



An Elegant Draw Near to Improve the Design of an E-commerce Website Using Web Usage Mining and K-Means Clustering

S. Divya Rajan

M. Tech. scholar, Department of Computer Science and Engineering,
Rungta College of Engineering and Technology, BHILAI, (CG) INDIA
sdivyarajan@gmail.com

Mr. Neelabh Sao

Associate Professor, Department of Computer Science and Engineering,
Rungta College of Engineering and Technology, BHILAI, (CG) INDIA
neelabhsao@gmail.com

Abstract— Web Mining is an enormous field that helps us to understand range of concepts of different fields. Web Usage Mining Techniques are attempted to motive about diverse materialized issues of Business Intelligence which include marketing proficiency as domain knowledge and are specifically designed for electronic commerce purposes. The growing reputation of e-commerce makes data mining requisite technology for several applications, especially online business competitiveness. The World Wide Web provides profuse raw data in the form of web logs. Nowadays many business applications utilizing data mining techniques to pull out useful business information on the web evolved from web searching to web mining. This paper introduces a web usage mining intellectual system to provide nomenclature on user information based on transactional data by applying association and K-means clustering data mining algorithms.

Keywords- Business Intelligence, CRM, e-Commerce, Web mining, Web usage mining, Web personalization

I. INTRODUCTION

As in classical data mining, the aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. There has been huge interest towards web mining *Sudhamathy G et.al* [4]. In web mining, huge web data

acts as the dataset. Web data include information like data about web log, web documents, web structure, and user profiles. Web Mining can be defined accurately on the basis of two concepts. One is process-based and the other is data-based. Data-based being the commonly used. In this perspective, web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. There are no differences between web mining and data mining compared in general. All of web data can be mined mainly in three different dimensions, which are web content mining, web structure mining, and web usage mining.

There are several reasons for the emergence of web mining *Jiawei Han et.al* [10]. First of all the World Wide Web is huge and effective source for data mining and data ware housing. The size of web is increasing continuously. It contains data in the order of several terabytes. Many organizations, individuals or societies provide their public information through web. Also, the content of the web pages are much more complex than any other traditional text documents. Web Mining can be broadly divided into three categories as shown in Fig 1 according to the kinds of data to be mined:-

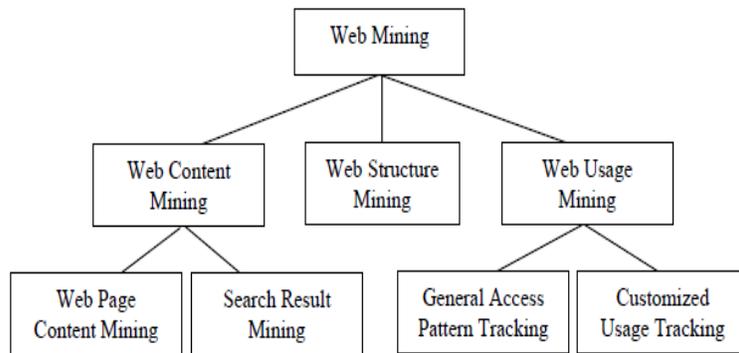


Figure 1. [17] Taxonomy of Web Mining

WWW can be accepted as a huge digital library and web mining as librarian of this digital library. There are several application areas of web mining; the important ones are listed in Figure-2. The most popular application area of web mining is e-commerce (business-to-customer) and web based customer relationship management. Web usage mining is most dominant application in this context. With the web mining, it is possible to record customer behavior for web-based business. It is also feasible to adapt web sites based on interesting patterns as a result of analysis on user navigation patterns *Kosala. R et.al* [12]. Web site topology can be customized to provide better facilities for the site user *Jiawei Han et.al* [10].

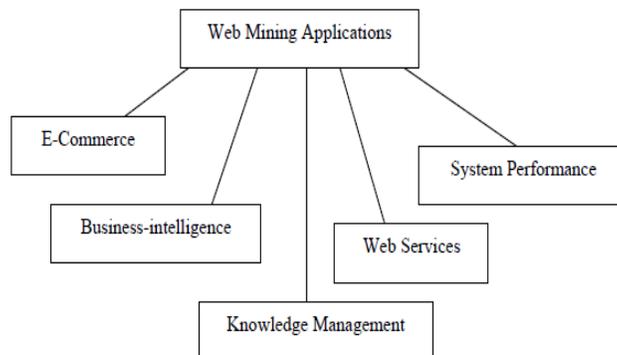


Figure 2. [15] Application Areas of Web Mining

A lot of previous work has focused on Web data clustering *G. Biswas et .al* [14]. Web data clustering is the process of grouping Web data into "clusters" so that similar objects are in the same class and dissimilar objects are in different classes. Its goal is to organize data circulated over the Web into groups or collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding users' navigation behavior, improving information retrieval and content delivery on the Web. The aim of web mining is to discover and retrieve useful and interesting patterns from a large dataset. Mining web log data can be used as a way to evaluate effectiveness of a website. Proposed algorithm is used to generate association rules that associate the usage pattern of the clients for an e-commerce website. In the proposed work we have combined the association mining with the clustering instead of mining association rules from the web log data directly we have mined the clusters. The goal of clustering is to organize data circulated over the Web into groups or collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding users' navigation behavior, improving information retrieval and content delivery on the Web.

The remainder of this paper is organized as follows: Section 2 provides a brief review of the related work. In Section 3, we explain problems in existing systems. In Section 4, we introduce our proposed algorithm and an illustration of the algorithm. Section 5 includes experimental setup and results. Finally, I conclude my work.

II. RELATED WORK

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest.

Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For example *B.Naveena Devia et.al* [5] has introduced a web usage mining intelligent system to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service which enables direct access of website functionalities to the third party.

Santosh kumar et. al. [6] concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of memory usage and time usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm.

Patel et al [7] discusses the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented Web Usage Mining applications and some Web Mining software.

An introduction to the field of Web mining and examination of existing as well as potential Web mining applications applicable for different business function, like marketing, human resources, and fiscal administration has been presented in [8].

Purandare [9] presents an overview of the web mining concept and shows how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. This paper contributes towards the various areas containing web sites on internet, which can make best use of different web mining

techniques to improve their business decisions based on the user behavior analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth.

In [11] the high level process of Web Usage Mining using basic Association Rules algorithm call Apriori Algorithm has been implemented. Web Usage Mining consists of three main phases, namely Data Pre-processing, Pattern Discovering and Pattern Analysis. Server log files become a set of raw data where it's must go through with all the Web Usage Mining phases to producing the final results. Here, Web Usage Mining, approach has been combining with the basic Association Rules, Apriori Algorithm to optimize the content of the serve log data. Finally, this paper will present a finding association Rule from server log which are useful in many application like cache for web page, Marketing, Targeted Advertising etc.

III. PROBLEM IDENTIFICATION

The explosive expansion of the World Wide Web (WWW) in recent years has turned the web into the largest source of available online data.

- There are various unrelated topics in the web page that may lead to confusion and make it difficult for the user to find the information he is looking for.
- The design of the whole site (interface, content, structure, usability, etc.) is one of the most important aspects for any institution that wants to survive in the cyberspace.
- Recognize the user navigation behavior of the site and identify the most recurrent used link ad pattern of using the features available in the site.

All these information is available online but are hidden for the users. Presently, there is no powerful tool that can analyze this hidden information and this Research work uses web usage mining (WUM) Apriori based approach for analyzing the visitor browsing behavior.

IV. PROPOSED SYSTEM

The main goal of the proposed system is to identify usage pattern from web log files of a website, collections of items bought by customers, or details of a website frequentation. In this paper we proposed a new algorithm which combines the concept of association mining and clustering instead of mining association rules from the web log data directly we have mined the clusters selected by user. Figure 3 represents proposed approach.

Algorithm Description

Input: A web log database
The Minimum-Support threshold

Output: Frequent item sets

Method:

- 1) Scan the database D and partition the transaction table into clusters using K-means algorithm. Apply the method from step 2 to 6 on user selected cluster.
- 2) The set of frequent 1 item sets say L1, can then be determined. It consists of candidate 1 item Sets which satisfy minimum support

- 3) To discover the set of frequent 2- item sets
- 4) The algorithm iterates to find upto n- frequent item sets
- 5) From user selected cluster find out the n-frequent item sets.

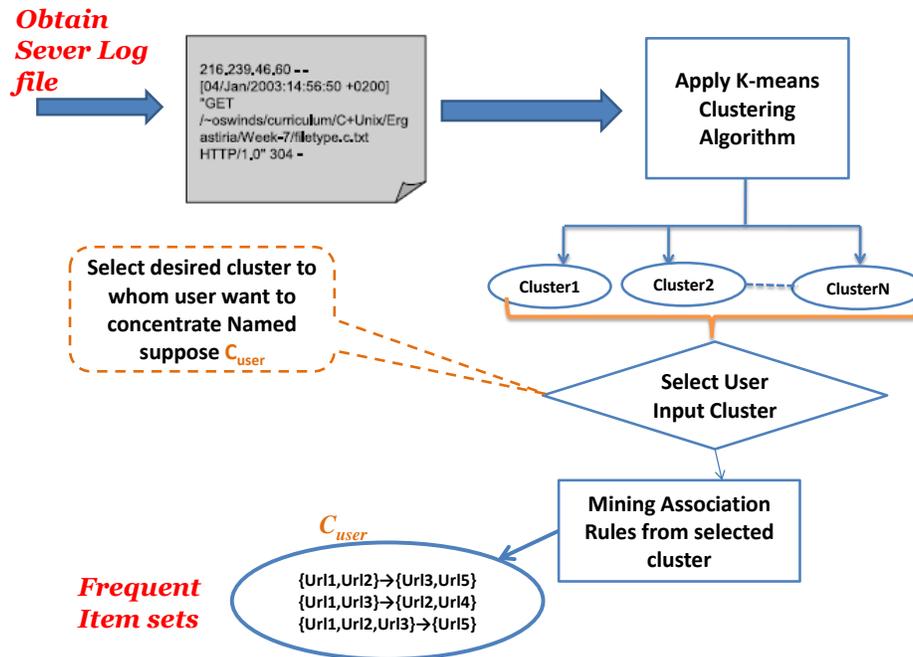


Figure 3: Proposed Approach

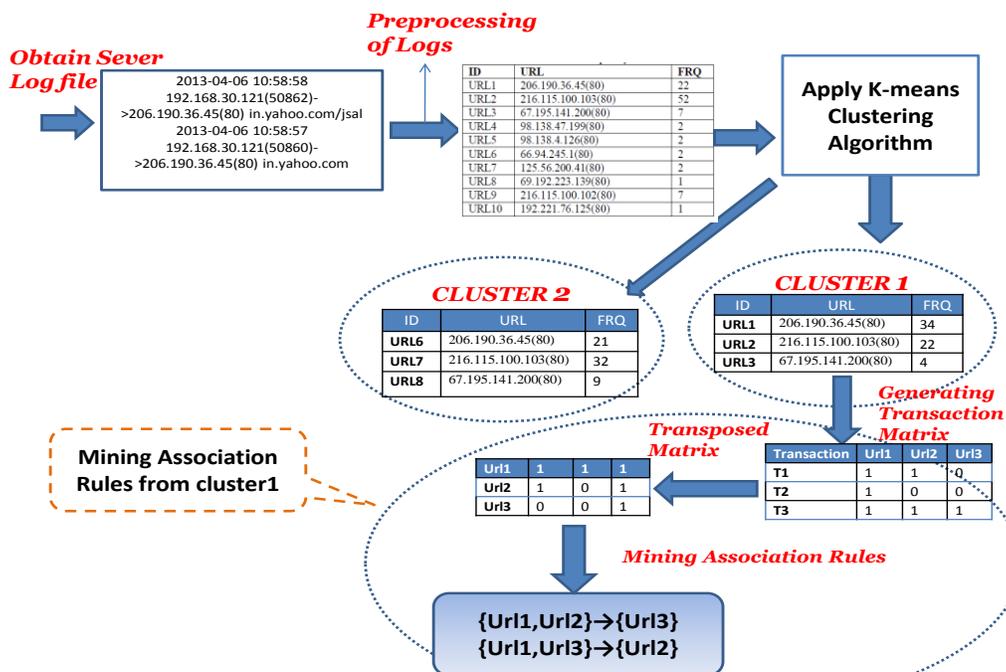


Figure 4: Example illustration

A. Clustering Algorithm

Clustering is a technique to search hidden patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. A popular clustering method that minimizes the clustering error is the k-means algorithm. It partitions the input dataset into k clusters. First select k initial centers based on desired number of clusters. The user can specify k parameter value. Each data point is assigned to nearest centroid and the set of points assigned to the centroid is called a cluster. Each cluster centroid is updated based on the points assigned to the cluster. The process will be repeated until the centroids remain the same or no point changes clusters. In this algorithm mostly Euclidean distance is used to find distance between data points and centroid.

Algorithm: The k -means clustering algorithm

Input: D: {d1,d2....dn} \\set of n items
K //Number of desired clusters

Output: A set of k-clusters.

Steps:

1. Arbitrarily choose k-data items from D as initial centroids;
2. **Repeat** assigns each item d_i to the cluster which has the closest centroid,
Calculate new mean for each cluster;

until convergence criteria are met.

3.3.2 Association Rule Mining

Given a server log files that represent user activities, the main purpose of Association Rules is to generate all Association Rules that have support and confidence greater than the user specified minimum support (called min_sup) and minimum confidence (called min_conf) respectively. An algorithm for finding all Association Rules, henceforth, referred to as the Apriori algorithm[2].

In Apriori algorithm, discovery of association rules require repeated passes over the entire database to determine the commonly occurring set of data items. Therefore, if the size of disk and database is large, then the rate of input/output (I/O) overhead to scan the entire database may be very high. We have proposed a new Algorithm, which improves the Apriori algorithm for repeated scanning of large databases for frequent itemsets generation. In our algorithm, transaction dataset will be used in the transposed form and the description of proposed algorithm is discussed in the following sub-sections.

Procedure Gen_candidate_itemsets (L_{k-1})

```

Ck = Φ
for all itemsets I1 ∈ Lk-1 do
for all itemsets I2 ∈ Lk-1 do
if I1[1] = I2[1] ^ I1[2] = I2[2] ^ ... ^ I1[k-1] < I2[k-1] then

```

$c = I_1 [1], I_1 [2] \dots I_1 [k-1], I_2 [k-1]$

$C_k = C_k \cup \{c\}$

End Procedure

Procedure Prune (C_k)

for all $c \in C_k$

 for all (k-1)-subsets d of c do

 if $d \notin L_{k-1}$

 then $C_k = C_k - \{c\}$

End Procedure

Algorithm: Association Rule Mining for each cluster

1. Read the database to count the support of C_1 to determine L_1 using sum of rows.

2. $L_1 =$ Frequent 1- itemsets and $k = 2$

3. While (k-1 \neq NULL set) do

 Begin

$C_k :=$ Call Gen_candidate_itemsets (L_{k-1})

 Call Prune (C_k)

 for all itemsets $i \in I$ do

 Calculate the support values using dot-multiplication of array;

$L_k :=$ All candidates in C_k with a minimum support;

$k := k + 1$

 End

4. End of step-3

End Procedure

V. EXPERIMENTAL RESULTS AND SCREENSHOT

Data Set: The web log data considered for evaluation is collected from web server during the period of May to August, 2011. Initially the log file consists of 9464 raw log entries with noisy entries like gif, jpeg etc which are not necessary for web log mining. So data cleaning is performed to remove the unnecessary log which will reduce the processing in determining the web usage pattern. This cleaning phase involves the removal of records with graphics and videos format such as gif, JPEG, etc., and records with robots traversal is also removed. A sample of dataset is shown below.

entry id	url	timestamp
1	146387 http://osa3.blog.ocn.ne.jp/project/2004/10/post_13.html	2004-10-15 00:00:00
2	146389 http://blog.livedoor.jp/kingcurtis/archives/8211528.html	2004-10-18 00:00:00
3	146524 http://valuation.jugem.cc/?aid=231	2004-10-13 00:00:00
4	146390 http://sugamo.jugem.cc/?aid=268	2004-10-14 00:00:00
5	146540 http://numazu.jugem.jp/?aid=333	2004-10-13 00:00:00
6	146541 http://radkyudan.jugem.jp/?aid=372	2004-10-13 00:00:00
7	146393 http://taai.blog.ocn.ne.jp/taai/2004/10/post_14.html	2004-10-14 00:00:00
8	146534 http://irisa.jugem.jp/?aid=27	2004-10-14 00:00:00
9	146538 http://brownshoes.jugem.cc/?aid=171	2004-10-13 00:00:00
10	146378 http://kiyomasa-hawks.blogzine.jp/annex/2004/11/11.html	2004-11-04 00:00:00
11	146391 http://blog.livedoor.jp/nakomayu/archives/8029253.html	2004-10-13 00:00:00
12	146397 http://blog.livedoor.jp/s126blue/archives/8109893.html	2004-10-15 00:00:00
13	146576 http://hoku-to.blog.ocn.ne.jp/worker/2004/10/post_55.html	2004-10-13 00:00:00
14	146388 http://zakankirokunin.blog.ocn.ne.jp/turedurezakki/2004/10/post_8.html	2004-10-17 00:00:00
15	146394 http://rockriver.xrea.jp/revirkcor/archives/2004/10/post_27.html	2004-10-14 00:00:00
16	146385 http://nakoya.blogzine.jp/area22/2004/11/post_2.html	2004-11-02 00:00:00
17	146579 http://blog.livedoor.jp/jizake/archives/8068991.html	2004-10-14 00:00:00
18	146386 http://osa3.blog.ocn.ne.jp/project/2004/10/post_19.html	2004-10-20 00:00:00
19	123588 http://osa3.blog.ocn.ne.jp/project/2004/11/post_4.html	2004-11-04 00:00:00

Figure 5: Input Real Time Data Set

Threshold value is taken as initial cluster centroid and the first step is done by comparing each row and so on. Initial cluster centroids are found which are used as centroids for second step. Fig 6 shows the input screen that takes the data and no. of clusters as input.

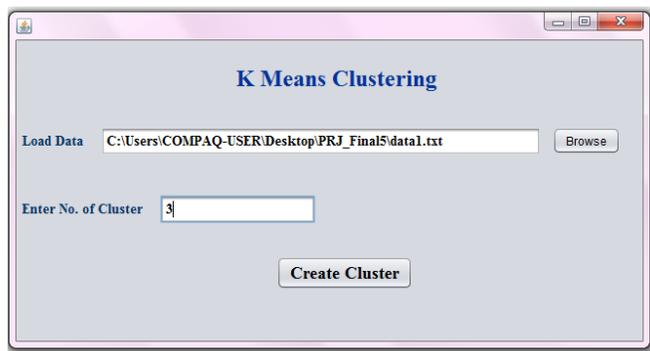


Figure 6: Home Screen of the System

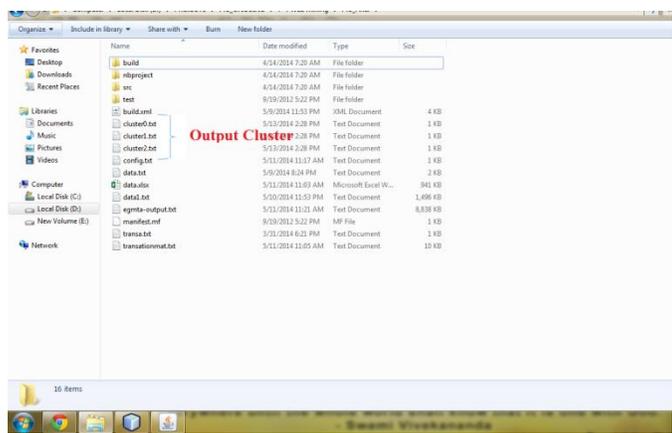


Figure 7: Output screen showing the output cluster

In the second step initial points are optimized by using Kmeans algorithm. Finally well-defined clusters with similar intra objects and dissimilar inter objects are obtained as shown in Fig 7.

VI. CONCLUSION

Extreme utilization of the internet has made automatic knowledge extraction from Web log files a necessity. Information provided are interested in techniques that could learn Web users' information needs and preferences. This can improve the effectiveness of the Web sites by adapting the information structure of the sites to the users' behavior.

The aim in web mining is to find out and retrieve useful and interesting patterns from a large dataset. A way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. Proposed algorithm is used to generate association rules that associate the usage pattern of the clients for a website. In the proposed work we have combined the association mining with the clustering instead of mining association rules from the web log data directly we have mined the clusters. Clustering aims to categorize data dispersed over the Web into groups in order to facilitate data availability and accessing. In future the algorithm can be extended to web content mining, web structure mining, etc. The work can also be extended to extract information from image files.

REFERENCES

- [1] Agrawal, R., Imielinski, T., and Swami, A. N. Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD, International Conference on Management of Data, pp.207- 216, 1993.
- [2] Agrawal. R., and Srikant. R., Fast Algorithms for Mining Association Rules, Proceedings of 20th International Conference of Very Large Data Bases. pp.487-499,1994.
- [3] Kosala and Blockeel, "Web mining research: A survey," SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000
- [4] Sudhamathy G, Jothi Venkateswaran C. "Fuzzy Temporal Clustering Approach for E-Commerce Websites" Sudhamathy G. et al. / International Journal of Engineering and Technology (IJET) Vol 4 No 3 Jun-Jul 2012
- [5] B.Naveena Devia, Y.Rama Devib, B.Padmaja Ranic, R.Rajeshwar Raod, Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce International Conference on Communication Technology and System Design 2011
- [6] B.Santhosh Kumar,K.V.Rukmani, Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms Int. J. of Advanced Networking and Applications Volume:01, Issue:06, Pages: 400-404 (2010)
- [7] Ketul B. Patel,Dr. A.R. Patel, Process of Web Usage Mining to find Interesting Patterns from Web Usage Data International Journal of Computers & Technology www.ijctonline.com ISSN: 2277-3061 Volume 3, No. 1, AUG, 2012
- [8] Rahi, Priyanka. "Business Intelligence: A Rapidly Growing Option through Web Mining." arXiv preprint arXiv: 1208.5875 (2012).
- [9] Pradnya Purandare, WEB MINING: A KEY TO IMPROVE BUSINESS ON WEB IADIS European Conference Data Mining 2008.
- [10] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques

- [11] Patel, Premal. "Implementing APRIORI Algorithm on Web serve log."
- [12] Kosala, R., Blockeel, H. (2000). Web Mining Research: A Survey. ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations. June, (2:1). Pp 1-10.
- [13] P. Bhattacharya and M. L. Gavrilova, "CRYSTAL: A New Density Based Fast and Efficient Clustering Algorithm," Proceeding of the 3rd International Symposium on Voronoi Diagrams in Science and Engineering, pp. 102-111, 2006.
- [14] G. Biswas, J.B. Weinberg, and D. Fisher. Iterate: A conceptual clustering algorithm for data mining. IEEE Transactions on Systems, Man and Cybernetics, 28:100–111, 1998.
- [15] D. Boley, M. Gini, J. Moore. Partitioning-based clustering for web document categorization. Decision Support Systems, 27:329–341, 1999.
- [16] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in PADKK '00: Proceedings of the 4th PacificAsia Conference on Knowledge Discovery and Data Mining,
- [17] Jaideep Srivastava , Prasanna Desikan , Vipin Kumar "Web Mining - Accomplishments & Future Directions".