the job done by the basic search engines [48]. Typical applications are content-based categorization and content-based ranking of the Web pages. Web Structure Mining is the area of Web Mining that focuses on discovering and modeling the link structure of the Web [6]. It aims to generate structural symmetry about websites and Web pages. Source data basically contains any structural information present in web pages, e.g., links to other pages. Typical applications are link-based categorization of Web pages, ranking of web pages through a combination of content and structure [16], and reverse engineering of web site models. Web Usage Mining focuses on mining the web server logs to extract knowledge about user access patterns to understand user behavior in interaction with web. Source data mainly consists of the logs collected when user interacts with the web servers, having standard formats namely: Common Log Format[50], Extended Log Format[51], LogML[50].Typical applications are those based on user modeling techniques e.g., Web Personalization, adaptive Web sites and user modeling. The main aim is to get the useful users' access information in logs in order to identify more potential customers, page popularity etc., so the sites can ultimately improve themselves with appropriate user requirements [7].

This paper is a survey of the recent developments in the area of Web Usage Mining. The focus is only on Web Usage Mining rather than Web Mining, Web Content Mining and Web Structure Mining. This survey is based on the research results reported specifically in the literature since 1995 and onwards. The paper is organized as follows. Initially, in section 2, we discuss the data sources that collect data from user navigation, for the mining purpose. It may contain data from server access logs, referrers log (containing information about the referring pages for each page reference), agent logs, client side cookies, etc [16, 22]. Section 3 explains the preprocessing of the collected web log data where it is filtered so as to be used for various purposes. Sections 4 and 5, respectively, contain the techniques for Web Usage Mining, and its applications. Section 6 discusses privacy issues raised by accurately tracking of user behavior through Web Usage Mining.

II. Data Sources

The data for Web Usage Mining is collected mainly from three sources:
(i) Web servers, (ii) proxy servers, and (iii) Web clients.

A. Web Servers

Web servers are the main and the richest source of data. They collect huge amount of navigation information in their huge databases as Web log files. The log files have standard formats e.g. Common log Format [50], Extended Log Format [51], LogML [50]. These logs are basically text files containing the basic information needed for the purpose of Web Usage Mining like name, IP address of the remote host, date and time of request, and even the exact request that was sent by the client. Sometimes, in case of massive log repositories instead of text files we use databases which helps improve the querying the log [15, 22].

The main issue in extracting log information from the web servers is the identification of users' sessions. All the users' clickstreams need to be grouped in order to clearly identify the paths followed by user while navigating through the website. Although, several types of information present in log files is helpful, it's still a difficult job. Commonly cookies are used to track the sequence of users' clickstrems [42]. If cookies are not available, several heuristics [23] can be employed to reliably identify users' sessions. However, in actual practice it is impossible to track the exact navigation paths since the use of back button is not tracked at the server level. Techniques to counter these problems are discussed in section 3.

Instead of Web logs, users' activities can also be tracked down by sniffing the TCP/IP packets. Although using this users' sessions cannot be identified exactly, still using packet sniffers is helpful in many ways [16] like: i) data are collected in real time; ii) information coming from different web servers can be easily combined into a single log; iii) use of special buttons, like the stop button, can be detected so to gather the information that is usually missing in log files. In spite of these advantages, packet sniffers are rarely put into practice as they raise scalability issues on web servers with high traffic [16], plus encrypted packets used in secured commercial transactions, cannot be accessed by them. Clearly, application of web usage mining to e-business suffers from this limitation severely [39].

The best approach for tracking web usage is probably, directly accessing the server application layer [32], which is not always possible as there may be issues related to the copyright of server applications. Plus the web usage mining application thus used must be server-specific so as to fulfil the specific tracking requirements.

B. Proxy Servers

Proxy Servers are provided to customers by their Internet Service Providers (ISPs) in order to improve navigation speed through caching. Even at proxy level usage data is collected in a similar way as that at the server levels, in many respects except from the fact that at this level data of groups of users accessing huge group of web servers is collected. Again, not all users' access paths can be recognized and the session reconstruction is complicated, due to caching between the proxy server and the clients.

C. Web Clients

At Web Clients, usage data can be tracked by using Javascript, Java applets, or by modified browsers [16]. These techniques do avoid the problem regarding users' navigation path identification and those caused by caching, by giving access to more in depth information about the actual user behavior [31, 40], but they, clearly, depend heavily on the user cooperation and might concern them about their privacy as they raise issues regarding strict privacy laws which is discussed in section 6.

III.Information Preprocessing

The prerequisite step to any type of usage mining is the identification of a set of server sessions from the raw usage data, in order to know the exact accounting of each user accessed what pages of the website and for what duration limited. Although it is a fundamental step in web usage mining yet literature on it is quite limited. Preprocessing [40] processes the usage, content and structure data obtained from several data sources into various data abstractions like user, page file (file served to the user), server session, [49] episodes(subset of pageviews from a single server session) etc.

Preprocessing consists of four steps: i) data cleaning, ii) users' session identification and reconstruction, iii) retrieving information about the page content and structure, finally iv) data formatting [16].

A. Data Cleaning

It involves merging logs from multiple servers and eliminating all the data irrelevant for mining purpose like the requests for graphical page content, any other files that are embedded into the webpage, common scripts such as count.cgi, and navigations performed by bots and web spiders. Former are easy to remove while robots and web spider navigation patterns are to be identified explicitly by referring to remote hostname or user agent, or by checking access to the robots.txt file, but some robots send a false user agent in HTTP request, needing heuristics based on navigational behavior to identify robot sessions from genuine users' sessions[46]. The proposed heuristics are based on earlier assumptions and a classification of navigations. The classifier is trained using well known navigation paths of robots and the resultant model is used for the classification of the navigational sessions without even having prior information about their user agent. However, a probabilistic model analysis algorithm namely, EM algorithm, based on Probabilistic Latent Semantic Analysis (PLSA) model is applied to the integrated usage data to infer the latent semantic factors as well as generate user session clusters for revealing user access patterns. Experiments have been conducted on real world data set to validate the effectiveness of the proposed approach.

B. Users' session identification and reconstruction:

It follows the steps (i) user and session identification from weblog data, and (ii) path completion within the previously identified sessions to reconstruct users' actual navigation path. It highly depends on the quality and quantity of data present in the web logs [41]. This task is made greatly complicated by the by the presence of local caches, corporate firewalls, and proxy servers. Proxy caching can be solved partially by using cookies [24], URL rewriting [33], or by making users enter a website through log in [34]. Cookies are source of a lot of information about the user, from which those can act as session identifiers are the ones we are interested in. However, cookies may also fail in cases like some browsers do not support cookies, and others allow their users to disable cookies. URL rewriting can be used, in such situations, by embedding the session id in all those URLs which will be written back to the browser; so whenever a user clicks on a link in a webpage, this rewritten URL is sent to the server and hence get stored in the weblog.

Path completion [23,25] is identifying the significant user accesses that are not recorded in the access log due to caching etc. It involves guessing cached navigation paths in the users' session on the basis of referring information from the log. SurfAid [16] introduced by IBM, to solve both proxy and web caching issues, has a javascript named Web Bug which has to be included in each page, asking server for a 1x1 pixel image generating parameters to identify that webpage and the overall process cannot be cached neither by proxy nor by the browser but logged by the server, hence solving the caching caused problem [16]. At a further level of granularity, users' session can be filtered by eliminating the episodes (very small server sessions) and low-support URI references (URIs which do not appear in a sufficient number of sessions) [40].

The sessionization problem arises for not knowing when to terminate session for a user, which is virtually impossible since the HTTP is stateless. Three heuristics for session termination identification have been discussed by [23]; two depending on duration between users' page requests, and one on the information about the referrer. Apart from these, we have Adaptive time out heuristic [16], setting different threshold for time oriented heuristics as suggested by [47].

C. Content and Structure retrieval

WUM applications mainly use visited URLs as the major source of data for mining but they do not suggest anything regarding actual page content. Experiments prove employing information regarding content and structure of the web site improves the effectiveness of the pattern analysis during preprocessing [21]. The web log data can be enriched with content based information [48], by introducing an additional categorization step in which web pages are categorized based on their content type. Even if adequate classification is unknown in advance, Web Structure Mining techniques can be employed to develop one. It classifies web pages based on their semantic areas and is used improve the information extracted from logs. Semantic Web can also be used for WUM where pages are mapped onto ontologies to add meaning to the frequently observed paths [24]. In this task, automatic extraction and classification of objects of different types into classes based on underlying domain ontologies, may be involved. There can be pre-specified domain ontologism or they may be learned automatically from available training data. Instead of user navigation paths, concept-based paths [33] can be used as they represent usual paths generalized over common concepts by means of intersection of raw users' path and similarity measures. An alternative way is to use information scent [34] for better user modeling, which is defined as " the imperfect, subjective perception of the value and cost of the information sources obtained from proximal cues, such as Web links, or icons representing the content sources".

D. Data Formatting

Once all previously mentioned steps for preprocessing have been applied to the server log, session data needs to be properly formatted according to the types of web usage mining technique to be employed. Data has been stored in many formats like, into relational database using a click fact schema [42], signature tree [22], WAP-tree

[43], FBP tree [25], cube structure [35], Frequent Link and Access Tree (FLaAT) [5] etc. in order to improve the extraction of frequent patterns.

[1] proposes a methodology not only preprocesses the data but offers a few advanced features in the form of statistics such as the elapsed time taken in obtaining the output, hit count by respective IP and above all it exports the preprocessed data to a file which can be easily imported in any data mining tool. [9] proposes a preprocessing methodology based on hierarchical clustering.

IV. WEB USAGE MINING TECHNIQUES

Several knowledge discovery techniques have been developed exclusively designed for the purpose of web usage mining. Most researches revolve around four paradigms: i) Association Rules, ii) Sequential Patterns Rules, iii) Classification, and iv) Clustering. A complete reference is provided in [36].

Association Rule Mining: It is one of the fundamental data mining techniques and most used web usage mining technique. Association rules are implications of the form X ⇒ Y where the rule body X and the rule head Y are set of items within a set of transactions. The rule X ⇒Y states that the transactions which contain the items in X are likely to contain also the items in Y. When applied to Web Usage Mining, association rules are used to find associations among Web pages that frequently appear together in users' sessions. The typical result has the form: "A.html, B.html ⇒ C.html",

which states that if a user has visited page A.html and page B.html, it is very likely that in the same session the same user has also visited page C.html. A modification in Apriori algorithm [36] produces such results. To evaluate the association rule mined from web usage data, a measure of interest can be calculated [30]. Ref. [37] exploits a mixed technique of association rules and fuzzy logic to dig out fuzzy association rules from weblogs. [17] For identifying the frequent sequence patterns (itemsets) from the pre-processed cache logs in linear time, association rules can also be used to mine for web access patterns of users using a variation of the Apriori algorithm[36].

Sequential Pattern Rules: Sequential patterns are used to discover maximal frequent subsequences from large amount of sequential data in appearing in users' session. The typical sequential pattern has the following form [16]: the 70% of users who first visited A.html and then visited B.html afterwards, have also accessed page C.html in the same session. Sequential patterns are similar to association rules to an extent that algorithms to extract association rules can also be used for sequential pattern mining. But sequential patterns include the concept of time, i.e., at which point of the sequence a certain event happened, while in association rules mining, information regarding the event sequence is not considered. Sequential analysis can be used to reveal four unique web navigation behavior categories, namely search-information browsing, social-information browsing, ecommerce-information browsing, and direct browsing [2]. Basically, two categories are used for association rule mining: one based on the association rules, and other based on tree – structure and Markov chains to represent the users' access patterns. Few well-known Association rule mining algorithms have been modified to mine sequential patterns e.g. AprioriAll [30] and GPS [16] are two extensions of Apriori algorithm [36]. Another algorithm [43] represents navigation patterns in the form of tree known as WAP-tree, for mining web access patterns and surpassed the above mentioned algorithms. Many extensions of WAP tree have been proposed and are doing well like [19] proposes parallel implementation of WAP-tree mining algorithm, [11] uses layer coded Breadth-First linked WAP-tree for mining maximal sequential patterns, [39] proposes a WAP-tree mining algorithm namely, DLT-mine (Doubly Linked Tree algorithm), to efficiently find all access patterns that satisfy user specified criteria. FreeSpan algorithm [45] focuses on the integration of frequent sequence mining while PrefixSpan [20] uses an approach based on data projection. [10] provides a survey on sequential pattern finding.

Clustering: This technique groups the similar items from huge amount of data based on a general idea of distance function that is a measure of similarity between the groups. It is widely used for web usage mining [35, 8, 14]. [3] combines clustering with fuzzy logic for enhanced usage mining, [12] applying fuzzy clustering to identify target group. [31] combines association rule mining and clustering into a method called association rule hypergraph partitioning. First, association rules are used to extract frequent patterns from user sessions; then the frequent patterns are used to build a graph in which: (i) nodes are the visited Web pages, (ii) edges connect two or more nodes if there is a frequent pattern which contains the pages represented by the nodes; (iii) edges are weighted depending on the relevance of patterns connecting the nodes.

V. APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize Web users). This information can be exploited later to improve the Web site from the users' viewpoint. The results produced by the mining of Web logs can used for various purposes [7]: (i) to personalize the delivery of Web content; (ii) to improve user navigation through pre-fetching and caching; (iii) to improve Web design; or in e-commerce sites (iv) to improve the customer satisfaction.

Personalization of web content. Web Usage Mining techniques can be used to provide personalized Web user experience. For instance, it is possible to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [47,31,38,30]. Personalized Site Maps [40] are an example of recommendation system for links [42]. Ref. [41] proposed an adaptive technique to reorganize the product catalog according to the forecasted user profile. An approach to integrate domain ontologies into the personalization process based on Web Usage Mining is proposed in [42], including an algorithm to construct domain-level aggregate profiles from a collection of semantic objects extracted from user transactions. A survey on existing commercial recommendation systems, implemented in e-commerce Web sites, is presented in [33].

Pre-fetching and caching. The results produced by Web Usage Mining can be exploited to improve the performance of Web servers and Web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time, as done in [34 – 36, 39, 47].

Support to the design. Usability is one of the major issues in the design and implementation of Web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of Web applications. Ref. [49] uses stratograms to evaluate the organization and the efficiency of Web sites from the users_ viewpoint. Ref. [39] exploits Web Usage Mining techniques to suggest proper modifications to Web sites. Adaptive Web sites represents a further step. In this case, the content and the structure of the Web site can be dynamically reorganized according to the data mined from the users' behavior [30,41].

E-commerce. Mining business intelligence from Web usage data is dramatically important for ecommerce Web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure[42,37,43].

VI. PRIVACY CONCERNS

As Web usage mining is being used for predicting users' behavior by gathering sensitive information about them from several sources, it is likely to raise privacy concerns. Many countries, European Union, and the United States are having strict laws about privacy [16]. The point regarding users' privacy was first discussed in [41] as a relevant and sensitive issue. [44] raised privacy issues linked to Web Personalization. A proposal to deal with these issues on web is P3P (Platform for Privacy Preferences) [26]. Its purpose is to enable websites to express their privacy practices in a standardized format that can be automatically retrieved and interpreted by the user agents. It enables Web users to understand what data will be collected by sites they visit, how that data will be used, and what data/uses they may "opt-out" of or "opt-in" to. But P3P does not completely solve the issue as it does not provide any mechanism to ensure that visited websites will actually act according to their declared policies, plus it does not address the issues raised by web mining techniques [16]. Web Usage Mining uses user profiling which is very important for e-business applications, and uses sensitive information about users' online behavior and activities. Research are being done to tackle these privacy issues so as to collect effective information of users appropriate for usage mining without accessing the information that violates users' privacy. Working with this approach, [44, 27] focus on decision tree techniques and, [18, 28, 29] on the Association rules. Solutions are yet to be proposed for the privacy issues from web usage mining point of view.

VII. CONCLUSIONS

This paper presented a survey of the recent developments in the area of Web Usage Mining. The data for Web Usage Mining is collected through several data sources as user navigation is recorded in several logs at server side as well as client side. Preprocessing process consists of: data cleaning, identification and the reconstruction of users_ sessions, the retrieving of information about page content and structure, and the data formatting. Several knowledge discovery techniques designed for analysis of web usage data revolve around four paradigms: Association Rules, Sequential Patterns Rules, Classification, and Clustering. The usage patterns obtained from Web logs can be used for various purposes: to personalize the delivery of Web content; to improve user navigation through pre-fetching and caching; to improve Web design; or in e-commerce site, to improve the customer satisfaction etc. Predicting users' behavior by gathering sensitive information through Web Usage Mining raises privacy concerns. Research are being carried out to tackle these privacy issues so as to collect effective information of users appropriate for usage mining without accessing the information that violates users' privacy.

ACKNOWLEDGEMENT

This study was supported by Department of Computer Science and Application at Kurukshetra University, Kurukshetra.

REFERENCES

[1] Neha Goel, C.K.Jha, "Preprocessing Web logs: A Critical phase in Web Usage Mining", Proceedings of the International Conference on Advances in Computer Engineering and Applications (ICACEA), 2015, 672-676.

[2] Qiqi Jiang et.al., "Using Sequence Analysis to Classify Web Usage Patterns across Websites", Proceeding of the 44th Hawaii International Conference on System Science (HICSS), 2012, 3490 – 349.

[3] Nayana Mariya Varghese; Jomina John, "Cluster optimization for enhanced web usage mining using fuzzy logic", World Congress on Information and Communication Technologies (WICT), 2012, 947 – 951.

[4] K. Sharma, G. Shrivastava, V. Kumar, “Web Mining: Today and Tomorrow”, Proceedings of 3rd International Conference on Electronics Computer Technology (ICECT), Vol. 1, 2011, 389-393.

[5] Cuifang Chen, “Discovery of User Preferred Access Patterns from Web Logs”, Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 1, 2011, 409 – 413.

[6] B. Singh, H. K. Singh, “Web Data Mining Research: A Survey”, Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2010, pp. 1-10.

[7] B. Singh, H. K. Singh, “Web Data Mining Research: A Survey”, Proceeding of IEEE on Computational Intelligence and Computing Research (ICCIC), 2010, 1-10.

[8] Manish Joshi; Pawan Lingras; Yiyu Yao; C. Bhavsar Virendrakumar, "Rough, fuzzy, interval clustering for web usage mining", Proceeding of the 10th International Conference on Intelligent Systems Design and Applications, 2010, 387 - 392.

[9] Tasawar Hussain; Sohail Asghar; Simon Fong, "A hierarchical cluster based preprocessing methodology for Web Usage Mining", Proceedings of the 6th International Conference on Advanced Information Management and Service (IMS), 2010, 462-467.

[10] Naseer Ahmed Sajid; Salman Zafar; Sohail Asghar, "Sequential pattern finding: A survey", Proceedings of the International Conference on Information and Emerging Technologies (ICIET), 2010, 1-6.

[11] Lizhi Liu; Jun Liu, "Mining maximal sequential patterns with layer coded Breadth-First linked WAP-tree", Conference on Computational Intelligence and Industrial Applications (PACIIA 2009), Vol. 1, 2009, 61 - 65.

[12] Jianxi Zhang; Peiying Zhao; Lin Shang; Lunsheng Wang, "Web usage mining based on fuzzy clustering in identifying target group", International Colloquium on Computing, Communication, Control, and Management (ISECS), Vol. 4, 2009, 209 – 212

[13] Qingyu Zhang, Richard S. Segall, “Web Mining: A Survey of Current Research, Techniques, and software”, The International Journal of Information Technology and Decision Making, Vol. 7, No. 4, (2008), pp. 683-720

[14] Chu-Hui Lee, Yu-Hsiang Fu, "Web Usage Mining Based on Clustering of Browsing Features", Eighth International Conference on Intelligent Systems Design and Applications, Vol.1, 2008, 281-286.

[15] K.P. Joshi, A. Joshi, Y. Yesha , “On Using a Warehouse to Analyze Web Logs”, Distributed and Parallel Databases, Vol. 13, No. 2, 2006, pp. 161-180.

[16] F.M. Facca, P.L. Lanzi, “Mining interesting knowledge from weblogs: a survey”, Data and Knowledge Engineering, Vol. 53, 2005, 225-240.

[17] S. R. Mohan, E.K. Park, Yijie Han, "Association Rule Based Data Mining Agents for Personalized Web Caching", Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC’05),Vol. 1, 2005, 37 – 38.

[18] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, "Privacy preserving mining of association rules", Information Systems, Vol. 29, Issues 4, 2004

[19] Ming Wu; Moon Jung Chung; H. D. K. Moonesinghe, "Parallel implementation of WAP-tree mining algorithm", Proceedings of the Tenth International Conference on Parallel and Distributed Systems(ICPADS 2004), 2004, 134 – 141

[20] J. Pei et.al., "Mining sequential patterns by patterngrowth: the PrefixSpan Approach", IEEE transactions on Knowledge and Data Engineering, Vol.16, Issue 11, 2004, 1414 – 1429

[21] R. Cooley, “The use of web structure and content to identify subjectively interesting web usage patterns”, ACM Transactions on Internet Technology (TOIT), Vol. 3, No. 2, 2003, 93–116.

[22] A. Nanopoulos, M. Zakrzewicz, T. Morzy, Y. Manolopoulos, “Indexing web access-logs for pattern queries”, 4th ACM CIKM International Workshop on Web Information and Data Management (WIDM’02), 2002.

[23] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, “The impact of site structure and user environment on session reconstruction in web usage analysis”, Proceedings of the 4th WebKDD 2002 Workshop Conference on Knowledge Discovery in Databases (KDD2002), 2002.

[24] G. Stumme, A. Hotho, B. Berendt, “Usage mining for and on the semantic web”, National Science Foundation Workshop on Next Generation Data Mining, 2002.

[25] E. Menasalvas, S. Millan, J. Pena, M. Hadjimichael, O. Marban, “Subsessions: a granular approach to click path analysis”, in: Proceedings of FUZZ-IEEE Fuzzy Sets and Systems Conference, at the World Congress on Computational Intelligence, 2002.

[26] "Platform for Privacy Preferences (P3P) Project", https://www.w3.org/TR/P3P/, 2002.

[27] Lindell, Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, Vol. 15, Issue 3, 2002, 177-206.

[28] Y. Saygin, V. S. Verykios , A. K. Elmagarmid, "Privacy preserving association rule mining", Proceedings of Twelfth International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems (RIDE-2EC 2002), 2002, 150-158.

[29] S.J. Rizvi, J.R. Haritsa, “Maintaining data privacy in association rule mining”, Proceedings of the 28th international conference on Very Large Data Bases, 2002, 682-693.

[30] X. Huang, N. Cercone, A. An, "Comparison of interestingness functions for learning web usage patterns", Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM Press, 2002, 617–620.

[31] B. Mobasher, H. Dai, M. Tao, "Discovery and evaluation of aggregate usage profiles for web personalization", Data Mining and Knowledge Discovery, Vol. 6, 2002, 61–82.

[32] S. Ansari, R. Kohavi, L. Mason, Z. Zheng, “Integrating e-commerce and data mining: Architecture and challenges”, in: N. Cercone, T.Y. Lin, X. Wu (Eds.), Proceedings of the 2001 IEEE International Conference on Data Mining(ICDM 2001), IEEE Computer Society, 2001.

[33] A. Banerjee, J. Ghosh, “Clickstream clustering using weighted longest common subsequences”, Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001.

[34] E.H. Chi, P. Pirolli, K. Chen, J.E. Pitkow, “Using information scent to model user information needs and actions and the web”, Proceedings of ACM CHI 2002 Conference on Human Factors in Computing Systems, ACM Press, 2001, 480–487.

[35] J. Z. Huang, et.al. , “A Cube Model and Cluster Analysis for Web Access Sessions”, WEBKDD 2001 — Mining Web Log Data across All Customers Touch Points, 2001, 48-67.

[36] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001

[37] S.S.C. Wong, S. Pal, "Mining fuzzy association rules for web access case adaptation”, Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case-Based Reasoning (ICCBR Ol), 2001.

[38] R. Kosala, H. Blockeel, “Web Mining Research: A Survey”, SIGKDD Explorations: Newsletter of Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, (2000), pp. 1-15.

[39] S. Ansari, R. Kohavi, L. Mason, Z. Zheng, “Integrating e-commerce and data mining: Architecture and challenges”, WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop, 2000.

[40] B. Mobahser , R. Cooley, J. Srivastava, “Automatic Personalization based on Web Usage Mining”, Communications of the ACM, Vol. 43, Issue 8, 2000.

[41] J. Srivastava, R. Cooley, M. Deshpande, P.N. Tan, “Web usage mining: discovery and applications of usage pattern from web data”, SIGKDD Explorations, Vol. 1, No. 2, 2000, 12-23.

[42] J. Andersen, A. Giversen, A.H. Jensen, R.S. Larsen, T.B. Pedersen, J. Skyt, “Analyzing clickstreams using subsessions”, International Workshop on Data Warehousing and OLAP (DOLAP 2000), 2000.

[43] J. Pei, J. Han, B. Mortazavi-asl, H. Zhu, “Mining access patterns efficiently from web logs”, Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000, pp. 386–397.

[44] R. Agrawal, R. Srikant, "Privacy-preserving data mining", ACM Sigmod Record, Vol. 29, Issue 2, 2000.

[45] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M. Hsu, "FreeSpan: frequent pattern-projected sequential pattern mining", Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000), 2000.

[46] P.-N. Tan, V. Kumar, “Modeling of web robot navigational patterns”, WEBKDD 2000—Web Mining for ECommerce—Challenges and Opportunities, Second International Workshop, 2000

[47] R. Cooley, B. Mobasher, J. Srivastava, “Data preparation for mining world wide web browsing patterns”, Knowledge and Information Systems, Vol. 1, Issue 1, 1999, 5–32.

[48] O. Etzioni, “The World Wide Web: Quagmire or gold mine?”, Communications of the ACM, (1996), pp. 65-68

[49] Margaret H. Dunham, “Data Mining Introductory and Advanced Topics”, Pearson Education

[50] Configuration file of W3C http, http://www.w3.org/Daemon/User/Config/ (1995).

[51] W3C Extended Log File Format, http://www.w3.org/TR/WD-logfile.html (1996).