International Academic Institute for Science and Technology

International Academic Journal of Innovative Research
Vol. 9, No. 1, 2022, pp. 52-60.

ISSN 2454-390X

International Academic Journal of Innovative Research

www.iaiest.com

Web Mining Using Improved Apriori Algorithm

C.K. Mohammed Malik [a]

[a] *Assistant Professor, Department of Computer Science and Engineering, Eranad Knowledge City Technical Campus, Manjeri, Malappuram, Kerala. E-mail: malikckbe@gmail.com*

Abstract

In this study, we will be concentrating on one of the more recent advancements in data mining, specifically mining online usage. The purpose of web use mining is to gain usable knowledge from the data that web servers keep about the actions of its visitors by mining the data that is stored on such servers. By using the association rule generation in the Web domain, the pages that are most frequently referenced together can be combined into a single server session. This is possible because of the interconnected nature of the Web. In association rule mining, a technique known as frequent set mining is one of the methods that may be used to discover regular patterns from a web log file. When it comes to mining the usage of the web, the term "association rules" refers to groups of web pages that are accessed together and have a support value that is higher than a given threshold. The support can be expressed as a proportion of total transactions that match a particular pattern. With the aid of the presence or absence of association rules, web designers are able to effectively reconstruct the websites they have created for their clients. In this research, we have introduced a method called Aprior for the purpose of extracting frequent patterns from online log files. The findings of the experiments that were carried out on data relating to people's use of the website indicate that general sequential patterns or frequent item sets are more suitable for use in Web customization and recommender systems.

Keywords: Web Usage Mining, Frequent Patterns, Frequent Set Mining.

DOI:10.9756/IAJIR/V9I1/IAJIR0917
Received: 16 December 2021/Revised: 31 January 2022/Accepted: 25 February 2022/Published: 30 March 2022

Introduction

As both the volume of data on the World Wide Web and its consumption continues to rise, there has been a significant growth in the number of study conducted on the extraction and discovery of information buried in connection with it. According to Internet World Stats, the number of people using the internet has increased by 528.1% between the years 2000 and 2011. The abundance of information, on the other hand, frequently makes it difficult for users to locate data that is both helpful and pertinent to their needs. An technique that has the ability to remedy this issue is to conduct an analysis of the navigational browsing behaviors of users. Additionally, it assists companies in providing customised recommendations of web pages to users based on the interests that the user now possesses.

Web mining refers to the process of using data mining methods to automatically retrieve, extract, and analyse information from websites, webpages, and other web-based services for the purpose of knowledge discovery. There are three distinct categories of web mining. There are three main types of web mining: mining content, mining structure, and mining usage. Web content mining is the process of extracting valuable knowledge from websites' data, documents, and services. As its name implies, Web Structure Mining extracts information by analyzing the web's internal connection structure. The graph of a link's structure within a site or between sites is what we mean by "structure." Data recorded in a web server's logs can be mined for useful insights using a technique called Web Usage Mining. Web usage mining is broken down into four distinct phases. Data collection, data preprocessing, pattern analysis, and pattern analysis are the four phases.

The World Wide Web (WWW) is continuing to undergo amazing growth both in terms of its role as a gateway to knowledge and as a platform for commercial activity. Web mining is the process of extracting information that is interesting and useful, as well as information that may be inferred from artifacts or activities that are connected to the World Wide Web [16]. "Web mining" refers to the process of extracting useful information from the World Wide Web and doing an analysis of such information. On the basis of the findings from a number of different research projects, we are able to essentially classify web mining into three categories: content mining, structure mining, and use mining [8]. Content mining refers to the process of extracting information about a website's content. [9]. It shouldn't come as much of a surprise to learn that the process of extracting information from the text of web documents is referred to as "web content mining." Web structure mining is the process of gleaning information that can be inferred from the organization of data [17]. Web use mining is the act of discovering and evaluating the patterns of user access to an information system on the web. [5] This can be done through a method known as "web use discovery." The patterns of web usage that have been uncovered by analyzing web data have been put to use for a wide range of purposes, such as web personalization, site alteration, system enhancement, business intelligence, and user characterization, amongst others.

The topic of Web usage mining is the exclusive focus of the discussions in this paper. When web servers receive requests for resources, they immediately begin recording and compiling data on the users who made those requests. The user behavior as well as the structure of the website can be better understood by analyzing the Web access logs. The process of employing data mining techniques in order to discover usage trends from Web data is referred to as web usage mining. This is done in order to better appreciate the requirements of web-based applications and to provide better support for those applications. The mining of online usage can be broken down into three parts, as indicated in figure 1: the preprocessing stage, the pattern discovery stage, and the pattern analysis stage.

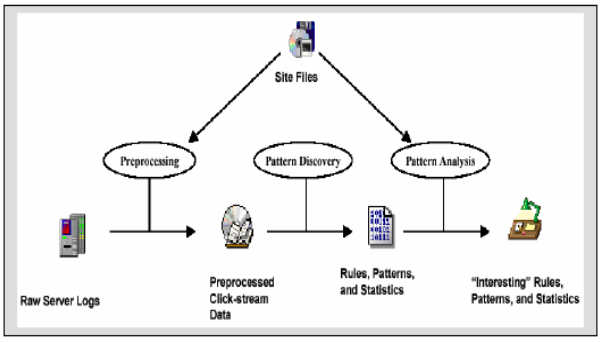


Figure 1: A High Level Web Usage Mining Process
(Adapted from Srivastava *et al.*, 2000)

Data for Web mining can be gathered in a number of different ways, including directly from the server, from the client, through a proxy, or as a consolidated set. The content and collection methods of these datasets may be different from one another. From single-user, single-site navigation patterns to multi-user, multi-site access patterns, all of these scenarios are reflected in the usage data acquired at various sources. The client side behavior in relation to the pages provided by the web server cannot be inferred with any degree of accuracy from the web server log. Data that has been cleansed and prepared for analysis can be put to work in a variety of ways, including discovery and analysis of patterns, Web usage statistics, and the creation of association and sequence rules. Many researchers have spent time and energy figuring out how to pull useful information out of Web logs, and the results have been put to use in everything from making websites more user-friendly to helping businesses run more smoothly [15, 20, 21, 22, 25].

Among the many aspects of Web mining, pattern discovery is essential. Data mining, machine learning, statistics, and pattern recognition are just few of the fields whose methods and algorithms come together in pattern discovery. Information about website users can be extracted most effectively through the use of statistical methods[11].

Cluster analysis is a method for organizing large amounts of data into subsets based on their shared properties, such as users or web pages. Data classification is the process of assigning each data point to a certain category. The goal of the sequential pattern technique is to identify the inter-session pattern in which one set of items is preceded by and then followed by another set of items in a series of episodes or sessions in chronological order. The association rule generation in the Web domain allows for the grouping of often referred-to pages into a single server session. Discovering unordered correlation between items in a transaction database can be done with the help of association rule mining techniques [12]. Association rules, as defined by [11]'s creators, are groups of pages that are frequently visited together and have support values greater than a predetermined cutoff. The backing is the share of transactions that include a particular pattern. With the aid of the presence or absence of the association rules, Web designers can effectively rearrange their Websites. Classification techniques can be utilized as a trigger for scheduling documents to reduce user perceived latency while loading a page from a remote site. A subset of association rule mining known as "frequent set mining" is used to the task of identifying all minimally supported pages that receive a high volume of user traffic.

The entire process of mining Web traffic culminates in pattern analysis. The output of pattern discovery is sifted through to remove irrelevant rules and patterns, leaving just the intriguing ones.

To start, we define a few key phrases in accordance with the World Wide Web Consortium's Web Characterization Activity's published guidelines. A user is a person who uses client software to obtain and display content from the World Wide Web in an interactive manner. A web page is a group of related web resources that are meant to be rendered all at once and have the same Uniform Resource Identifier (URI). Consequently, a URI is the label for a digital or concrete item. A user session is a collection of clicks by a user across one or more web servers, while a user visit or server session is a collection of clicks by a user in a single web server. Last but not least, a click-stream is the rendered web document in a particular client application arranged by time.

An entry in a log file typically contains the user's request URL, the user's IP address, the time the request was made, and maybe other fields as indicated in Table 1.

Table 1: The Entry of a Web Server Log Following the IIS Format

Session no	IP Address	Time | Method |URL Protocol |Status | Size | Referrer Agent
1	123.456.78.9	-[25/Apr/2021:03:04:41-0500]GET *https://www.msn.com/en-in/shopping* HTTP/2.0" 2000 3280 - Mozilla/78.10 (Win8.1)
2	123.456.78.9	-[25/Apr/2021:03:05:34 -0500] "GET *http://www.msn.com/en-in/shopping/items* HTTP/1.0" 2000 2040 Mozilla/78.10 (Win8.1)
3	123.456.78.9	-[25/Apr/12021:03:05:39 -0500] "GET *http://www.msn.com/en-in/shopping/medicine* HTTP/1.0" 2000 4120- Mozilla/78.10 (Win8.1)

In this context, data regarding user access has a few advantages over other types of data provided by users, such as the fact that it is more specific. In particular, they are automatically collected "over the user's shoulder," they are of a huge volume delivering a rich data source, they cannot be changed or removed by the users, and they protect the users' privacy as long as the user identity is not disclosed and cannot be determined. When processing web access

logs, some of the limitations that can occur include missing document references as a result of caching mechanisms, incorrect interpretation of IP addresses as a result of the use of proxy servers that assign the same IP address to all users, and the difficulty of updating the user model utilizing only access information. These limitations can occur because of the following: the use of proxy servers that assign the same IP address to all users; the use of caching mechanisms; and the use of proxy servers.

There has been a substantial amount of work done in the area of Web Usage Mining[11], which directly or indirectly addresses the issues involved in the extraction of web navigational patterns [26], ordering relationships [18], prediction of web surfing behavior [23], and clustering of web usage sessions based on web logs, possibly supplemented by web content or structure information. Because the raw data are all presented in the form of sequences rather than vectors, there is going to be a big obstacle in the way of resolving any of these issues. Because of this, mining techniques that are particularly well established for vector-type data cannot be utilized directly. In the past few years, research has been conducted looking specifically at the issue of web usage clustering. The authors of paper [13] investigated the concept of clustering online usage sessions by employing a generalization-based methodology and making use of a hierarchical clustering algorithm known as BIRCH. In the paper [24], the concepts of a path feature space and path angles are presented, and the K-means algorithm is used to cluster the various pathways as a final step. On the other hand, a conclusive solution has not yet been found.

In their study on visitor click sequences, Jesperson et al. [14] suggested using a hybrid method of analysis. Lumberjack, which was introduced by Chi et al. [10], constructs user profiles by integrating user session clustering with conventional statistical analysis through the use of the K-means algorithm. The user access records could be segmented using computational intelligence paradigms such as self-organizing maps and fuzzy C-means algorithms, and the user access trends could be analyzed using self-organizing maps. Abraham and Wang et al.[15] have unequivocally demonstrated the significance of the clustering approach in order to investigate the user access patterns [1]. Using the Apriori algorithm, this research generates patterns of user behavior that are frequently observed on a website. Mining all of the frequent sets was just the first step, but it was also the phase that presented the most challenge. The Apriori method, however, solves the entire problem of association rule mining.

This paper is structured in the following manner: The topic of frequency set mining is covered in Section 2. In Section 3, we will discuss the Apriori method. Following an explanation of the Apriori algorithm's working approach in Section 4, as well as its most significant results and performance, the conclusion and recommendations for further research are offered in Section 6.

Frequent Set Mining

The mining of association rules is one of the most common topics in Data Mining[7], but frequent sets play a crucial part in a wide variety of other tasks that aim to discover interesting patterns in databases. One of the most fundamental jobs in Data Mining is the identification of groups of things, products, symptoms, features, etc. that frequently occur together in the supplied database.

The requirement to study so-called supermarket transaction data, or customer behavior in terms of purchased products, was the primary impetus for searching frequent sets[4]. Frequent product sets characterize the frequency with which particular products are bought together. Let's τ formally that is a collection of things.

T = (tid, I) is the pair that represents a transaction over, where tid is the transaction identifier and I is the set of items involved in the transaction.

Information stored in a database D over is the collection of all possible transactions (τ) over where each transaction is uniquely identified. Whenever τ is obvious from the surrounding text, we omit.

If X I, then the set X is said to be supported by the transaction T = (tid, I). If you have a set X in the database D, its cover is the set of transaction identifiers for all the transactions in D that support X. The number of operations in the cover of a set X in D is its support in D. Probability that a set X occurs in a transaction, or the frequency of X in D, is the support of X divided by the total number of transactions in D. When it's already obvious, we don't include the letter D.

When its support is greater than some absolute minimal support threshold σ_{abs}, where $0 \leq \sigma_{abs} \leq |Dj|$, we say that a set is frequent. D is a database of purchases over the products in set I, and σ is the lowest acceptable level of support.

$F(D, \sigma) := \{X \subseteq \tau \mid support(X,D) >= \sigma\}$ as well as only F if D and are obvious.

Apriori Algorithm

At the same time as the frequent set mining problem was presented, the first solution to tackle it was proposed; this approach would be known as AIS [2]. An enhanced version of the algorithm, dubbed "Apriori," was developed soon after. The use of the monotonicity condition of the support of sets was the key advance [4]. Mannila et al. [19] independently proposed the same method. After that, the two pieces were fused into one [2]. It's important to keep in mind that mining all frequent sets was just the first and, arguably, most challenging step of the entire association rule mining problem, which is solved in its entirety by the Apriori algorithm.

The algorithm is also known as the level-wise algorithm. Agrawal and Srikant [4] first put forth the idea. Popularity-wise, it is the algorithm of choice for discovering all the frequent sets. The property of downward closure is exploited in this method. The algorithm is a bottom-up one, in that it searches lattice levels vertically upwards, as the name suggests. The method is sophisticated because it thoughtfully eliminates numerous sets that are not expected to be frequent sets before scanning the database at every level.

The most common groupings of things are identified in the first pass of the algorithm, which merely counts the occurrences of each item. For the sake of argument, let's imagine that the next pass, pass k, also includes two stages. To begin, the candidate itemsets Ck are generated a priori by using the frequent itemsets Lk-1 discovered in the (k-1)th pass. The next step is to go through the records and tally the votes cast for each contender in Ck. In order to quickly tally, we need an efficient way to find all of the candidates in Ck that are part of a specific transaction, t. In order to ensure that all subsets of the candidate sets are already known to be a frequent itemset, we prune the set of candidate itemsets. The most crucial steps of this algorithm are the generation of candidates and the subsequent pruning. In the following paragraphs, we'll each describe one of these procedures in detail.

Unless specified otherwise, the elements of transactions and sets are maintained in lexicographic order. Algorithm 1 details the Apriori algorithm's set mining process.

A. Candidate Generation

We wish to create a superset of the set of all frequent k-itemsets given Lk-1, the set of all frequent (k-1)-itemsets. The idea underlying the a priori candidate-generation process is that all subsets of an itemset X have minimum support if it. Below is a description of the candidate generating process:

Algorithm 1.

Gen − candidate- items with the given L_{k-1} *as follows:*
$C_k = \varnothing$
For all itemsets $l_1 \in L_{k-1}$ *do*
For all itemsets $l_2 \in L_{k-1}$ *do*
If $l_1[1] = l_2[1] \wedge \ldots \wedge l_1[k-1] < l_2[k-1]$
Then $c = l_1 1]$, $l_1[2] \ldots l_1[k-1]$, $l_2[k-1]$
$C_k = C_k \cup \{c\}$

B. Pruning

Extensions of (k-1)-itemsets that are not deemed to be common are removed from consideration for counting support during the pruning process. Here is a description of the pruning algorithm:

Algorithm 2.

Prune (C_k)

For all $c \in C_k$
For all (k-1) subsets d of c do
if $d \subset L_{k-1}$
then $C_k = C_k \setminus \{c\}$

Algorithm 3.

Initialize: k: = 1,C_1 = all the 1- itemsets;
Read the database to count the support of C_1 to determine L_1.
L_1: {frequent 1- itemsets};
K: = 2 ; // k represents the pass number //
While ($L_{k-1} \neq \varnothing$) do
Begin
C_k: = Gen – candidate- items with the given L_{k-1}
Prune(C_k)
For all transactions $t \in T$ do
Increment the candidates in C_k that are contained in t.
L_k: = All candidates in C_k with minimum support ;
K: = k+1;
End
Answer: = $\cup_k L_k$;

Experimental Results

Timer frequency The challenge of mining associations between rules is a well-known one[4]. Here, the server-side web log file is retrieved, and a record is produced in the database for each user's pageviews. As a result, the user's history of browsing the web is stored in database D. Each database record represents a session's worth of web page views. The goal of this mining project is to identify every subset of D's most-visited websites. In the field of association mining, these subsets are referred to as itemsets. Using Table1 as an example, think about the database of web logs. We simply keep track of the URLs that were actually viewed by each user, not their full journey. The goal of this mining operation is to identify all clusters of frequently visited pages. All k-itemsets that have appeared in at least three separate user transactions (minimum support = 3) are displayed in Tables 2, 3, and 4. Among the most common sets of items, ABC, AF, and CF are the most common.

On a sample of 26 web pages, we ran the APriori association mining algorithm. It was found that there were a total of 150 user sessions, with an average of 12 unique nodes each session. Our mining with a 9% minimum took 30 seconds. Support. An example frequent sets found are shown below:

Table 2: Set Database Extracted from Web Log File

User id	Pages visited
User 1	ADE
User 2	ACDEFG
User 3	ACD
User 4	ABCDE
User 5	ACDEF
User 6	AD

Frequent sets generated are given in Table 3,4,5 and 6.

Table 3: Frequent 1-Itemsets

Page visited	Support	Frequency
A	6	100%
C	4	67%
D	6	100%
E	4	67%

Table 4: Frequent 2-Itemsets

Pages Visited	Support	Frequency
AD	6	100%
AE	4	67%
CD	4	67%
CE	3	50%
DE	4	67%

Table 5: Frequent 3-Itemsets

Pages Visited	Support	Frequency
ACD	4	67%
ACE	3	50%

Table 6: Frequent 3-Itemsets

Pages Visited	Support	Frequency
ACDE	3	50%

An Example frequent set found is shown below:

http://localhost/lib-01/language.html

http://localhost/lib/language.html

http://localhost/lib/os/windows.html

http://localhost/lib/animation/3dsoftware.html

Performance Analysis

Empirical studies have demonstrated that it is difficult to determine which metric to use for different problems; yet, each of the metrics includes particular properties that measure various parts of the algorithms that are being evaluated. Despite this, it might be difficult to choose which meter to employ. Due to the huge weighted disparities that frequently occur between a projected value and its actual counterpart or otherwise, it is frequently challenging to determine which metric is the most appropriate one to use when evaluating algorithms in web log information. The success of machine learning algorithms is evaluated based on their prediction accuracy, which is frequently inadequate in the situation of imbalanced data because mistake costs might vary quite a bit. The following line chart compares the total time needed for different minimum support levels.

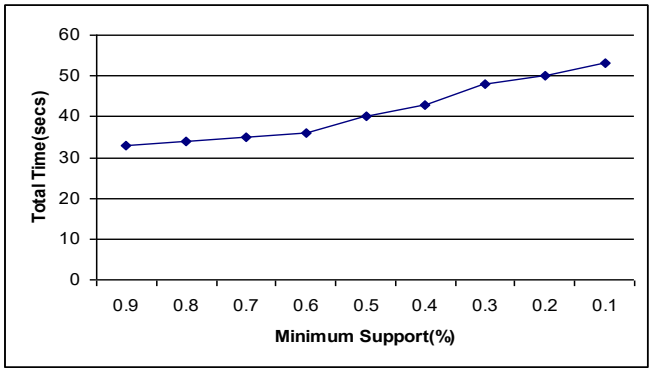


Fig. 1: Frequent Set Mining

Conclusion

There is a lot of interest in the topic of "Web Usage Mining," and its proponents believe it will open up exciting new opportunities for online enterprises. The Web Usage Mining data processing stage is a complex undertaking. In this research, we use the Apriori method to extract patterns in web server logs. Since the Apriori algorithm is a standard data mining method for mining frequent sets, it was chosen for this task of mining Web usage. This experimental investigation yielded useful data that can be applied to the enhancement of websites in light of typical user behavior. Our web data mining algorithm has produced some preliminary results, which we present here in the form of typical user behavior on a website. We also came to the realization that the distinction between web usage mining, web content mining, and web structure mining is a phony one. Many web usage mining methods rely on data gleaned from web content and web structure mining processes. The majority of recent studies have zeroed down on just one of these areas. A holistic strategy that integrates prior learning from other fields is required for effective deployment of online usage mining.

References

Abraham A, i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems. *The IEEE International Conference on Fuzzy Systems FUZZ-IEEE'03*, IEEE Press, pp. 1129-1134, 2003.

Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases. *In Buneman, P. and Jajodia, S., editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,* volume 22(2) of SIGMOD Record, pages 207–216.ACM Press, 1993.

Agrawal, R., Mannila, H., Srikant, R., Nandita Agrawal, Anand Jawdekar, "User-Based Approach For Finding Various Results In Web Usage Mining", *in IEEE* 2015.

Hong-Yi Chang, Yih-Jou Tzang, Zih-Huan Hong, "A Hybrid Algorithm for Frequent Pattern Mining Using Map Reduce Framework", *in IEEE* 2015.

Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya, "Association Rule Mining for Web Usage Data to Improve Websites", *In IEEE* 2014.

Baumgarten M., A.G. Bchner, S.S. Anand, M.D. Mulvenna, J.G. Hughes,Navigation Pattern Discovery from Internet Data. M. Spiliopoulou, B. Masand (eds.) Advances in Web Usage Analysis and User Profiling, Lecturer Notes in Computer Science, Vol. 1836, Springer-Verlag, ISBN: 3-540-67818-2, July 2000.

Burdick, D., Calimlim, M., and Gehrke, J., MAFIA: A maximal frequent itemset algorithm for transactional databases. *In Proceedings of the17th International Conference on Data Engineering,* pages 443–452. IEEE Computer Society, 2001.

Chakrabarti S, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers, 2003.

Chang, G., Healey, M.J., McHugh, J.A.M., Wang, J.T.L, Web Mining, Mining the World Chi E.H., Rosien A. and Heer J, Lumber Jack: Intelligent Discovery and Analysis of Web User Traffic Composition. *In Proceedings of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles,* Canada, ACM Press, 2002.

R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000. ivastava. *Data preparation for mining worldwide Web browsing patterns. Knowledge and Information Systems,* 1(1), 1999.

Y. Fu, K. Sandhu, and M. Shih, 'A Generalization-Based Approach to Clustering of Web Usage Sessions', *In Intl. Workshop on Web Usage Analysis and User Profiling (WEBKDD'99),* pages 21--38, San Diego CA, August 1999.

Jespersen S.E., Thorhauge J., and Bach T, Hybrid Approach to Web Usage Mining, Data Warehousing and Knowledge Discovery, LNCS 2454, Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.), pp. 73-82, 2002.

Kitsuregawa, M., Toyoda, M., Pramudiono, I, Web Community Mining and Web Log Mining: Commodity Cluster Based Execution. *Proceedings of the 13th Australasian Database Conference (ADC'02),* Australia, 2002.

Kosala R and Blockeel H, Web Mining Research: A Survey, ACM SIGKDD Explorations, 2(1), pp. 1-15, 2000.

Kuo Y.H., M.H. Wong, Web Document Classification Based on Hyperlinks and Document Semantics. *PRICAI 2000 Workshop on Text and Web Mining,* Melbourne, pp. 44-51, August 2000.

H. Mannila and C. Meek, 'Global partial orders from sequential data'*, In Proc. 6th Intl. Conf. on Knowledge Discovery and Data Mining (KDD2000),* pages 161-168, Aug 2000.

Mannila, H., Toivonen, H., and Verkamo, A, Efficient algorithms for discovering association rules. In Fayyad, U. and Uthurusamy, R., editors, *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases,* pages 181–192. AAAI Press, 1994.

Masseglia, F., Poncelet, P., Cicchetti, R. An Efficient Algorithm for Web Usage Mining. *Networking and Information Systems Journal (NIS),* vol. 2, no. 5-6, pp 571-603, 1994.

Perkowitz, M., Etzioni, O, Adaptive Web Sites: Automatically Synthesizing Web Pages. *Proceedings of the 15th National Conference on Artificial Intelligence,* pp. 727-732, 1998.

Pirolli, P., Pitkow, J., Rao, R, Silk From a Sow‘s Ear: Extracting Usable Structures from the Web. *Proceedings on Human Factors in Computing Systems (CHI’96),* ACM Press, 1996.

J. Pitkow and P. Pirolli, Mining longest repeating sub sequences to predict World Wide Web surfing’, *In Proc 2nd USENIX Symposium on Internet Technologies & Systems (USITS'99),* Oct 1999.

C. Shahabi, A.M. Zarski, and V. Shah J. Adibi,’ Knowledge discovery from users web navigation’. *In Proc. Conf on Research Issues in Data Engg.,* pages 20.29, 1997.

Smith K.A. and Ng A, Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems,* Volume 35, Issue 2, pp. 245-256, 2003.

M. Spiliopoulou and L.C. Faulstich. WUM, A tool for web utilization analysis’, *In Extended version of proc. of EDBT Workshop WebDB'98,* pages 184-203. Springer Verlag, 1999.