AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Advancement in Web Usage Mining by Analyzing Web Log Files Using Clustering

[1]Senthil Pandian, P. and [2]Dr.Srinivasan, S.

[1]*Research Scholar,Anna University, Regional Office, Madurai*
[2]*Professor & Head, Dept. of CSE, Anna University, Regional Office, Madurai*

ARTICLE INFO

Article history:
Received 8 August 2014
Received in revised form
12 September 2014
Accepted 25 September 2014
Available online 2 November 2014

Keywords:
Web log file, Data Preprocessing, Clustering

ABSTRACT

An absolute framework and findings are presented in mining web usage patterns from web log files of a real web site that has all the challenging aspects of real-life web usage mining, including evolving user profiles and external data describing ontology of the web content. The behavior of a web site's users may change so quickly that attempting to make predictions, according to the frequent patterns coming from the analysis of an access log file, becomes challenging. In order for the obsolescence of the behavioral patterns to become as possible, the ideal method would provide frequent patterns in real time, allowing the result to be available immediately. Even though the web site under study is part of a nonprofit organization that does not "sell" any products, it was crucial to understand "who" the users Ire, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, this is an approach for discovering and tracking evolving user profiles here also describes how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from web log data. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior. In this a method allowing finding frequent behavioral patterns in real time, whatever the number of connected users has been measured.

To Cite This Article: Senthil Pandian, P. and Dr. Srinivasan, S., Advancement in Web Usage Mining by Analyzing Web Log Files Using Clustering. ***Aust. J. Basic & Appl. Sci.,*** *8(16): 125-131, 2014*

INTRODUCTION

Customer Relationship Management (CRM) can use data from within and outside an organization to allow an understanding of its customers on an individual basis or on a group basis such as by forming customer profiles. The short-term model is learned from the most recent observations only, whereas the long-term (default) model represents the user's general preferences. Specifically, there are a number of issues in pre-processing data for mining that must be addressed before the mining algorithms can be run. Analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the web site. Because of many unique characteristics of the client-server model in the World Wide Web, including differences between the physical topology of web repositories and user access paths, and the difficulty in identification of unique users as well as user sessions or transactions, it is necessary to develop a new framework to enable the web usage mining process.

1.1 The Web usage mining process:

Web usage Mining contains of three basic steps. These steps are preprocessing, pattern discovery and pattern analysis. To successfully complete an analysis of a web site, it is must to obtain data suitable for data mining at the beginning of a process. Mostly data preprocessing step is the most time-consuming step in web usage analysis.

The task of preprocessing is to prepare the data for the application of some data mining algorithm. After data has been preprocessed, it is ready for the application of knowledge extraction algorithms. The process is to analyze the obtained results in order to distinguish trivial, useless knowledge from knowledge that could be used for web site modifications, system improvement and web personalization. After preprocessing the data is ready for knowledge extraction and it is extracted by using some pattern analysis method or techniques. The trivial result makes a clear distinction between trivial useless knowledge from knowledge which is useful for improvement of systems. The pattern analysis reveals the frequency of visits per document, most recent visit per

Corresponding Author: Senthil Pandian, P.,Research Scholar,Anna University, Regional Office, Madurai

document, who is visiting which documents, frequency of use of each hyperlink, and most recent use of each hyperlink. Patterns discovery implements techniques such as data mining, psychology, and information theory association, path analysis, cookies, sequential patterns, and so on.

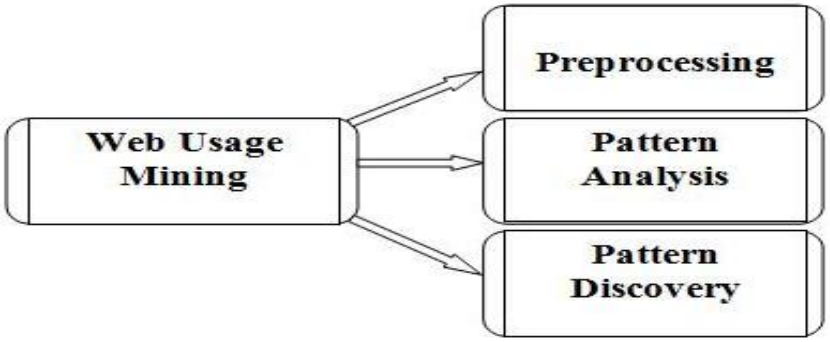


Fig. 1.1: Web Usage Mining Process.

2. Contemporary system:

In earlier research efforts in web usage mining have worked with the assumption that the web usage data is static. However, the dynamic aspects of web usage have recently become important. This is because web access patterns on a web site are dynamic due not only to the dynamics of web site content and structure but also to changes in the user's interests and, thus, their navigation patterns. Thus, it is desirable to study and discover web usage patterns at a higher level, where such dynamic tendencies and temporal events can be distinguished.

Mining evolving click streams is the subject of only a few recent research efforts. Raiyani and Jain(2012) stated that a user profiling system was developed based on monitoring the user's web browsing and e-mail habits. This system used a clustering algorithm to group user interests into several interest themes, and the user profiles had to adapt to changing interests of the users over time. The adoption approach based on periodical batch mining that has the advantage of being easy to adapt to use any other unsupervised learning tool that automatically discovers clusters in static or dynamic data. In this work, it is good to use the full memory (periodical or window based), in part, because our goal was to describe the user profiles in certain periodical increments (about two weeks each). Hence, it was essential to fully mine the web logs from each period and then compare the subsequent results.

In web usage logs data's were enriched with semantics derived from the content of the web site's pages. Content keywords were first mapped to the categories of a manually constructed domain-specific taxonomy through the use of a thesaurus, and then the web documents were clustered based on the taxonomy categories. The enhanced web logs, called C-Logs, were then used as input to web usage mining. Relying only on web usage data for user modeling or for personalization can be inefficient, either when there is insufficient usage data for the purpose of mining certain patterns or when new pages are added and thus do not accumulate sufficient usage data at first. Nasraoui, Krishnapuram, and Joshi (2001) mentioned the lack of usage data in these cases can be compensated by adding other information such as the content of web pages or the structure of a web site. In keywords that appear in web pages are used to generate document vectors, which are later clustered in the document space to further augment user profiles. Borges and Levene (2009) the web site's own hierarchical structure is treated like an implicit taxonomy or concept hierarchy that is exploited in computing the similarity between any two web pages on the web site. Most of the efforts cited on the previous page rely on an explicit taxonomy that needs to be handcrafted by an expert before the analysis. On the other hand, the implicit taxonomy, as used in inferred automatically and quickly from the web site directory structure via URL tokenization. Furthermore, this implicit taxonomy does not require any modification to the underlying data mining algorithm, since it is only incorporated within the similarity measure used to cluster the user sessions.

3. Description of the proposed system:

The process of web usage mining and a road map to the rest is summarized in Fig. 1.1, which starts with the integration and preprocessing of web server logs and server content databases, pattern analysis and pattern discovery. This is followed by a post processing of the clustering results to obtain web user profiles and finally ends with tracking profile evolution. The automatic identification of user profiles is a knowledge discovery task consisting of periodically mining new contents of the user access log files and is summarized in the following steps:

1. Preprocess web log file to extract user sessions.
2. Cluster the user sessions by using Hierarchical Unsupervised Niche Clustering (H-UNC).

3. Summarize session clusters/categories into user profiles.
4. Enrich the user profiles with additional facets by using additional web log data and external domain knowledge.
5. Track current profiles against existing profiles.

3.1 Preprocessing the web log file to citation user sessions:

Each URL in the site is assigned a unique number where NU is the total number of valid URLs. The h user session is then encoded as an NU-dimensional binary attribute vector with the following property: In addition to URLs, encode the search query terms from the initial request's REFERRER field and take advantage of the

Power Law properties of session lengths (with the majority tending to be short) to implement sessions as lists instead of vectors, thus saving on memory and computational costs. The access log of a web server is a record of all files (URLs) accessed by users on a web site. Each log entry consists of the access time, IP address, URL viewed, REFERRER (the web page visited just prior to the current one), etc. The first step in preprocessing is explained in Nasraoui, Cardona, Rojas, and Gonzalez, (2011), consists of mapping the URLs on a web site to distinct indices. A user session consists of requests from the same IP address within a predefined time period.

3.2 Pattern Analysis:

In analysis first of all cookies can be deleted by the user, cache busting defeats the speed advantage that caching was created to provide and can be disabled, and user registration is voluntary and users often provide false information. A related but much harder problem is determining if there are important accesses that are not recorded in the access log. Mechanisms such as local caches and proxy servers can severely distort the overall picture of user traversals through web site Current methods to try to overcome this problem include the use of cookies, cache busting, and explicit user registration. As detailed in, none of these methods are without serious drawbacks. Methods for dealing with the caching problem include using site topology or referrer logs, along with temporal information to infer missing references.

Another problem associated with proxy servers is that of user identification. Use of a machine name to uniquely identify users can result in several users being erroneously grouped together as one user. An algorithm presented in checks to see if each incoming request is reachable from the pages already visited. If a page is requested that is not directly linked to the previous pages, multiple users are assumed to exist on the same machine. In, user session lengths determined automatically based on navigation patterns are used to identify users. Other heuristics involve using a combination of IP address, machine name, browser agent, and temporal information to identify users. Techniques to clean a server log to eliminate irrelevant items are of importance for any type of web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses of the web site. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, JPG, and map can be removed.

3.3 Pattern Discovery:

Conversion starts from file name to page titles after the completion of defining the titles. Page titles are easier than

URL's so page titles are visualized on reports in addition to URLs. The first type is *navigation-content*, where each transaction consists of a single content reference and all of the navigation references in the traversal path leading to the content reference. These transactions can be used to mine for path traversal patterns. The second type of transaction is *content-only*, which consists of all of the content references for a given user session. These transactions can be used todiscover associations between the content pages of a site. A given page reference is classified as either navigational or content, based on the time spent on the page. This kind of "page typing" is further delineated in, where various page types such as index pages, personal home pages, etc. are used in the discovery of user patterns. Each transaction is defined to be the set of pages in the path from the first page in the log for a user up the page before a backward reference is made. A new transaction is started when the next forward reference is made. A forward reference is defined to be a page not already in the set of pages for the current transaction. Similarly, a backward reference is defined to be a page that is already contained in the set of pages for the current transaction. For example, an access sequence of A B C D C B E F E G would be broken into three transactions, i.e. A B C D, A B E F, and A B E G. The transactions created with this algorithm are similar to the *navigation-content* transactions of and can be used to mine for path traversal patterns. Before any mining is done on web usage data, sequences of page references must be grouped into logical units representing web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, as discussed above. A transaction differs from a user session in that the size of a transaction can range from a single page reference to the entire page references in a user session, depending on the criteria used to

identify transactions. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method of clustering page references into transactions smaller than an entire user session.

3.4 Identifying Sets by Clustering Sessions:

To cluster user sessions, it is must to use H-UNC from Mishra and Choubey, (2012) a divisive hierarchical version of a robust clustering approach (Un-supervised Niche Clustering (UNC)) that uses a Genetic Algorithm (GA)to evolve a population of candidate solutions through generations of competition and reproduction. The main outline of the H-UNC algorithm is sketched in the following. The reason that use H-UNC instead of other clustering algorithms is that unlike most other algorithms, H-UNC can handle noise in the data and automatically determines the number of clusters. In addition, evolutionary optimization allows the use of any domain-specific optimization criterion and any similarity measure, in particular a subjective measure that exploits domain knowledge and ontologies. However, unlike purely evolutionary search-based algorithms, H-UNC combines evolution with local Piccard updates to estimate the scale of each profile, thus converging fast (about 20 generations). H-UNC is outlined. This web similarity takes into account not only the hierarchical structure of the web site content as inferred from the URL address itself (for example, URLs a/b/c and a/b/d are related from the hierarchical structure aspect) but also how different content items on the web site relate to each other according to an externally defined web site ontology are semantically related from an external ontology aspect if these two URLs can be mapped to A/B and A/C, that is, if they refer to content areas B and C that share the same parent A).

Table-3.4.1:

Descp. of Activity	Date of completion	DURATION	First Month			Second Month			Third Month				Fourth Month				Fifth Month		
			2nd WEEK	3rd WEEK	4th WEEK	1st WEEK	2nd WEEK	3rd WEEK	4th WEEK	1st WEEK	2nd WEEK	3rd WEEK	4th WEEK	1st WEEK	2nd WEEK	3rd WEEK	4th WEEK	1st WEEK	2nd WEEK	3rd WEEK
Project Approval	10-12-08	-																		
Abstract	09-01-09	29 days																		
Req. Specific. & Design	23-01-09	14 days																		
Coding and testing	21-02-09	11 days																		
Demo and draft project Report	13-03-09	23 days																		
Final project subm.	01-04-09	20 days																		

3.5 Clustering Used by Comparison Measure:

This approach only implicitly incorporates information about the web pages' content. This is different from methods based on the explicit content of the web pages, as infer this information from the hierarchy knowledge that is external to the web logs. A good survey with extensions in a fuzzy-set theoretic framework can also be found in. Thus, normalization is done by dividing by the average of the path lengths, whereas our similarity divides by the maximal length. Hence, our similarity is more restrictive and penalizes more for widely differing path lengths that correspond to concepts at widely different levels of specificity and generality. In addition, IC requires a rigid taxonomy and reliable corpus to be able to accurately estimate the probabilities. Thus, the combination of hierarchical site structure and external ontology occurs naturally in two stages: First, each URL is parsed to extract the structure, and then, each remaining dynamic URL (that is, taxonomy) between individual

dynamic URLs and higher level categories encoded in web site ontology. This similarity is used in our clustering algorithm (H-UNC) to group similar user sessions into clusters or profiles.

3.6 User Profiles by Post Processing and Upgrading of Session Clusters:

After automatically grouping sessions into different clusters, I summarize the session categories in terms of user profile vectors. The k^{th} component/ weight of this vector (pik) captures the relevance of URL k in the I^{th} profile, as estimated by the conditional probability that URL k is accessed in a session belonging to the I^{th} cluster (this is the frequency with which URL k was accessed in the sessions belonging to the i^{th} cluster). The profiles are then converted to binary vectors (sets) so that only URLs with weights> 0:15remain. The model is further extended to a robust profile based on robust weights In addition to the cluster-induced user profiles above here in several descriptors of the users in each cluster.

4. Methodologies:

4.1 Relevant Query Terms and Probing Related Information for User Profiles:

In addition to the relevant URLs that are extracted from the sessions that can extract the explicit information need of the users in each profile from the queries that they could have typed prior to visiting the web site when this information is available from the readily available REFERRER field in the web log files. Hence, for each profile, it accumulates all the search phrases extracted from the REFERRER fields of the assigned user sessions. This allows us to describe each profile in terms of either a set of significant URLs or a set of explicit search query phrases and terms. In addition to the relevant URLs that are extracted from the sessions assigned to each profile, this can extract information about which companies or organizations tend to visit the web site and fall in this profile to extract this information from two complementary sources:1) by getting the company information that corresponds to an ID in the server content Database, where the ID is extracted from the web log if the visitors did not sign in through the registration page, then an attempt is made to obtain the company affiliation from a specialized web service (www.whois.com). The web site under study provides a virtual meeting point between different companies providing various services that are related to the portal's subject. Hence, it was important to know not only which companies take part in each cluster of activities but also what company information seemed to be relevant to users in each cluster.

4.2 Polygonal User Profiles from Web Usage Facts Experimental Results of Mining:

The discovered profiles in frequent patterns will provide one way of forming a summary of the input data. As a summary, profiles represent a reduced form of the data that is, at the same time, as close as possible to the original input data. This description is reminiscent of an information retrieval possible to the original session data. Closeness should take both of the following into account:

1. Precision. A summary profile's items are all correct or included in the original input data; that is, they include only the true data items.
2. Coverage /Recall. A summary profile's items are complete compared to the data that is summarized; that is, they include all the data items.

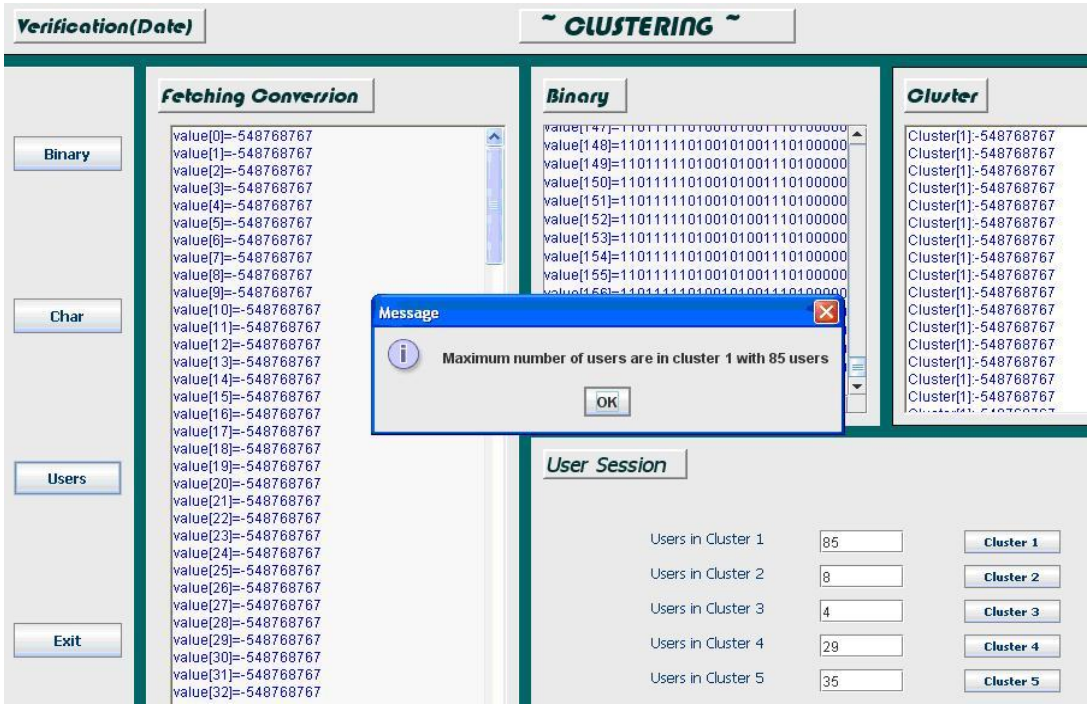
These criteria are clearly contradictory, since precision will favor only the smallest profiles, eventually with a single URL, whereas coverage will favor the largest possible profiles. Ideally, each data query should be answered by a profile that is identical to this query. H-UNC was applied on a set of web sessions preprocessed from web log data for several months in 2008 and 2009. After filtering out irrelevant entries, the data was segmented into sessions based on the client IP address and a time-out threshold between two consecutive accesses in the same session of 45 minutes. After filtering irrelevant URLs (for example, graphics) and requests from web crawlers, should get a number of unique sessions varying from 800-1,500 per week, accessing between 3,000-6,000 URLsin obtaining (not countingthe pt. graphics). Thus for the studied periods, H-UNC was applied to the web sessions by using a maximal number of levels L¼10 in the hierarchy and the following parameters that control the final resolution on Zaiane, Xin and Han (2008): N split and split H-UNC partitioned the web user sessions of each period into several clusters (that ranged from 20 to 35 clusters, depending on the period), and each cluster was characterized by one of the user profile.

Conclusion:

The tracking and validating evolving multifaceted user profiles on web sites that have all the challenging aspects of real-life web usage mining, including evolving user profiles and access patterns, dynamic web pages, and external data describing ontology of the web content.

A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies. The choice of the period length for analysis depends on the application or can be set, depending on the cross-period validation results. Even though it did not focus on scalability, the latter can be addressed by following an approach similar to where web click streams are considered as an evolving data stream, or by mapping some new sessions to persistent profiles and

updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.



REFERENCES

Bianco, A., G. Mardente, M. Mellia, M. Munaf and L. Muscariello, 2012. "Web User Session Characterization via Clustering Techniques", Computer Networks, Special Issue on Long-Range Dependent Traffic, 40(3): 319-337.

Borges, J. and M. Levene, 2009. "Data Mining of User Navigation Patterns", Web Usage Analysis and User Profiling, pp: 92-111.

Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa, 2010. "Web Data Mining: Effective personalization based on association rule discovery from web usage data.

Cooley, A., Lu. Mobasher and K Srivastava, 1999. "Web Mining: Information and Pattern Discovery on the World Wide web, in the Proceedings of the 1999 Ninth IEEE International Conference, pp: 558-567.

Desikan, P. and J. Srivastava, 2010. "Mining Temporally Evolving Graphs,"Proc. Workshop In Mining and Web Usage Analysis (Web KDD' 2010).

Gupta, M.R. and P. Gupta, 2011. "Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree", in the International Journal of Computer Science & CommunicationNetworks, vol. 1.

Mishra, M.R. and M.A. Choubey, 2012. "Discovery of frequent patterns from web log data by using FP-growth algorithm for web usage mining," in the International Journal of Advanced Research in Computer Science and SoftwareEngineering vol. 2.

Mobasher, B., Honghua Dai, Tao Luo, Nakagawa, 2000. "Using sequential and non-sequential patterns in predictive web usage mining tasks",EPA 105-R-95-2000.

Nasraoui, O., A. Krishnapuram and T. Joshi, 2001. "Mining web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimatorin the Proceedings of the 2001 Eighth Internaional World Wide Web Conference, pp: 57-65.

Nasraoui, O., R. Krishnapuram, H. Frigui and A. Joshi, 2004. "Extract-ing Web User Profiles Using Relational Competitive Fuzzy Clustering", in the International Journal of Artificial Intelligence Tools, 9(4).

Nasraoui, O. and R. Krishnapuram, 2009. "A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining", in the International Journal of Computational Intelligence and Applications, special issue on Internet intelligent systems, 2(3): 339-348.

Nasraoui, Cardona, Rojas and Gonzalez, 2011. "Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm," Proc. Workshop Web Mining as a Premise to Effective and Intelligent Web Applications (IbKDD '11): 71-81.

Raiyani, S.A. and S. Jain, 2012. "Enhance Preprocessing Technique Distinct User Identification using Web

Log Usage data", in the International Journal of Computer Science & Communication Networks, (2): 526-530.

Srivastava, J., R. Cooley, M. Deshpande and N. Tan, 2010. "Web Usage Mining: Discovery and Applications of Usage Patterns from Ib Data,"SIGKDD Explorations, 1(2): 1-12.

Sheetal, A. Raiyani, Shailendra Jain and G. Ashwin Raiyani, 2012. "Advanced Proprocessing using Distinct User Identification in web log usage data", in the International Journal of Advanced Research in Computer and Communication Engineering, 1(6).

Sudheer Reddy, K., G. Partha Saradhi Varma and M. Kantha Reddy, 2014. "An Effective Preprocessing method for web usage mining", in the International Journal of Computer Theory and Engineering", Vol.6.

Uma Maheswari, B. and P. Sumathi, 2014. "A New Clustering and Preprocessing for web log mining", in the Journal of Computing and Communication Technologies, pp: 25-29.

Wu, Z. and M. Palmer, 2007, "Verb Semantics and Lexical Selection," Proc. 32nd Ann. Meeting of the Assoc.

Computational Linguistics, pp: 133-138.

Zaiane, O., M. Xin and J. Han, 2008. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs,"Proc. Advances in Digital Libraries, pp: 19-29.