A Clustering and Associativity Analysis Based Probabilistic Method for Web Page Prediction

Er. Manish Bhanot[1], Er. Jasdeep Singh Mann[2]

P.G. Student, Department of Computer Engineering, BMS Engineering College,
Sri Muktsar Sahib, India [1]
Associate Professor, Department of Computer Engineering, BMS Engineering College,
Sri Muktsar Sahib, India[2]

Abstract:- Today all the information, resources are available online through websites and web page. To access any instant information about any product, institution or organization, users can access the online available web pages. In this work, a three stage model is provided for more intelligent web page prediction. The method used the clustering and associativity analysis with rule formulation to improve the prediction results. The CMeans clustering is applied in this prior stage to identify the sessions with high and low usage of web pages. Once the clustering is done, the rule is defined to identify the sessions with page occurrence more than average. In the final stage, the neuro-fuzzy is applied to perform the web page prediction. The result shows that the model has provided the effective derivation on web page visits.

Keywords: *HMM, Cmeans, Kmeans.*

____*****____

1. INTRODUCTION

WWW web environment to all users globally and publically. It provides the easy availability to all users over the world. The web is the respiratory for different text documents, image media and the video information. Each host over the web can access this information publicly and provide the resource access to the system with short and global identifiers. To access this information, the web server is identified a specific address or the domain name called Uniform Resource Identifier. This information access to the system is defined in a simple and effective way. The URL is actually the pool of HTML documents that connected through the Hyperlinks. These interlinked Web Services actually represent a complete web site domain.

Web Mining: Web Miningis the application of data mining techniques used to extract useful patterns from the web. Web Miningcan be divided into three different types.

1. **Web Content Mining:** describes the automatic explore of information resources available online, and involves mining of web data content.
2. **Web Structure Mining:** is the process of analyzing hyperlink and tree-like structure of a web site using graph theory.
3. **Web Usage Mining:** is the process of extracting effective information from web server logs.

Working of Web Usage Mining: Web usage mininguses data mining techniques to discover useful access patterns from web server logs. Web log data is a record of all URLs accessed by users on a Web site.

Log entry consists of –: A.) Access Time B.) IP Address C.) URL Viewed D.) Referrer

Purpose & Goal of Web Usage Mining

Web Usage Mining: Automatic discovery of patterns in click streams and associated data collected or generated as a result of user interactions with one or more Web sites. The Goal of Web usage mining is to analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests.

Data Input Process in Web Usage Mining:

a.) Web Server Logs b.) Site Contents

c.) Data about the visitors, gathered from external channels d.) Further application data.After that various data mining algorithms can be applied.

Web Usage Mining Process:

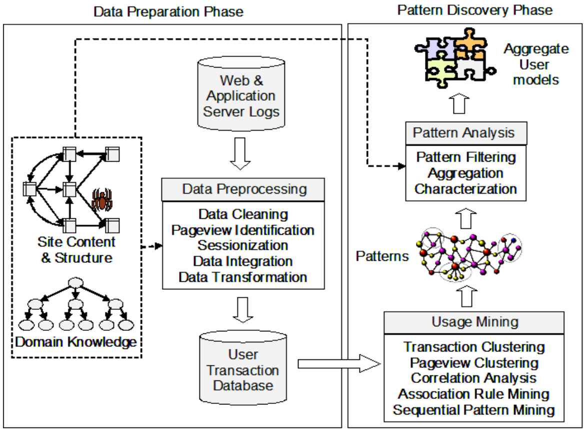

Figure1: Web Usage Mining

Web Crawler: Global web architecture is actually the client server based system.One of the important mediators between the web server and the client is defined by the Web Service Identification. The Web Service Identification basically accepts the user query and performs the search over all servers available over the web.

Web crawlers basically work on the theory of graph structure to move between the fetched Services. As the search request is performed, the relative web pages are retrieved on which the user request contents are present. The process of searching the information from these servers is defined by Web Crawler. The web crawler is also called a spider, worm, walker etc. It is basically about to perform an effective web search. A Crawler actually performs the identification of some text or the information over the web. To perform this search it uses the URL database to access different URLs one by one. One a URL is taken, the search is performed on that particular URL and the information retrieval is performed.

2. RESEARCH METHODOLOGY

To present the effective globalize work, a dynamic model is presented on dummy dataset. The page number is assigned to process effectively.

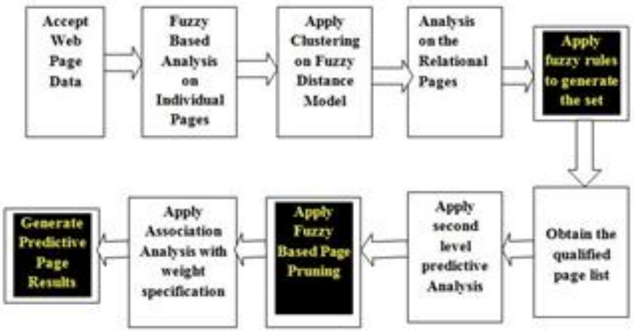
Accept Web Page Data
Fuzzy Based Analysis on Individual Pages
Apply Clustering on Fuzzy Distance Model
Analysis on the Relational Pages
Apply fuzzy rules to generate the set
Obtain the qualified page list
Apply second level predictive Analysis
Apply Fuzzy Based Page Pruning
Apply Association Analysis with weight specification
Generate Predictive Page Results

Figure:2:Proposed Model

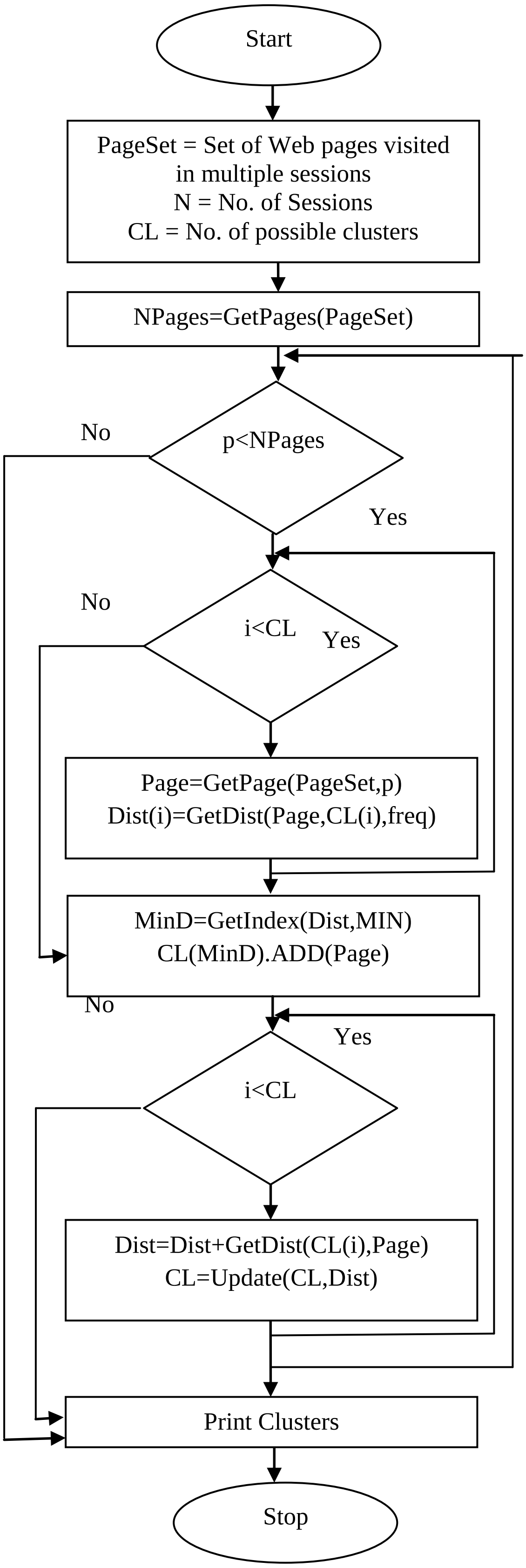

Figure 3: Flow Chart

ISSN: 2321-8169
213 - 217

In this work, a fuzzy clustering integrated improved model is defined with individual and the paired page analysis. The work is about to improve the clustering process for the web page prediction.

3. RESULTS

In this section the different problems are faced and all these problems are resolved with different objectives. The MATLAB Environment is used to implement work .The compiled result snap shots are given below:

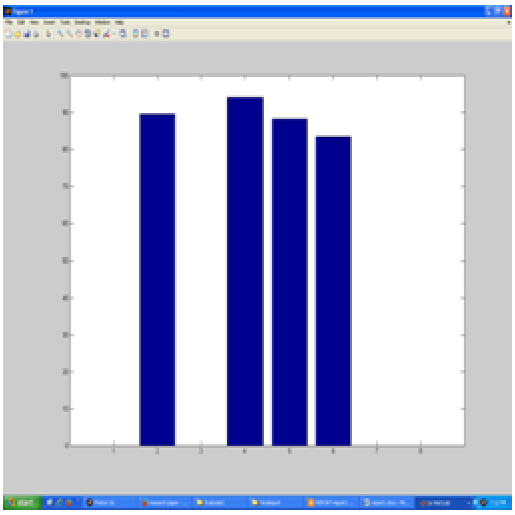

Figure3:Web Page Visit Probability (Page I)

Here figure 3 is showing the web page visit frequency among number of defined pages. Here x axis is showing the page index and y axis is showing the web page visit. The figure is showing the web page visits so that the ratio based visited page visits can be obtained.

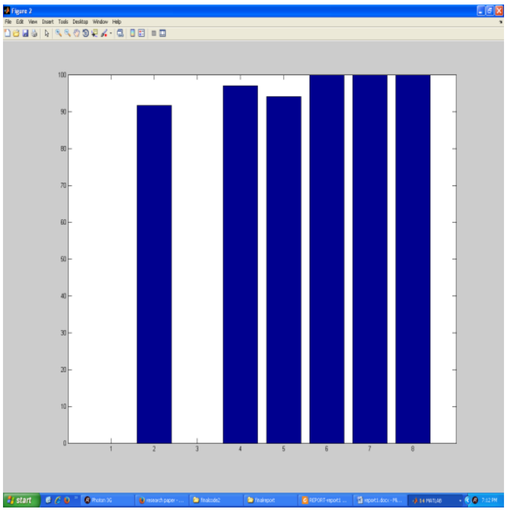

Figure 4: Web Page Visit Probability (Associated Page)

Here figure 4 is showing the frequency measure obtained for different web pages for associated page 2. The Fuzzy Clustering Improved model is here applied to analyze the first page visit identification. Here x axis is showing the page number or the page itself and y axis is showing the possibility of page visit. The ratio based analysis is here performed to identify the page visit. The ratio is obtained respective to maximum visited page. As the first page is identified, the possibility of relative page visit is required to identify. The figure is showing the association adaptive analysis.

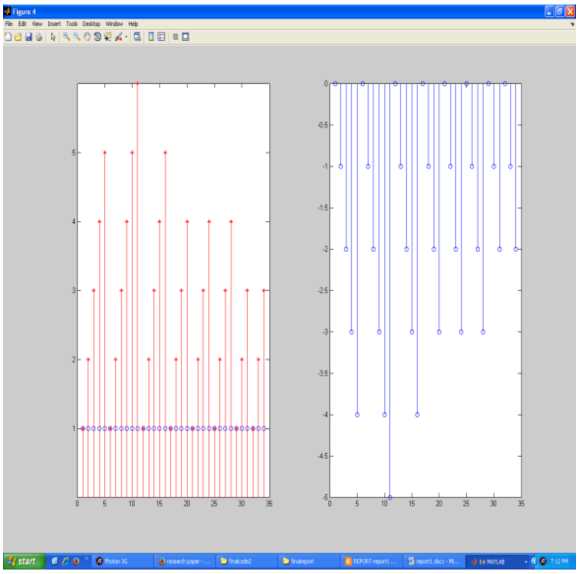

Figure 5: Neural Predictive Page Results

Here figure 5 is showing the neural predictive results respective to different pages. Here figure is showing the page 10 is having the maximum frequency so that the chances of this page visit are high. The blue dots in first section are showing the lower limit of the page. After that the stage adaptive page visit is defined. Higher the line more frequent a page can be visited. The right side is the negative form of same model as the positive and negative weight ages are processed.

Comparative Analysis: TestSet I

Table 1: TestSet Properties

Properties	Values
Size of Training Data	238
Size of Testset	40
Number of Features Collected	20

Number of Converted Text Features	14
Number of Numerical Features	6
Number of Class	5

Here table 1 is showing the number of training and testing feature based on which the classification process is implied. The parameters of neurofuzzy processing applied are shown in table 2. These parameters include the static and dynamic parameters.

Table 2: Neuro-Fuzzy Parameters

Parameters	Values
Number of Iterations	100
Number of Linear Parameters	252
Number of Nodes	527
Number of Nonlinear Parameters	480
Number of Fuzzy Rules	12

After applying these neuro-fuzzy parameters, the classification process is applied on training and testing data. The accuracy obtained from the system is 61.53. Some variations on number of iterations is applied to generate more results. These results are described in table 3

Table 3: Iteration Based Results

Number of Neuro-Fuzzy Iterations	Accuracy %
100	61.53
200	66.6667
500	69.2308
1000	72.69

Here table 3 has showed that the accuracy of the work is improved as the numbers of iterations are increased. The accuracy driven results are here shown in figure 6.

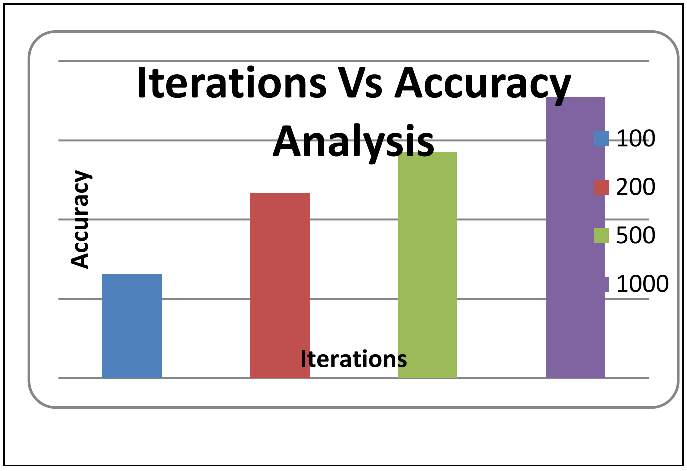

Figure 6: Accuracy Analysis

Here figure 6 is showing the accuracy driven results for the presented work. The result shows that the model has provided effective results as the number of iterations in the system increases. The method provided the accuracy upto 70% for this sample set.

3.1Comparative Evaluation

The hierarchical clustering technique with Levenshtein distance evaluation method. Author used the page rank with access time, frequency and markov model prediction for effective web page prediction. The usage based page ranking algorithm was applied for effective web page prediction. The agglomerative hierarchical clustering method was applied on different session and similarity distance analysis was defined. The comparative evaluation of existing and proposed method under accuracy analysis is shown in figure 5.13. The evaluation is also taken against Frequency based page rank (TFPR), Usage based ranking(UBR) and Kmeans clustering Euclidian distance and Markov Model (KEDM) methods.

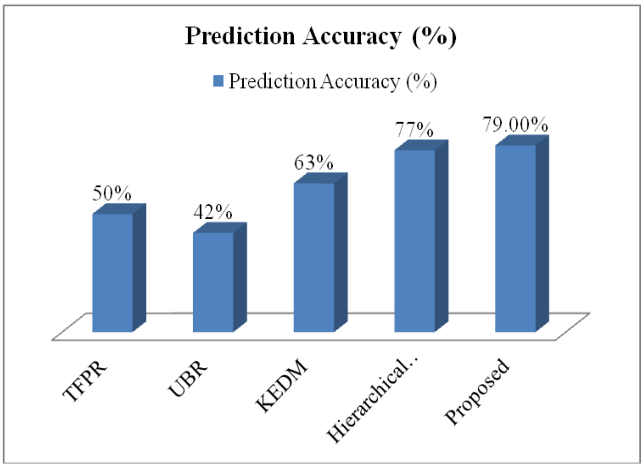

Figure 7: Comparative Analysis

Here figure 7 is showing the comparative results of proposed method against four existing methods. The accuracy ratio observation shows that the proposed method has improved the prediction results.

4. CONCLUSIONS

Web content mining is one of the effective applications required to improve the web page access and the web processing speed. In this work, an improved fuzzy clustering based featured model is defined for web page prediction. The work model is divided in three main stages. In first stage, the individual page analysis is done and the group is done under frequency distance parameter. Here fuzzy clustering is used for identifying the visiting page list The result shows that the model has provided the effective derivation on web page visits.

5. REFRENCES

[1] A. C. E. S. Lima and L. N. de Castro, "Automatic sentiment analysis of Twitter messages," Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, Sao Carlos, 2012, pp. 52-57.

[2] Ah Chung Tsoi," Adaptive Ranking of Web Pages", WWW2003, May 20–24, 2003, Budapest, Hungary.ACM 1-58113-680-3/03/0005.

[3] Alon Altman," Ranking Systems: The PageRank Axioms", EC'05, June 5–8, 2005, Vancouver, British Columbia, Canada. ACM 1-59593-049-3/05/0006

[4] B. Y. Ong, S. W. Goh and C. Xu, "Sparsity adjusted information gain for feature selection in sentiment analysis," Big Data (Big Data), 2015 IEEE International Conference on, Santa Clara, CA, 2015, pp. 2122-2128.

[5] Bin Gao," Semi-Supervised Ranking on Very Large Graphs with Rich Metadata", KDD'11, August 21–24, 2011, San Diego, California, USA. ACM 978-1-4503-0813-7/11/08

[6] C. Pino, I. Kavasidis and C. Spampinato, "Assessment and visualization of geographically distributed event-related sentiments by mining social networks and news," 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2016, pp. 354-358.

[7] Chen Wang." Extracting Search-Focused Key N-Grams for Relevance Ranking in WebSearch",WSDM'12, February 8–12, 2012, Seattle, Washington, USA. ACM 978-1-4503-0747-5/12/02

[8] Dungeon Choi," An Approach to Use Query-related Web Context on Document Ranking", ICUIMC '11, February 21–23, 2011, Seoul, Korea. ACM 978-1-4503-0571-6

[9] ElisabettaFersini," Granular Modeling of Web Documents: Impact on Information Retrieval Systems", WIDM'08, October 30, 2008, Napa Valley, California, USA ACM 978-1-60558-260-3/08/10

[10] Franco Scarselli," Adaptive Page Ranking with Neural Networks",WWW 2005, May 10–14, 2005, Chiba, Japan. ACM 1-59593-051-5/05/0005

[11] Guangyu Zhu," Mining Rich Session Context to Improve Web Search", KDD'09, June 28–July 1, 2009, Paris, France. ACM 978-1-60558-495-9/09/06

[12] Haixuan Yang," Diffusion Rank: A Possible Penicillin for Web Spamming", SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007.

[13] Haixuan Yang," Predictive Ranking: A Novel Page Ranking Approach by Estimating the Web Structure", WWW 2005, May 10–14, 2005, Chiba, Japan. ACM 1-59593-051-5/05/0005.

[14] Harish Kumar, Vibha L, Venugopal K R, "Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model", IEEE Region 10 Symposium (TENSYMP), pp 1-6, 2016.

[15] Honghan Wu," Ranking Domain Objects by Wisdom of Web Pages", WIMS'12, June 13-15, 2012 Craiova, Romania ACM 978-1-4503-0915-8/12/06