International Journal of advanced studies in Computer Science and Engineering
IJASCSE Volume 4, Issue 11, 2015

Analyzing the User Behaviours by Mining Web Access Log Files

Dr.R.Chinnaiyan
Professor, Department of Computer Applications. New Horizon College of Engineering, Bangalore

Dr.V.Ilango
Professor, Department of Computer Applications, New Horizon College of Engineering, Bangalore

Abstract: Analyzing the usage of web sites and the web traffic by the users is vital in the growing world of internet. The growth of the World Wide Web (WWW) has showed the way of diverse client and server side tools development that mines the information resources to extract the intelligent knowledge. Analyzing this data will helps the enterprises for realizing the value of their customers and provides them with a more sophisticated structure of web sites and services. Web Mining is a process of determining the knowledge and vital information from the World Wide Web. Web Mining consists of three different categories, namely Web Content Mining, Web Structure Mining, and Web Usage Mining. The main aim of Web Usage Mining is to analyze the users' navigation patterns and their use of web resources. The primary focus of this paper is to analyze the user behaviours by mining the web access log files.

Keywords
WWW, Web Mining, Content, Structure, Usage.

1. INTRODUCTION

Web servers store information of each page requested by web visitors called the web access log. Web Usage Mining addresses the problem of extracting behavioral patterns from one or more web access logs. Web Usage Mining (also called as click-stream analysis) Edelstein [5] is the process of applying data mining techniques to the discovery of usage patterns from Web data, and is targeted towards applications Srivastava, et al. [8]. It tries to make sense of the data generated by the Web surfer's sessions or behaviors. While the web content and structure mining use the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users during Web sessions. Web usage data includes the data from web server access logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, mouse clicks, and any other data as the result of interaction with the Web. Figure 1.depicts the Web Usage Mining process.

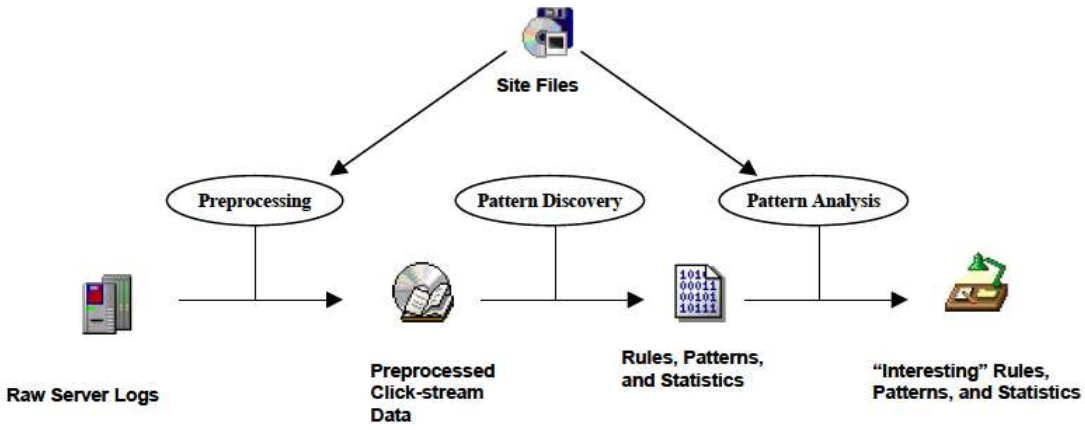

Figure 1. Process of Web Usage Mining

1.1 Web Access Logs And Web Usage Mining

In order to manage a web server effectively, it is necessary to get feedback about the activity and performance of the server as well as any problems that may be occurring. Web server creates and maintains log files for this purpose . A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site.

2. Motivation

In the recent times more and more users are using the internet services for their necessity. This helps to do research in extracting useful information and user interest from the web using mining techniques. The web logs are one of the most utilized features to extract the user's interest measure. The web log mining is used more frequently in order to identify the user behavior based on the extent to which a user is visiting a particular web site. Since web logs are updated each time the user visits a particular web site, so it is considered as moving data. Thus algorithms which are capable of processing moving data are to be considered for the mining of the web logs.

2.1 Related Work

Web content mining typically involves searching through web based information repositories. Broadly speaking, there have been two approaches to searching content. The data base approach is an extension of traditional data base querying, though adapted for the nature of the Web (OKS Group, 2003; Biggs, 2003). The agent based approach is inspired by autonomous agent research and the creation of web-bots for adaptively searching or spidering through web repositories (Menczer, 2003). The driving force behind Web usage mining was business to consumer (B2C) e-commerce. By searching for usage patterns from within weblogs, one could adapt a web site to maximise the opportunity for increased sales (Doherty,2000; Iyer et al, 2002; Manchester, 2001; Mulvenna et al, 2000; Thelwall, 2001). Both web content and web usage mining can take a retrieval approach of known or anticipated content or a discovery of new or unexpected content. There is some debate over whether content retrieval is true web mining (Hearst, 1999), however on balance, the term has been typically used to cover both aspects.

The Special Edition on Web Retrieval and Mining for Decision Support Systems (Chen, 2003) identifies a strong future for WM, with the prospect of applications supporting a multi-lingual web, a multi-media web and a wireless accessible web. The following sections will review the literature in terms of technology trends, application trends and product trends. Paola Britos et al., [18] described the capacity of use of Self Organized Maps, kind of artificial neural network, in the process of Web Usage Mining to detect user's patterns. The process detail the transformations necessaries to modify the data storage in the Web Servers Log files to an input of Self Organized Maps. Mehrdad Jalali et al., [13] presented an approach which is based on the graph partitioning for modeling user navigation patterns.

In order to mining user navigation patterns, they establish an undirected graph based on connectivity between each pair of the web pages and also proposed novel formula for assigning weights to edges of the graph. Kobra Etminani et al., [9] applied ant-based clustering algorithm to pre-processed logs to extract frequent patterns for pattern discovery and then it is displayed in an interpretable format. N. Sujatha et al., [16] have proposed a new framework to improve the web sessions' cluster quality from k-means clustering using Genetic Algorithm (GA). Mahdi Khosravi et al., [11] proposed a novel approach for dynamic mining of users' interest navigation patterns, using naïve Bayesian method

V.V.R. Maheswara Rao and Dr. V. ValliKumari [20] have proposed an extensive learning algorithm to get the desired information. In their paper they introduce an extensive research frame work capable of pre processing web log data completely and efficiently. The learning algorithm of proposed research framework can separates human user and search engine accesses intelligently, with less time.

Xiaohua Hu et.al [21] have developed two approaches, exact match and relatedness-match, the main aim is to extract the related information from Wikipedia. These two techniques are utilized to map text documents to Wikipedia concepts, and further to Wikipedia categories. Then the text documents are clustered based on a similarity metric which combines document content information, concept information as well as category information.

Yanjun Li, Soon M chung and John D Holt (22) have devised two algorithms on the basis of frequent word meaning. The clustering techniques proposed are namely clustering based on frequent word sequence and clustering based on frequent word meaning sequence. They have tried to cluster

documents finely on the basis of the words which are making the sentence .This proposed technique is very different from the conventional techniques which used vector space models and treats whole documents not words.

Mehrdad Mahdaviand and Hassan Abolhassani [10] have proposed a Fast and high quality document clustering for crucial task in organizing information, search engine results, enhancing web crawling, and information retrieval or filtering. They have used the most commonly used partitionbased clustering algorithm, the *K*-means algorithm, which is more suitable for large datasets. However, the *K*-means algorithm can generate a local optimal solution. But they propose a novel Harmony *K*means Algorithm (HKA) that deals with document clustering based on Harmony Search (HS) optimization method.. Dimitrios Pierrakosand and Georgios Paliouras [3] have proposed a knowledge discovery framework for building Web directories according to the preferences of user communities.

3. Pattern Analysis

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition . They are similar to those developed for non-Web domains such as statistical analysis, clustering, and classification, but those methods must take into consideration the different kinds of data abstractions and prior knowledge available for Web Mining. In Web Usage Mining, a server session is an ordered sequence of pages requested by a user. Pattern analysis is to filter out the uninteresting rules or patterns from the dataset found in the pattern discovery phase. The exact methodology used for analysis is usually governed by the application for which Web mining is to be done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can highlight patterns. The content and structure information can be used to filter out patterns which contain pages of a certain use type or content, or pages that match a certain hyperlink structure.

3.1 Pattern Extraction

It deals with extracting interesting patterns from the pre processed web logs. This is the key component of web usage mining. Table 1 depicts the general activities statistics of a website

Type	Description
General Data	1) Total Number of Hits 2) Total Number of Visitor Hits 3) Total Number of Spider Hits 4) Average Number of Hits per day 5) Average Number of Visitor Hits per day 6) Total Number of Successful Requests 7) Total Number of Failed Requests 8) Total Number of Incomplete Requests 9) Total Number of Error Reports
Activity Data	1) Daily Activity 2) Activity by Hour of Day 3) Activity by Day of Week 4) Activity by Week 5) Activity by Month
Access Data	1) Page Views 2) File Views 3) Image Access 4) Directory Access

Table 1. General Activities Data of a Website

4 .NUMERICAL RESULTS

This following statistics is prepared with Web Log Expert application. It is a pace and commanding web access log analyzer tool used for predictive analytics. It also produces information about the web site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. It also helps for producing easy to read reports including text information and charts. The Table 1 shows the general activity statistics of the website. Results of general statistics shows that there are 30474 hits, 29191 visitors, 4435 page views,2932 IPs.

Description	Count
Hits	
Total Hits	30474
Visitor Hits	29191
Spider Hits	1283
Average Hits per Day	4353
Average Hits per Visitor	8.18
Cached Requests	3979
Failed Requests	233
Page Views	
Total Page Views	4435
Average Page Views per Day	633
Average Page Views per Visitor	1.24
Visitors	
Total Visitors	3570
Average Visitors per Day	510
Total Unique IPs	2932

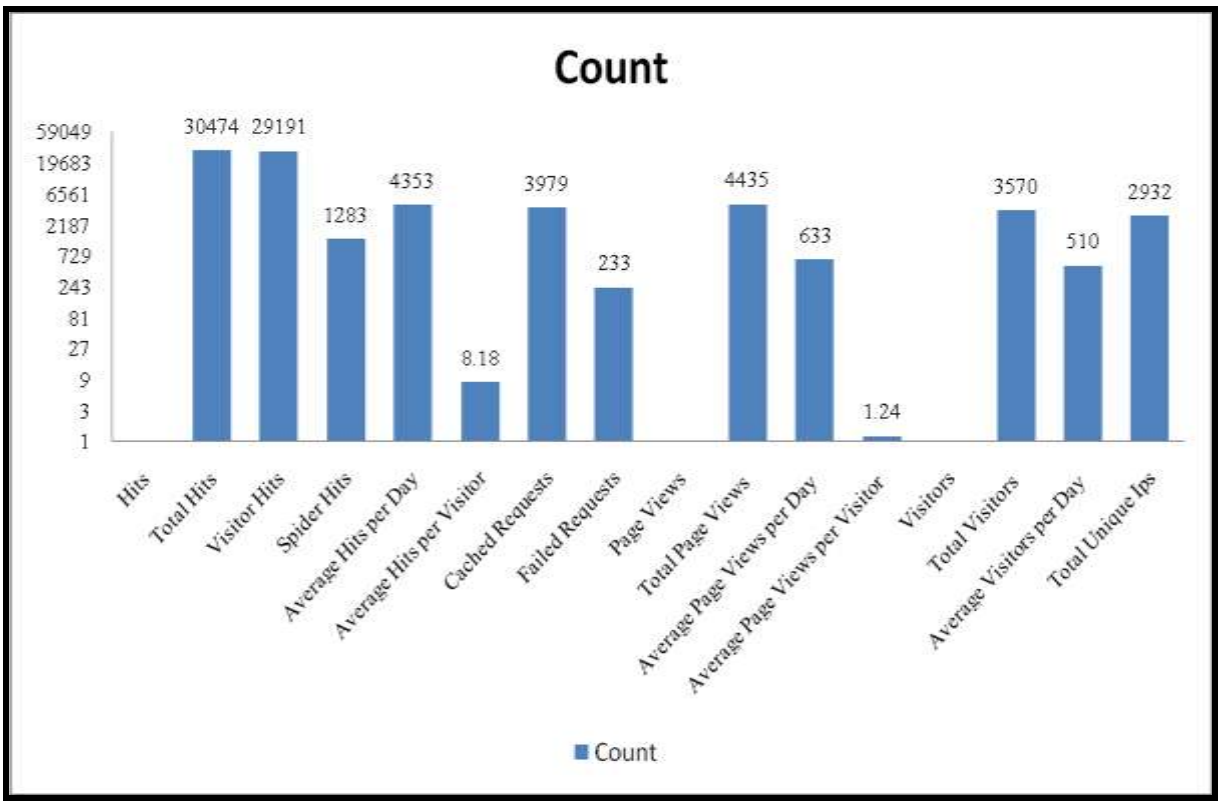

Figure 2. General Activity Statistics of a Web Site

The Table 2 shows the most requested file types in the web site. Results of most requested file types shows that there are 18403 requests for gif file. 4148 requests for html files. 2186 requests for jpg files. 1746 requests for ico files and 1032 requests for css files etc.,

Table 2. Most Requested File Types

S.No	File Type	Hits	Incomplete Requests
1	gif	18403	5
2	html	4148	0
3	jpg	2186	2
4	ico	1746	0
5	css	1032	4
6	exe	890	539
7	php	224	0
8	xml	152	0
9	pdf	124	103
10	swf	94	0
11	txt	32	0
12	zip	12	0
13	asp	7	0
14	rtf	1	0
	Total	**29051**	**653**

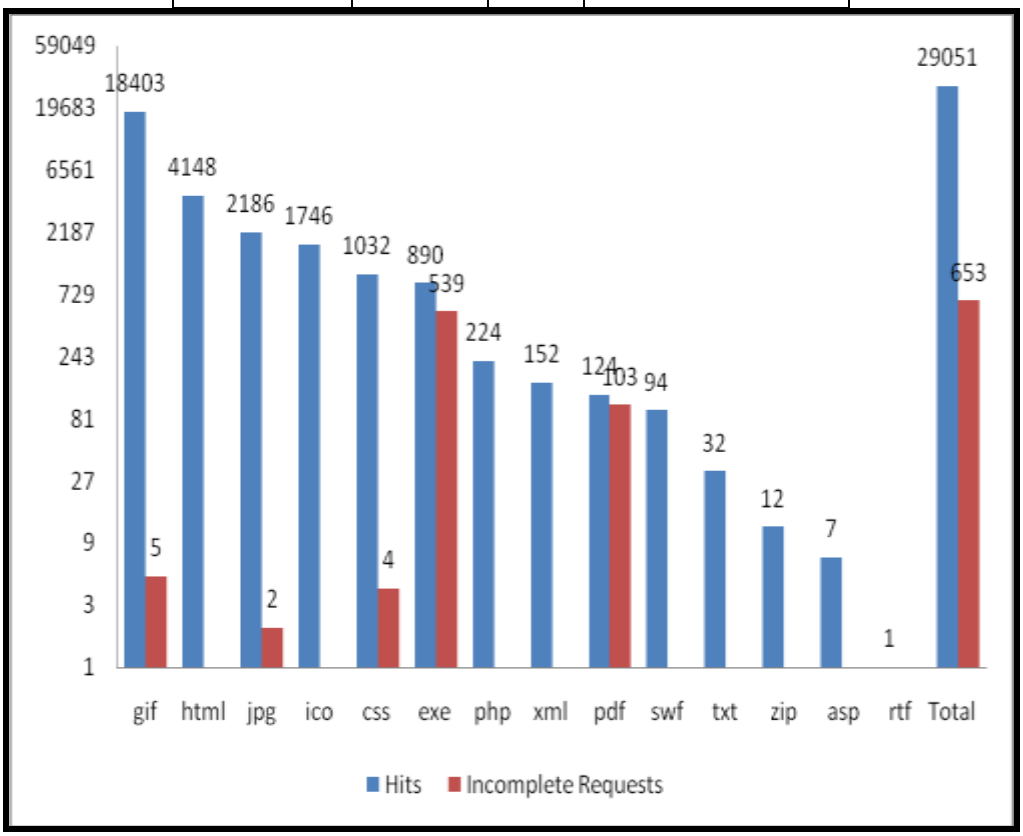

Figure 3. Most Requested File Types

The Table 3 shows the Most used Web Browsers. Results of Most used Web Browsers shows that the customers used Internet Explorer for 21302 times, Fire Fox for 6461 times, Opera 400 times , Other web browsers 148 etc.

Table 3. Most Used Browsers

S.No	Browsers	Hits	Visitors	% of Total Visitors
1	Internet Explorer	21302	2260	61.43
2	Firefox	6461	883	24.00
3	Others	275	153	4.16
4	Opera	400	100	2.72
5	Google Desktop	60	59	1.60
6	Mozilla/4.0 (compatible;)	148	38	1.03
7	ActiveRefresh	19	19	0.52
8	Gecko/20070308 Minefield/3.0a1	96	10	0.27
9	Mozilla	36	9	0.24
10	Avant Browser	8	7	0.19
11	Other Browsers	386	141	3.84
	Total	**29191**	**3679**	**100%**

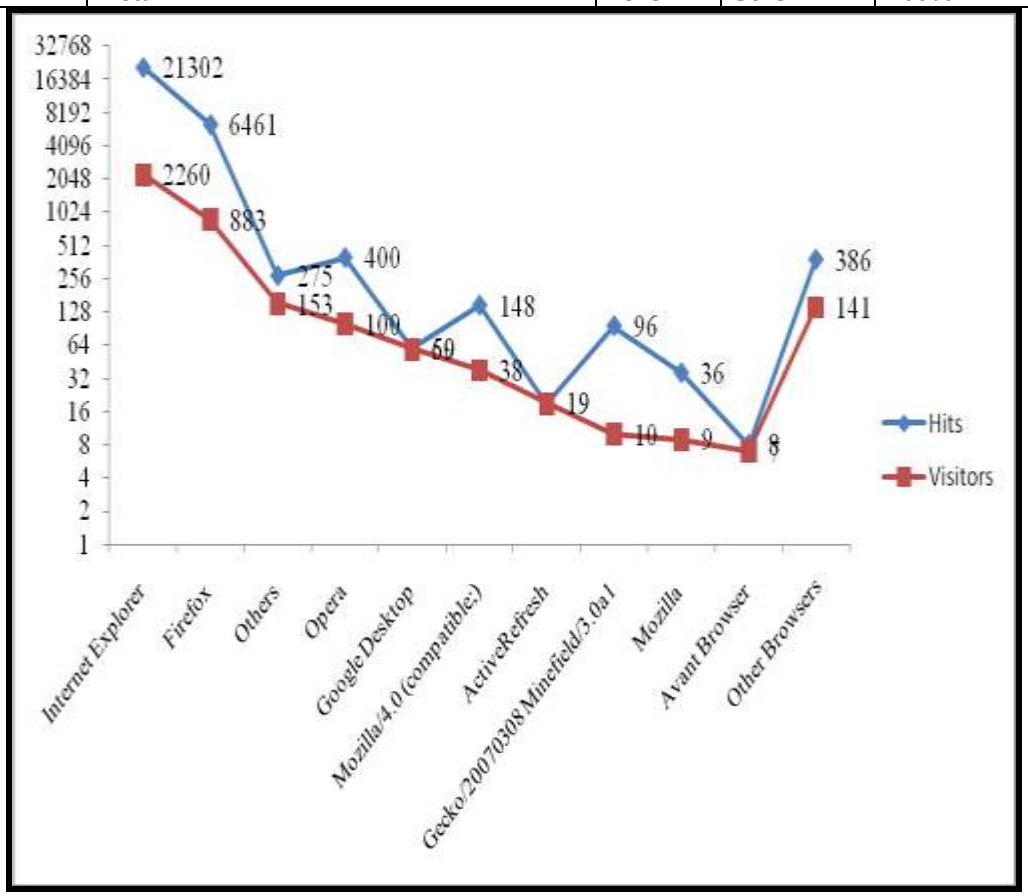

Figure 4. Most Used Browsers

The Table 4 shows the Most used Operating Systems by the Customer. Results show that the customer used 21449 times Windows XP, 3921 times Windows Vista, 1386 times Windows 2000, 890 times Windows Server 2003, 482 times Windows 98, 713 times other Operating Systems.

Table 4. Most Used Operating Systems

S.No	Operating Systems	Hits	Visitors	% of Total Visitors
1	Windows XP	21449	2383	66.75
2	Windows Vista	3921	440	12.32
3	Others	713	299	8.38
4	Windows 2000	1386	181	5.07
5	Windows Server 2003	890	94	2.63
6	Mac OS	107	54	1.51
7	Windows 98	482	50	1.40
8	Windows NT	55	28	0.78
9	Linux	174	28	0.78
10	Windows 95	5	5	0.14
11	Windows ME	4	4	0.11
12	Chrome OS	3	2	0.06
13	FreeBSD	1	1	0.03
14	Windows 3.x	1	1	0.03
	Total	**29191**	**3570**	**100**

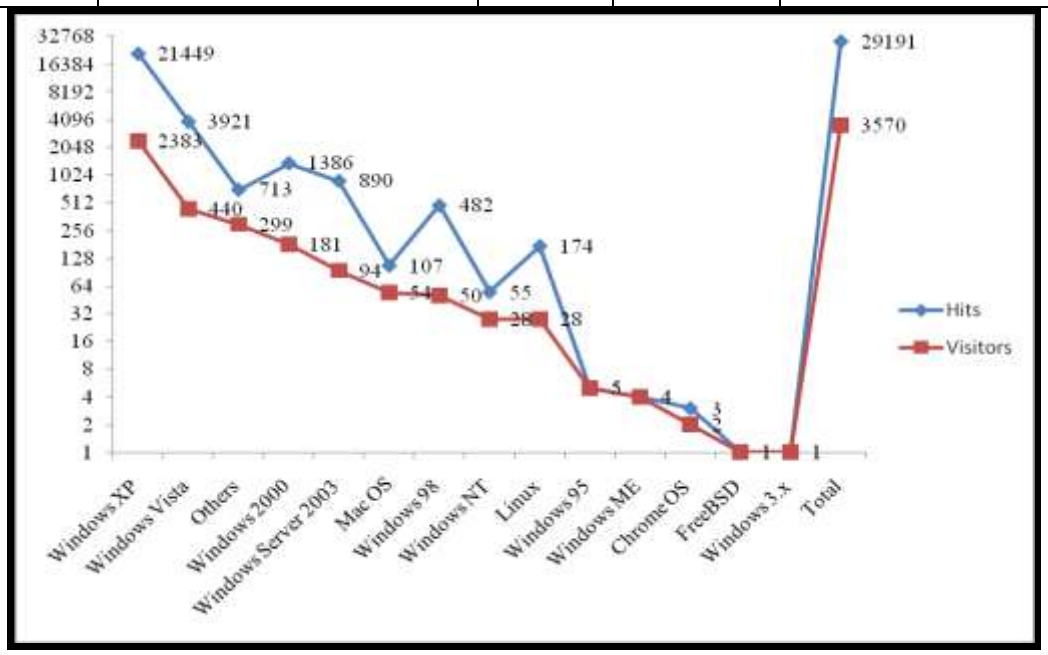

Figure 5. Most Used Operating Systems

5. CONCLUSION

As the web and its usage continue to grow, it is important to analyze web data and extract all manner of useful knowledge from it. Now a days web mining is a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper a survey is made in the areas of Web mining, focusing on the category of Web mining. Since this is a huge area, and there a lot of work to do, Thus this paper could be a useful starting pack for the academicians

and the researchers for identifying the various opportunities for further research

6. REFERENCES

1. Biggs, M. (2003). "Data Mining outside the firewall." InfoWorld Sept 5th 2003.
2. Chen, H. (2003). "Special Issue: "Web retrieval and mining "" Decision Support Systems 35 : pp.1-5.
3. Dimitrios Pierrakos and Georgios Paliouras. "Personalizing Web Directories with the Aid of Web Usage Data". IEEE Transactions on Knowledge and Data Engineering , vol 22(9):pp 1331-1344, 2010
4. Doherty, P. (2000). "Web Mining - the E-Tailer's Holy Grail . " DM Review.
5. Edelstein, H. A. (2001, March 12, 2001). Pan for Gold in the Clickstream. Information Week, 77 - 91.
6. Hearst, M.: Untangling Text Data Mining. In the Proceedings of the ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland (1999).
7. Iyer, G., A. Miyazak i, et al. (2002) . "Linking Web-based segmentation to pricing tactics . " Journal of Product & brand Management 11 (5): pp.288-302
8. J. Srivastava, R. Cooley, M. Deshpande, and P. -N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," SIGKDD Explorations , Vol. 1, No. 2, pp. 12-23,2000
9. Kobra Etminani, Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009
10. M. Mahdavi and H. Abolhassani, "Harmony k-means algorithm for document clustering," Data Mining and Knowledge Discovery, vol. 18, no. 3, pp. 370 – 391, 2009
11. Mahdi Khosravi, Mohammad J. Tarokh, "Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method", 978-1-4244- 8230-6/10/$26.00 ©2010 IEEE
12. Manchester, P. (2001). "Every detail of the way potential customers use a website is re cordable and this information can be used to turn visitors into buyers ." Financial Times
13. Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman, "WEB USER NAVIGATION PATTERN MINING APPROACH BASED ON GRAPH PARTITIONING ALGORITHM", Journal of Theoretical and Applied Information Technology
14. Menczer, F. (2003). Complementing search engines with online Web mining agents. Decision Support Systems, 35(2)
15. Mulvenna, M., Anand, S., & Bchner, A. (2000). Personalization on the net using Web mining. Communications of ACM, 43(8)
16. N. Sujatha, K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476
17. OKS Group (2003), "Web Mining", www.oksgroup.com
18. Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García-Martínez, "Web Usage Mining Using Self Organized Maps", International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007
19. Thelwall, M. (2001). "Web log file analysis : backlink sandqueries . " Aslib Proceedings 53 (6)
20. V.V.R. Maheswara Rao and Dr. V. Valli Kumari, An Enhanced Pre- Processing Research Framework For Web Log Data Using A Learning Algorithm, netcom 2010,CSCP 01, pp. 01-15, 2011
21. Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. 2009. Exploiting wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining , pages 389–396. ACM
22. Yanjun Li, Soon M. Chung, John D. Holt. 2008. "Text document clustering based on frequent word meaning sequences". Data & Knowledge Engineering. 64, 1 (Jan.2008). Elsevier Science Publishers B. V. Amsterdam,The Netherlands. 381 -404.