SURVEY OF WEB PAGE ACCESS PREDICTION TECHNIQUES

[1]Pooja M. Bharti, [2]Prof. Tushar J. Raval
[1]Student, [2]Associate Professor
[1]Department of Computer Engineering,
[1]L. D. College of Engineering, Ahmedabad, India

Abstract : Web is a huge collection of resources. Web users often face the problem of information overload. So there is a need to recognize the web users' behaviors on web sites using web mining for improving the user's experiences on these sites. Prediction of the web page that may be visited by the web user is important, since this knowledge can be used for pre-fetching of web pages, recommendation or personalization of the web page for that user. The objective of the work is to study various method used in past, at present and which can be used in future to minimize the search time of a user on the network.

IndexTerms **- Web Usage Mining, Web Content Mining, Web Page Access Prediction.**

I. INTRODUCTION

Due to the advancement in technology there are large amount of unprocessed information. It is time consuming to view or extract the needed information. In such a situation we are in need to develop a strategy which is useful to obtain the necessary information. Since there are large amount of data, decision making process is tedious. To overcome these pitfalls the concept of Data Mining is used. [13]

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and Intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular web localities.

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. Web mining research can be classified into three categories: Web content mining (WCM), Web structure mining (WSM), and Web usage mining (WUM). Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, etc. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main applications of web usage mining is to learn user profiles. The Taxonomy of web mining is as in the figure.

Fig. 1.1 : Categorization of Web Mining

II. WEB PAGE ACCESS PREDICTION

Internet nowadays is the most democratic of all the mass media. Millions of users access different Websites all around the world. When they access the network, a large amount of data is generated and is stored in Web log files which can be used efficiently as many times user repeatedly searched the same type of Web pages recorded in the log files. These series can be considered as a web access pattern, helpful to find the user behavior. Through this personalized information, it's quite easy to predict the next set of pages user might visit based on the previously searched patterns, thereby reducing the browsing time of a user. [13]

Web is a huge collection of resources. Web users often face the problem of information overload. There is need to recognize the web users' behaviors on web sites using web mining for improving the user's experiences on these sites [6]. Prediction of web page that may be visited by the web user is important, since this knowledge can be used for pre-fetching of web pages, recommendation or personalization of the web page for that user [6]. Next potential web page that might be visited by the web user can be predicted by web mining. Firstly the web log data is collected and pre-processed. Then the model is constructed by applying data mining techniques. When the web user's request comes, the prediction module predicts the web page that may be visited by that web user.

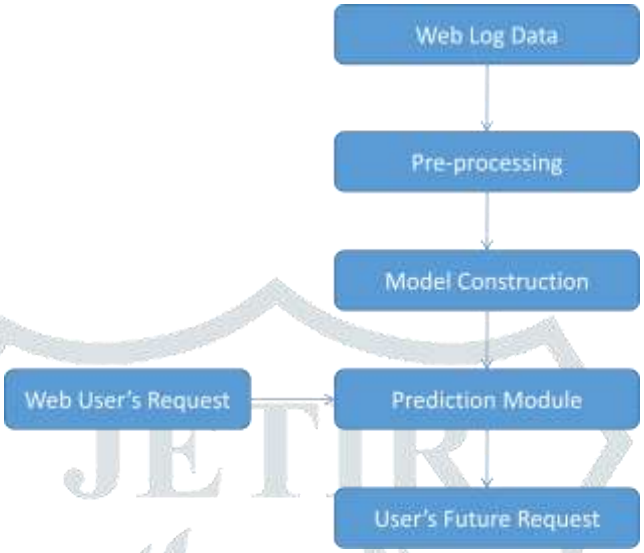

Fig. 2.2 : Flow chart of web page access prediction

2.1 Application of web page access prediction

- To improve user navigation through pre-fetching and caching
- To improve web design or in e-commerce sites
- Dynamic web page recommendation
- Personalization of web page for specific user or group of users
- Dynamic hyperlink generation

2.2 Advantages of web page access prediction

- The web page access time is reduced.
- Reduces network latency by pre-fetching the predicted pages.
- Web user's search time is reduced.

III. TECHNIQUES OF WEB PAGE ACCESS PREDICTION

The focus of this section of the paper is to study and contrast different available techniques to predict user's next web page access.

3.1 A New Web Usage Mining Approach for Next Page Access Prediction

This approach use sequential access data for improving recommendation accuracy. PNN clustering is applied on cleaned web log data. The pair wise nearest neighbor approach is a bottom-up hierarchical clustering technique, by which every object belongs to individual clusters initially, pair wise merging of objects is done at every step based on their similarity. Finally, resulting in a single cluster. The distance calculations are replaced by similarity measure. Similarities between two transactions are given by the ratio of, total number of unique pages referenced by them to the number of common references. Then a table representing similarity value between every pair of transaction is created. For every transaction, their first k nearest access sequences are identified. Among the whole set of k neighbors, the pair of sequences having high similarity is identified and merged. The merging is continued until no more merging is possible. In this process, only the pair of access sequence having similarity value greater than a pre-defined threshold is selected for merging process. By this approach, the distant objects that are irrelevant to mining process are eliminated resulting in homogeneous access patterns. [1]

After clustering, sequential pattern mining is performed using dynamic support pruned k-th order Markov model. For new test session the cluster having similar access sequences is selected. Start with highest possible value of k and Apply Markov model and find out the kth order states for the test session from its cluster. If the support is very less, calculate next lower order states for the test session from its cluster. Then repeat the steps until states are generated with enough support. Display the page with highest probability as the recommended page.

Advantages: This approach use sequential access data for prediction and has good prediction accuracy than traditional Markov Model with less state space complexity.

Limitation: By this approach, the distant objects are eliminated resulting in homogeneous access patterns. It does not consider loosely connected (non-continuous) access sequences so if these access sequences are important for prediction then they are ignored by this approach.

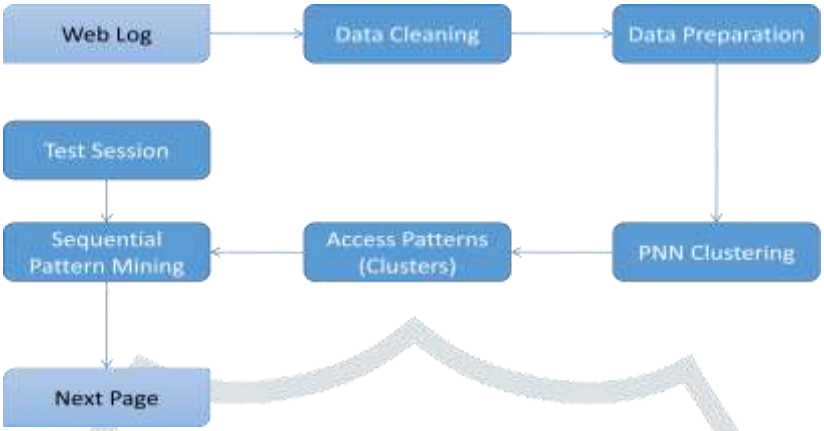

Fig. 3.1 : Steps in New web usage mining approach for next page access prediction

3.2 Clustering Frequent Navigation Patterns from Website Logs by Using Ontology and Temporal Information

It combines concept-based sequence clustering with time-spent information to predict the web page that may be visited by the user. Navigation information of users is integrated with the set of concepts defining web pages. Each session is a sequence of web pages and each web page is represented with a set of concepts from the taxonomy defined. Sessions are clustered using direct k-way clustering to find meaningful partition with the aim of maximizing intra-cluster similarity while minimizing inter-cluster similarity. Two different kinds of similarity measures are used: Rada distance to measure the similarity between two concepts, and average set similarity to measure the similarity between two sets of concepts. [2]

To improve the similarity score between two web pages, the importance component is introduced and used together with the similarity component [2]. The importance component computes how important the similarity between two web pages and how much it should contribute to the overall similarity between two sessions containing these two web pages. It uses the fraction of time spent at these pages [2]. Then, Needleman–Wunsch dynamic programming algorithm is used to measure similarity between two sessions, which is used for clustering of sessions. Computed clusters represent groups of sessions of users with similar contents. These clusters can be used to understand the behavior of users. Then k-nearest Neighbor classification algorithm is used for recommendation of web page to the user.

Advantages: Similarity measure used in this method has more accuracy than URL-equality measure. Intent of the user can be determined by this method because it makes use of both web usage mining and content of the web pages as the set of concepts.

Limitation: Each web page is represented with a set of concepts from the taxonomy defined. This taxonomy is manually defined using the keywords. It is not always possible to build taxonomy manually because you need to study whole web site to extract the keywords and then make taxonomy using the relationship between the keywords.

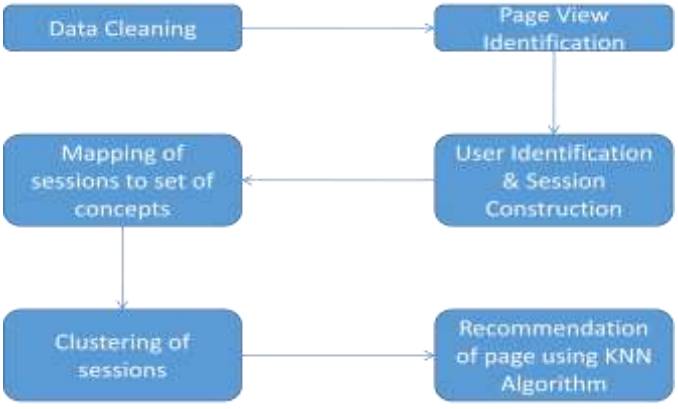

Fig. 3.2 : Recommendation system using Ontology and Temporal information

3.3 Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method

The K-Nearest Neighbor model is the simplest and most straightforward for class prediction, it is the most popular similarity or distance based text and web usage classification and recommendation model. K-Nearest Neighbor could be described as learning by analogy, it is learnt by comparing a specific test tuple with a set of training tuples that are similar to it. It is classified based on the class of their closest neighbors, most often, more than one neighbor is taken into consideration hence, the name K-Nearest Neighbor (K-NN), the ‘‘K’’ indicates the number of neighbors taken into account in determining the class. [3]

The K-NN is often referred to as ‘‘Lazy learner’’ in the sense that it simply stores the given training tuples and waits until it is given a test tuple, then performs generalization so as to classify the tuple based on similarities or distance to the stored training tuples. In this work, the number of class C of user X that can be recommended by the recommendation model is set at 5, ‘‘5’’ indicates different news categories headlines and user classes that could be presented to the active user, based on information from the user’s click stream. However, this number could be increased or decreased depending on the available options at a given time. Euclidean distance is used as the similarity measure. [3]

Advantages: Overcome the scalability problems and provide precise recommendation to user based on his current navigation pattern.

Limitation: It provides classifications and recommendations to the user at any time based on his immediate requirement rather than information based on his previous visit to a site. So this method ignores the previous visits of the user for prediction.

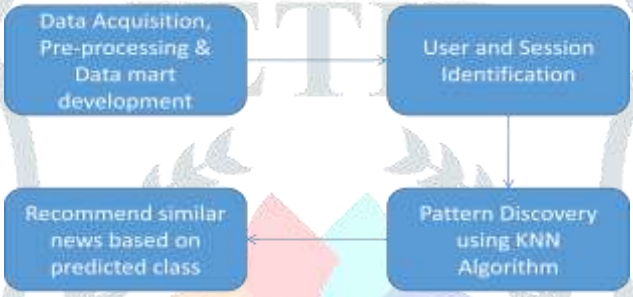

Fig. 3.3 : Recommendation system using KNN classification

3.4 Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model

The methodology uses the Agglomerative Hierarchical Clustering. To cluster the sessions, measuring the similarity between the sessions is necessary. This is achieved by using an appropriate distance metric. Modified Levenshtein distance is used as the distance metric which is an improved form of Edit distance (Levenshtein Distance). Levenshtein distance has a severe drawback when applied to the web sessions. It does not consider the page sequence into account. Page sequence is very essential in clustering and prediction. Modified Levenshtein distance technique can be used for checking the string similarity and it also takes into account the page visit sequence. Set of sessions obtained from the user and session identification algorithm are given as input to the Hierarchical clustering algorithm to get *k* clusters. Modified Levenshtein distance is used to find the similarity between the sessions and sessions with higher similarity are merged hierarchically. Cluster representative is the input to the Markov Model algorithm and gives the Transition Probability Matrix which gives the probability of moving to the next page from the current and previous pages. [4]

Advantages: It eliminates the difficulties in deciding the number of clusters and choosing initial random centers. It also considers the order of page visits for prediction.

Limitation: It makes use of Higher order Markov Model so accuracy is good but state space complexity is high.

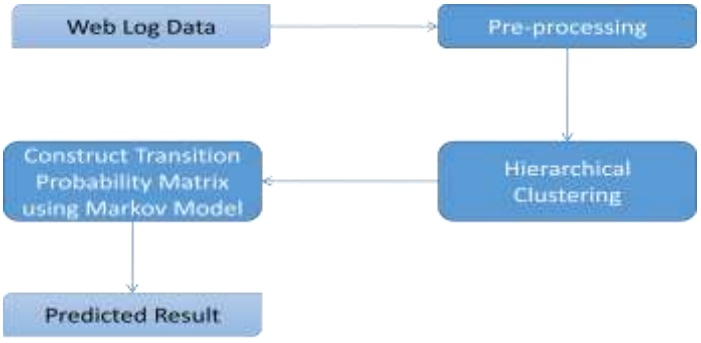

Fig. 3.4 : Web page access prediction using Hierarchical Clustering and Higher order Markov model

3.5 An effective web page recommender system with fuzzy c-mean clustering

A web user may have multiple interests for which he needs to be put into multiple clusters. In fuzzy clustering or soft clustering, data elements can go to more than one cluster, and linked to each item is a set of membership levels. These specify the strength of the relationship between that data element and a particular cluster. Fuzzy clustering is a procedure of assigning these membership levels, and then using them to assign data elements to one or more clusters. [5]

After applying fuzzy c-mean clustering, we get cluster center values for each cluster upon which processing is done instead of entire data of users present in the cluster. We are required to find out the cluster number in which each user occurs. Since we are dealing with soft clusters, we may get a number of cluster numbers for each user. After choosing target user we now need to find all users similar to target user as they would have a direct impact on user's recommendations. The similar users are found for the target user using Fuzzy c-means clustering algorithm which generates soft clusters. A user can be present in more than one cluster. After that, Top N-clusters are produced for the user. After that the weight for each page category is determined. Higher the weight, higher is the probability of the page to be recommended to the user. [5]

Advantages: It uses sequential information together with content for recommendation and fuzzy c-mean algorithm is used so that multiple interests of the users can be determined. For new websites the system is of great significance because website owners can track what the user is going to purchase and can provide related recommendations to the user without their explicitly searching over their website.

Limitation: Lacks in privacy and trust because website owners can track that what user is going to purchase.

Fig. 3.5 : Web page recommender system using fuzzy c-mean clustering

3.6 Improving the prediction of page access by using semantically enhanced clustering

This method considers Semantics as a set of concepts corresponding to the web page and use concept-based semantic similarity method for clustering. The web log data is collected and pre-processed. After that annotation of web pages with concepts corresponding to that page's semantics is done and sessions are generated from web logs. URL's of visited pages in the session are replaced with the concepts associated with that page. User session generated at the end of this phase is a sequence of sets of concepts associated with web pages. Semantic session similarity calculation is done next. Rada distance is used for concept similarity and Modified Jaccard Set Similarity is used to measure the similarity between two concept sets and similarity between two sessions is determined by Needleman-Wunsch dynamic programming algorithm. [6]

After similarity calculation Direct k-way clustering is applied to form the clusters using content-based semantic similarity. The centroids of clusters as potential representation of common user navigation behaviors are calculated. For prediction the current user session is compared against the centroids of the clusters formed in the model construction process. The most similar centroid is chosen. The current user session is compared against all of the sessions of the selected cluster. Choose the most similar session from that cluster. The next web page is predicted using the matching session's unmatched suffix.

Advantages: Clustering reduces the amount of search time while drops the accuracy slightly. But the content-based semantic similarity method compensates the difference and improves the accuracy.

Limitation: All the sessions are kept for prediction but it is not possible to keep all the sessions. This method manually associates set of keywords for each concept in the ontology.

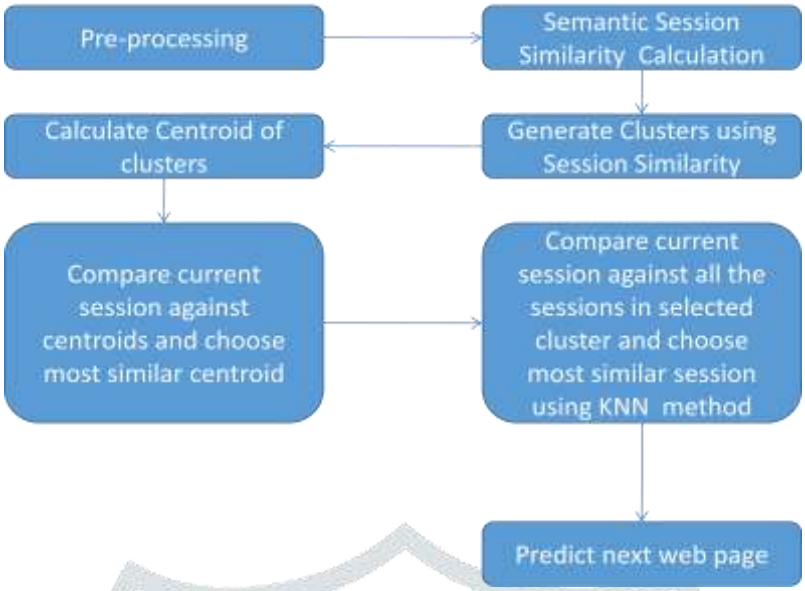

Fig. 3.6 : Page access by using semantically enhanced clustering

IV. COMPARATIVE ANALYSIS

Table 4.1: Comparative Analysis of different techniques of web page access prediction

Sr. No.	Paper Title	Methodology	Objective	Similarity Measure	Limitation
1	A New Web Usage Mining Approach for Next Page Access Prediction	Hierarchical Clustering (PNN) based on variant of Markov Model	Use sequential access data for improving recommendation accuracy	Ratio of no. of unique references and no. of common references	Non-continuous access sequences are ignored
2	Clustering Frequent Navigation Patterns from Website Logs by Using Ontology and Temporal Information	Direct k-way Clustering and K-Nearest Neighbor Classification	Combining concept-based sequence clustering with time-spent information	Importance Component, Similarity Component	Taxonomy is manually defined only considering the keywords of the web pages
3	Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method	k-Nearest Neighbor Classification Method	Overcome the scalability problems and recommendation based on current navigation pattern	Euclidean Distance	Ignores the previous access sequences for prediction
4	Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model	Hierarchical Clustering and higher order Markov Model	Eliminate the difficulties in deciding the number of clusters and choosing initial random centers. Consider the order of page visits.	Modified Levenshtein Distance	State space complexity is high
5	An effective web page recommender system with fuzzy c-mean clustering	Fuzzy c-mean Clustering	Use sequential information together with content for recommendation	Membership levels	Lacks in privacy and trust

6	Improving the prediction of page access by using semantically enhanced clustering	Direct k-way Clustering using content-based semantic similarity	Consider Semantics as a set of concepts corresponding to the web page for capturing intention of users	Modified Jaccard Set Similarity	All the sessions are kept for prediction and manually associates set of keywords for the concepts

In [1] sequential data and the Markov Model are used for prediction. It uses variant of Markov Model which has less state space complexity than traditional Markov Model without compromising accuracy. It ignores distant access sequences for prediction. The similarity measure used in [2] has more accuracy than URL-equality measure and Euclidean distance because it can capture the intent of the user. Hakki Toroslu in [6] extended the work done in [2] and improved the prediction accuracy using the content of the web pages as the set of concepts for prediction. The methodology in [3] overcomes the scalability problems common to many data mining techniques. It can provide best recommendation to users based on his immediate requirement but does not consider previous access sequences. In clustering techniques like k-way clustering or c-mean clustering there is need to choose initial random centers and decide the number of clusters. In [4] these difficulties are eliminated but the state space complexity is high due to the use of higher order Markov Model. Fuzzy c-mean clustering considers multiple interests of the users by putting the same objects into multiple clusters unlike all the other methodologies which puts the object into only one cluster but it lacks in privacy and trust.

V. CONCLUSION

The information generated in log files can be extensively used through Web Usage Mining, to efficiently increase the prediction time without compromising accuracy. By only considering the URLs of web pages, it is not possible to predict the accurate web pages so content of the web pages is equally important for prediction. Markov Model achieves good prediction accuracy but the state space complexity is high. By only considering the URLs of the web pages, it is not possible to capture the intent of the web user so content of the web pages is also needed to predict the next web page that may be visited by the web user. So prediction of page access by using semantically enhanced clustering can be used in future to improve the accuracy of prediction and reduce the search time of the web user because it doesn't require full content of the web pages but the web pages can be represented by the set of concepts which is enough to capture the intent of the web user.

REFERENCES

[1] A. Anitha 2010. A New Web Usage Mining Approach for Next Page Access Prediction. International Journal of Computer Applications, IEEE.
[2] Sefa Kilic, Pinar Senkul, Ismail Hakki Toroslu 2013. Clustering Frequent Navigation Patterns from Website Logs by Using Ontology and Temporal Information. Computer and Information Sciences III, Springer.
[3] D.A. Adeniyi, Z. Wei, Y. Yongquan. 2015. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, Elsevier.
[4] Harish Kumar B T, Dr. Vibha L, Dr. Venugopal K R. 2016. Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model. Region 10 Symposium, IEEE.
[5] Rahul Katarya, Om Prakash Verma. 2016. An effective web page recommender system with fuzzy c-mean clustering. Science+Business Media New York, Springer.
[6] Erman Sen, Hakki Toroslu, Pinar Karagoz. 2016. Improving the prediction of page access by using semantically enhanced clustering. Science+Business Media New York, Springer.
[7] Monika Dhandi, Rajesh Kumar Chakrawarti. 2016. A Comprehensive Study of Web Usage Mining. Symposium on Colossal Data Analysis and Networking (CDAN), IEEE.
[8] Federico Michele Facca, Pier Luca Lanzi. 2004. Mining Interesting Knowledge from Weblogs: A Survey. Data & Knowledge Engineering 53 (2005) 225–241, Elsevier.
[9] Dr.Sanjay Kumar Dwivedi, Bhupesh Rawat. 2015. A Review Paper on Data Preprocessing: A critical phase in Web Usage Mining, IEEE.
[10] Sunena, Kamaljit Kaur. 2016. Web Usage Mining – Current Trends and Future Challenges. International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE.
[11] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. 2000. Web Usage Mining: Discovery and Applications of usage patterns from web data. SIGKDD Explorations.
[12] Jiawei Han, Micheline Kamber (Second Edition). Data Mining: Concepts and Techniques. Elsevier.
[13] Avneet Saluja, Dr. Bhupesh Gour, Lokesh Singh. 2015. Web Usage Mining Approaches for User's Request Prediction: A Survey, International Journal of Computer Science and Information Technologies.