International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979
Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 4; Issue 12; December 2017; Page No. 74-79

Cluster analysis to predict web page using k-means and affixed agglomerative approach (CAPKAAA)

Jothish Chembath
Ph.D. Research Scholar, Department of computer science, Karpagam University, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India

Abstract

Web mining is about combining information collected from the World Wide Web using data mining methods and technologies. Predicting the subsequent web page that may be visited by a particular user has become the most wanted area of research as the need for maximum accuracy is mandatory in the sustenance of business in the World Wide Web. Several models are in use nowadays for prediction which focus on the detecting the users subsequent visit of a web page. Basically all prediction mechanisms concentrate on the basic web usage mining principles of clustering. Obviously we need the services of a prediction model like Markov model which has to be trained with the clusters created by cluster algorithms. Server logs help us to understand the user behavior and the possibility of their next web page visit. In this paper, we prove from the experiments conducted that good formation of clusters will lead to better predictions. Here we present an integrated cluster approach from data sets for the prediction of user requests. This clustering approach defines clusters which can be used for predicting the next user request. We focus on predicting the next request of web users by utilizing Basic Agglomerative Hierarchical Clustering Techniqueor "Bottom-Up" Algorithm and K-Means algorithm along with the prediction algorithm of Markov model. Experimental results reveal that Markov-based models when combined with K-means and Agglomerative approach for clustering produce more accurate results.

Keywords: agglomerative clustering, markov model, next page prediction**,** web log data, web usage mining, K-Means

1. Introduction

Internet evolution has made the web developers and internet class as a whole to rethink their strategies in forcing or understanding the necessities and requirements of a customer/user. Many an attempt by researchers have provided tools which help in improving the internet surfing of users. The focal point here is concentrated on web mining. It locates this knowledge required from the web servers basically from web logs. Although several mechanisms are there to predict the move what a user is going to make, the intention of this paper is to use the best clustering method K-means [2] with Agglomerative Cluster Technique along with Markov model, so as to arrive at the best prediction possible. When dealing with clusters of web pages, the similarity of clusters have to be taken into an account. Clustering deals with pages, consisting of pages which are similar. During Clustering, we have given due consideration to the inter-class and intra-class cluster similarity.

Here in this paper, as researchers we have first clustered the web log files using the clustering methods discussed above for analysis and subsequently continued it with the prediction process using Markov model. Markov algorithm reduces the prediction time. The steps that are involved in CAPKAAA system is Preprocessing, Clustering and Pattern analysis. This paper analyses two cluster algorithms that can predict future next web requests of the user, one simple K-means and the other a combined effort of using both K-means and Agglomerative algorithm for enhancing efficiency and quality of clusters used for predicting mechanisms. Then these two algorithms selected is combined with Markov Model. This is referred to as CAPKAAA model in this paper. The rest of the paper is organized as the section 2 presents review of literature and the related work, the section 3 presents the existing work, the section 4 presents the detailed description of the model selected, the section 5 discusses the experimental results obtained during the performance evaluation, the section 6 presents the findings from the experiments, the section 7 presents the applications/improvements from this research paper, and the section 8 concludes the work with future research directions.

2. Review of Literature

Clustering of web pages is a field which can be explored more as far as research and implementation is concerned. Clustering is an unsupervised learning problem, and its aim is to find grouping in a collection of disintegrated data [3]. A cluster is thichhe group of similar objects that are 'dissimilar' and belonging to other clusters [4]. Conceptual clustering is a type of technique where items are believed to be part of same cluster, if it defines concepts which are common to the items. Grouping of objects is completed by its nature and not according to simple Common Subsequence algorithm [5, 6]. K-means clustering and Regression Analysis algorithms which are used to predict the future request by Vedpriya Dongre [1] and Jagdish raekwal [1] and have proved that this combination produced efficient results. There are limitations associated with K-means [2] which reiterates that K-means is sensitive to cluster center initialization which needs to be improved. The current clustering techniques [3] faced problems in solving the problem of measuring the distances of clusters created. There are lot of benefits for Agglomerative clustering [4] when it can reduce the number of clusters created from a large number of datasets. Inter cluster relationship is not concentrated in the K-means algorithm [6] A new system was proposed that improves K-means algorithm by an improved centroid estimation for the clusters created. This paper is an attempt to improve the prediction accuracy of K-Means clustering algorithm when combined with Hierarchical Clustering Technique (Agglomerative Clustering), thereby taking both inter cluster and intra clusters for clustering. Prediction of web pages is carried out through Markov model which implements the clusters so produced, to give precise prediction of the future visits by a user.

3. Existing Work

The previous work comparing fuzzy c means [12] and k means indicates that the former produces better performance. It has been indicated that bisecting K-means is as good as the hierarchical approach [11], which clearly indicates when K-means can be combined with the best of the clustering methods it can give optimal result. Hierarchical algorithm provides good quality of results corresponding to k-means [10] when used with query redirection method. When K-means is found to be better for large datasets hierarchical algorithms works better for small datasets. Presently, the aim of the researchers here is to use the better qualities of both K-means and Agglomerative clustering to bestow the best performance and quality, as K-means gives performance and hierarchical clustering gives quality.

4. Cluster Analysis to Predict Web Page using K-means and Affixed Agglomerative Approach

The present paper represents Agglomerative clustering that improved the performance and precision of K-means clustering algorithm. We have attempted to combine K-means clustering algorithm with its threshold value and also applied Agglomerative clustering on this K-means algorithm and then compared it individually performance wise, precision wise and f1-measure wise. We have demonstrated that the K-means algorithm when combined with Agglomerative clustering produced good results when compared to the existing k-means algorithm, through experimental results. We could conclude that our work on these methods, the first one using K-means and the other one on CAPKAAA (Clustered Analysis to Predict web page using K-means and Agglomerative Approach) produced results that favored combined clustering (CAPKAAA) on the basis of its performance in terms of accuracy, precision and F1 measure parameters.

4.1 CAPKAAA

The CAPKAAA is designed as a hybrid model that combines three algorithms, namely, K-means clustering, Agglomerative clustering and Markov model. The impetus of using this model is to increase the performance and quality of the prediction model. The number of transactions is reduced through the use of the K-means and Agglomerative clustering algorithm. Then Markov model prediction is performed on each cluster. CAPKAAA model uses K-means and Agglomerative clustering along with prediction model MARKOV's algorithm. We have also used another model for comparing CAPKAAA that uses simple K-means and Markov prediction algorithm. While prediction, both the models begin by performing Markov model analysis on each cluster obtained by applying K-Means clustering algorithm on user sessions. Let C = {c_1,..., c_m) be the set of pages in a website and N be a user session which has the series of web pages visited by the user. If the user has visited 'v' pages, then the probability that a user can visit a page p_i is denoted as prob(c_i|N). The probability of page C_{v+1} can thus be estimated using Equation (1).

$$CP_{r+1} = \operatorname{argmax}_{p\in P}\{C(C_{v+1} = C|N)\} = \operatorname{argmax}_{C\in C}\{C(C_{v+1} = C|C_v, C_{v-1},..., C_1)\} \qquad (1)$$

From Equation (1), it can be seen that the probability is estimated by using sequences of all users in history (or training data) denoted as N. CAPKAAA reduces the complexity, when assumed that all visited pages follows a Markov process with K-means and Agglomerative. Thus, while using Markov model with K-means, the probability of visiting a page depends only on a small set of k preceding pages and not on all the pages in the session. Using this assumption, Equation (1) can be rewritten as Equation (2).

$$N_{v+1} = \operatorname{argmax}_{n\in N}\{N(Nv_{+1} = n|nv, nv_{-1},..., nv_{-(k-1)})\} \qquad (2)$$

Here, k is the number of previously visited pages and it is used to identify the order of Markov model. The model using Equation (2) is referred as kth Order Markov Model (KMM). Thus, in order to use the KMM model, the learning of nv+1 is needed for each sequence of k web pages. Let S_j^k be a state with k number of preceding pages denote the Markov model order and j be the number of unique pages on the website. The probability of N (ni|S_j^k) is estimated using Equation (3) which is obtained from the historical or training dataset.

$$N(n_i \mid S_j^k) = \frac{frequency\ (< S_j^k, n_i >)}{frequency\ (S_j^k)} \qquad (3)$$

The above equation calculates the conditional probability as the ratio of number of times a sequence S_j^k occurs in the training set to the number of times the page p_i occurs immediately after S_j^k. All pages that satisfy the condition probability are selected as pages which may be visited by the user. In order to improve the coverage and reduce the complexity of the model, two modified Markov models are used. They are, All kth Markov Model, Agglomerative Markov Model and Kth model (CAPKAAA). The main aim of CAPKAAA is to determine a Markov model that leads to high accuracy with low state space complexity. Most of web usage mining algorithms are based on data mining techniques like Association Rule Mining, Markov Modeling and clustering. Navigation patterns are analyzed by taking into consideration the current user browsing activities for predicting their future requests. The various phases of doing so in arriving at a better prediction of web page visit is chronologically arranged. Initially web log file is preprocessed to remove unwanted entries. Secondly, potential users are identified from non-potential users. Thirdly, clustering is performed using K-means and Agglomerative Clustering technique [7] through Data Similarity factors. In the fourth and final step an attempt on prediction using this modified Algorithms of K-Means combined with Agglomerative Algorithm and K-means algorithm [8] is made. Markov Models [10] are generally used to predict the next page that might be accessed using the users' browsing history which is stored in Web-logs. This research work proposes an improved cluster mechanism that takes into consideration the proximity similarity measure. Although we are having many clustering schemes in use for web mining, we have attempted to apply cluster algorithms' like K-means and the agglomerative approach clustering together along with Markov's prediction technique for achieving better prediction. K-means clustering algorithm alone may not be sufficient enough for prediction of the next web page, hence we have attempted to use the agglomerative approach together with K-means for arriving at accurate predictions. Results point out that that the selected clustering algorithms of K-means and Agglomerative paved way for arriving at exact predictions. As the question in this ever changing world of the inability to predict the user's next web page visits, web page prediction gathered importance. In this paper, prediction is got by the preprocessing the web logs and thereby integrating the two techniques of K-means and Agglomerative or bottom up clustering approach with the efficient use of Markov Model. Initially we have used clustering technique for preprocessing the given data into meaningful clusters, using both the cluster algorithms of K-means and Agglomerative algorithm. Secondly, these clusters so created are used as training data while predicting web pages using the prediction model Markov. We present the comparison study between these K-means and K-means with Agglomerative approach in predicting a web page using Markov prediction algorithm. Hence after the experiments conducted we conclude that this study will help in reaching at a better prediction of the user's next web page visit.

4.2 Research Methodology

The research aims to increase the Web page prediction accurately by combining two clustering techniques like K-means and Agglomerative approach with Markov prediction model. An approach is introduced to integrate [9] three algorithms K-Means Clustering, Agglomerative clustering approach and the Markov Model [10]. We are aiming at both efficiency and quality which we can get by using K-means and agglomerative clustering as proved in the experiments conducted by us. The basic step combining k-means clustering and Agglomerative approach is given below. Initially the counts of clusters K are obtained and at the same time assume the centroid from the so obtained clusters. The combination of K-means and Agglomerative algorithm for obtaining the data that would be used to train the prediction algorithm Markov Model for prediction of web page. The steps required are shown in the figure given in the next page.

Steps included in CAPKAA

Step 1. Initialize data set of the web page visited on the server.

Step 2. Define K number of clusters.

Step 3. Determine from the dataset its centroid.

Step 4. Compute the proximity matrix between two clusters taking centroids as a measure to combine and group the clusters which are equidistant

Step 5. Repeat this process starting from the Step 2 until all clusters are not complete.

Step 6. Using Agglomerative approach find the proximity between clusters and merge the two clusters which are the closest

Step 7. Update this proximity matrix that reflect similarity between the current cluster and the previous cluster

Step 8. Iterate till all clusters are updated except for one.

Step 9. Next these K clusters are used as a basic training set for algorithm used for prediction.

Step 10.Determine the subsequent visits by a user using Markov model.

Step 11. Isolate the value that has less value than the average value.

Step 12. Iterate again to isolate the value which is below than the average value

Step 13. Locate the subsequent and preceding page accessed from the group.

Step 14. Compute the potent value between the obtained values

Step 15. The value that has is most potent will be represented as the strongest value.

5. Experimental results

Web users do face the problem of resource overload because of the voluminous amount of information and the increase in the count of internet people. It's estimated that there are well over one billion sites on the Web today, an amazing number. As of July 2016, the Indexed Web contains at least 4.75 billion pages, according to WorldWideWebSize.com. Clustering algorithms on different entropy factors are compared in this paper.

5.1 Calculation of Entropy of K-means, Agglomerative and Combined K-means (CAPKAAA)

Table below shows the entropy value of k-means, Agglomerative and combined K-means with Agglomerative which have been obtained by calculating value iteratively as against the number of web pages to get the accurate value of entropy. If the high the entropy values are received, the system will be more disordered. Hence, the Table 1 is given below that, to get the quality of Agglomerative technique and the efficiency of K-means, combining both would give the edge for performance improvement. K-means as the entropy values are higher as the number of web pages increase, compared to the other two clustering techniques even though it is efficient. On the other hand the Agglomerative values are comparatively better than K-means and also quality wise better than K-means. Since the both the efficiency aspect and quality aspect of both the algorithms are used, the combined algorithms are tested to get a result which has both the efficiency and quality of the K-means and Agglomerative clustering as shown in CAPKAAA.

Table 1: Comparison of Entropy for K-means, Agglomerative and CAPKAAA

Number of Web pages	k-mean entropy	Agglomerative entropy	Combined K-means (CAPKAAA)
1000	0.350	0.141	0.246
2000	0.359	0.146	0.253
3000	0.358	0.155	0.257
4000	0.360	0.162	0.261
5000	0.369	0.169	0.269
6000	0.372	0.173	0.273
7000	0.402	0.177	0.290
8000	0.411	0.189	0.300

Figure 1 in the next page denotes that entropy increases as the web page clusters increase which means obviously, the quality of the clusters do decrease as the number of records increase. Combined K-means with Agglomerative algorithm (CAPKAAA) provides better efficient and quality clusters when compared to the conventional k-means and Agglomerative algorithm. Although Agglomerative algorithm seems to be better, we arrive at a conclusion that combining it with K-means certainly boosts the performance of better clustering as it can be visualized in the graph below, which is neither at the bottom nor at the top but has all the better qualities of K-means and Agglomerative clustering. We already know that agglomerative approach gives quality [10] and k-means give us efficiency.

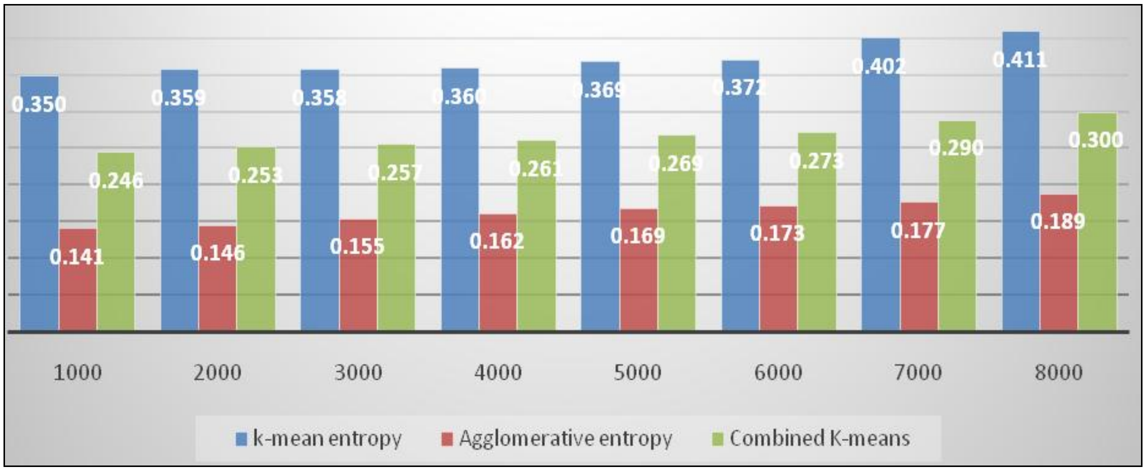

Fig 1: Entropy of Cluster Techniques used

Performance is evaluated by its performance on preprocessing algorithms. Four datasets from web logs as shown in Table 2 in the next page is used for concluding at a prediction. The K-means and K-means combining Agglomerative clustering with Markov algorithms shows the implications by measuring window size 4, threshold of 90% confidence and with minimum support of 4%. The experiments are designed here for assessing the prediction algorithms in line with performance metrics namely, precision, coverage, F1-Measure and speed. Figure 2 shows the efficiency of the prediction models with respect to accuracy performance measure. From the accuracy results, it can be visualized that the performance of the K-means with Agglomerative clustering is excellent. Maximum performance is shown by the model that combines clustering by Agglomerative approach and K-means clustering algorithms and Markov prediction model. This model improved the precision process by 17.15%, 12.17% and 6.44% respectively over the conventional K-means and Markov based prediction model. A related tendency was visualized with coverage (Figure 3) and F1-Measure (Figure 4) performance measures also. But this drift changed while analyzing the algorithms with execution speed (Figure 5).

Table 2: Datasets Used

Datasets	Codes	Dated	Size (MB)	Total Records
NASA Kennedy Center Space (http://ita.ee.lbl. gov/html/contrib/NASA-HTTP.html)	NASA	01-07-95 to 31-08-95	205.2	34,61,612
University of Saskatchewan's (http://ita.ee.lbl. gov/html/contrib/Sask-HTTP.html)	SASK	01-06-95 to 31-12-95	233.4	24,08,625
ClarkNet Internet Service Provider (http://ita.ee. lbl.gov/html/contrib/ClarkNet-HTTP.html)	CN	24-08-95 to 10-09-95	171	33,28,587
University of Calgary's, Department of Computer Science (http://ita.ee.lbl.gov/html/contrib/ Calgary-HTTP.html	CL	24-10-94 to 11-10-94	52.3	7,26,739

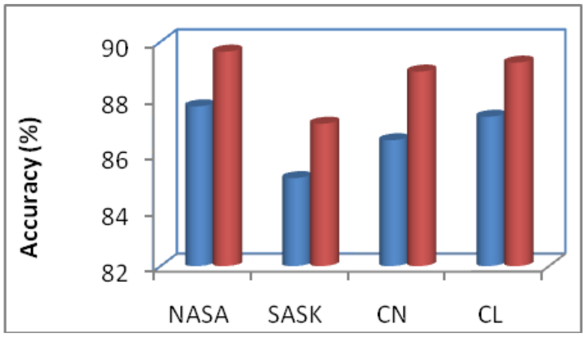

Fig 2: Precision (%)

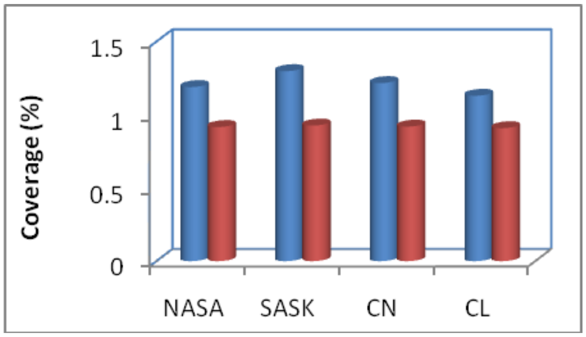

Fig 3: Coverage (%)

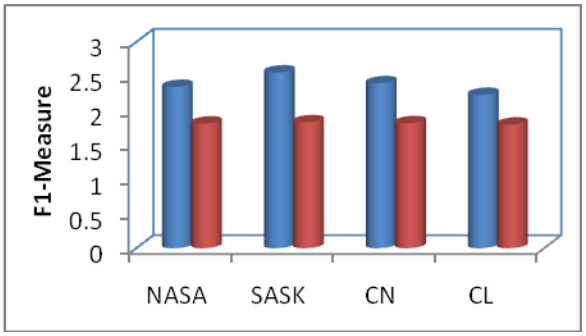

Fig 4: F1-Measure

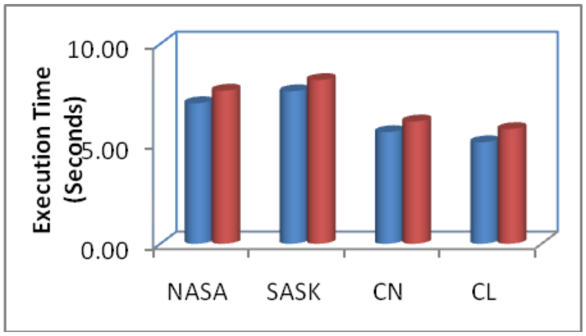

Fig 5: Execution Speed (Seconds)

🟦 Markov with K-means method 🟥 Markov with K-means and Agglomerative method

6. Findings of the study

The researchers explored to analyze the browsing pattern of users to figure out whether the traditional Agglomerative clustering would be a reasonable trial for web page prediction. The selected model displayed better results as against the other cluster approaches. The results also proved that the K-means cluster model when combined with Agglomerative Clustering, produced good accuracy when used with Markov model which was as good as any other model in this contemporary period. The experiments conducted also showed a remarkable performance improvement. The results after the experiment prove that through Hierarchical Clustering technique (Agglomerative) with the traditional K-Means algorithm has escalated the performance of Markov model in its prediction accuracy as compared with other clustering mechanisms with Markov model.

7. Conclusion

The K-Means Clustering algorithm combined with Agglomerative Cluster Technique proves better than other Clustering techniques being used for prediction. Future research ideas may also test model-based, and other cluster analysis algorithms like Divisive approach, Density based, Grid based, model based, constraint based with the aim of further improving the accuracy of web page prediction.

Various types of users browse internet and it is intricate to understand the browsing habits and their intention. There are various models available in this current age for prediction of web pages that may be visited by these users. Hence it becomes even more difficult to predict the same. To solve this issue, the selected web page prediction model Markov was used to predict user's next access using a combination of K-means clustering, and Agglomerative approach. The Markov models combined with clustering and agglomerative algorithms are more efficient in terms of accuracy when compared to other model using K-means algorithm. Future research ideas include methods to identify other clustering algorithms like top down model with the aim of further improving the accuracy of web page prediction.

8. References

1. Dongre V, Raikwal J. An Improved User Browsing Behavior Prediction using Regression Analysis on Web Logs International Journal of Computer Applications, 120(19):0975-8887.
2. Piyush Rai C. Data Clustering: K-means and Hierarchical Clustering –5350/6350, 2011
3. Bagiwa AM, Dishing SI. International Journal of Computer Trends and Technology-March to April issue ISSN: 2231-2803-1-IJCTTA Conceptual Framework For Extending Distance Measure Algorithm For Data Clustering.
4. Madhulatha TS. An overview on clustering methods. IOSR J Eng. 2012; 2(4):719-725.
5. Singh M, Kaur K, Singh B. Cluster algorithm for genetic diversity. World Acad Sci Eng Technol. 2008; 2(6):432-436.
6. Vijayakumar M, Prakash S, Parvathi RMS. Inter cluster distance management model with optimal centroid estimation for K-means clustering algorithm, Article in WSEAS Transactions on Communications. 2011; 10(6):182-191.

7. Improved Web Prediction Algorithm Using Web Log Data, Megha Jharkad P, Prof. Mansi, Bhonsle, International Journal of Innovative Research in Computer and Communication Engineering An ISO 3297: 2007 Certified Organization, 2015, 3(5). Copyright to IJIRCCE DOI: 10.15680/ijircce.2015.0305176 4902
8. Meenu Brala, Mrs Mamta Dhanda. International Journal of Computer Science & Communication Networks - An Improved Markov Model Approach to Predict Web Page Caching, ISSN: 2249-5789,
9. Khalil F, Li J, Wang H. Integrating recommendation models for improved web page prediction accuracy, Thirty-First Australasian Computer Science Conference ACSC 2008, Conferences in Research and Practice in Information Technology CRPIT. 2008; 74:91-100.
10. Manpreet kaur, Usvir Kaur. Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection, International Journal of Advanced Research in Computer Science and Software Engineering
11. Michael Steinbach, George Karypis, Vipin Kumar. Comparison of Document Clustering Techniques, Department of Computer Science / Army HPC Research Center, University of Minnesota.
12. Ananthi Sheshasayee, Sharmila P. Comparative Study of Fuzzy C Means and K Means Algorithm for Requirements Clustering, Indian Journal of Science and Technology.