Enhanced Model of Web Page Prediction using Page Rank and Markov Model

Soumen Swarnakar
Assistant Professor
Department of IT
Netaji Subhash Engineering College,Technocity,Garia, Kolkata-152, India

Anjali Thakur, Debapriya Misra
Student, Department of IT
Netaji Subhash Engineering College, Technocity, Garia, Kolkata-152, India

Debopriya Paul, Moutrisha Pakira, Sreyashi Roy
Student, Department of IT
Netaji Subhash Engineering College, Technocity, Garia, Kolkata-152, India

ABSTRACT
Now-a-days, with the massive size of the web it has become hectic as well as time-consuming for the users to search for the most appropriate page in least time. Prediction of the next page saves users' time and it becomes easy for the user to reach the most suitable or correct page. In this paper, web page prediction technique has been improved by combining clustering with markov rule and page ranking algorithm. The K-means clustering technique is used for the accumulation of the similar web pages. Page Rank Algorithm is used here to assign probabilities to web-pages from beforehand according to their importance. Markov rule has been used on each cluster to evaluate occurrences of each web pages visited under different sessions and markov model is applied to predict the next web page from the current web page. The rule of transition probability of markov model has been used to predict the next web page from the current web page.

Keywords
Web page prediction, Markov Model, K-means clustering algorithm, Page Rank algorithm, Transition probability.

1. INTRODUCTION
In today's world, Internet with a vast ocean of information and data which connects various personal and different computer networks. It consists of millions of private, public, government networks from local to global source. One of the major advantages of the internet is that it can be accessed anytime and anywhere. Now-a-days Internet or the Web is used for a variety of purposes which includes finding a location, search maps, finding routes, sending and receiving mails. Whenever a user accesses the same web-page a number of times, the data is saved in the web-server log files. The next time when the user wants to access the same web-page, the usage patterns from web-data is used to serve the needs of the users. It not only saves time for the users but also helps to find the most appropriate information.

In this paper, Web page search is optimized using the concepts of page rank algorithm and mean value calculation. It uses the history of web page visit. The objective is to identify the successive requests of the user, given the current request that the user has made. By doing this, it controls the load on the server and thus reduces the access time.

In the K-Means Clustering Algorithm, the number of clusters are pre-determined, is one of the simplest unsupervised learning clustering algorithm. In this clustering algorithm, cluster-centers are arbitrarily selected at first, then calculating Euclidean Distance or any other distance measures between each data-set and the cluster-centers the data-points are assigned to the most suitable cluster.

The Page Rank Algorithm is used to rank the web-pages according to their comparative importance. The more the number of in-bound links (i.e. incoming links) to a page from other pages the more is its importance. A Page that is linked by many pages has a high Page Rank value.

Web-server Log files are the files that are robotically created and maintained by a server. It has the navigated web-pages by a user in a user's session where various information including client IP Address, requested page address, date/time of the request, HTTP Code, bytes are saved. Analysis of the log files gives information about users' navigational behavior, time spent on a page, etc.

The paper has been divided into four sections. Section 2 describes the related work, section 3 describes methodology containing proposed model while Section 4 illustrates the analysis of the result. The conclusions are summarized in Section 5.

2. RELATED WORK
Web Page Prediction is such an interesting topic to cruise upon, quite a lot of algorithms have been already proposed. Ruma Dutta et al. [1] used markov and clustering algorithms to determine next web page. In that paper it has been proposed that the predicted web page would be the next web page if and only if the prediction accuracy is higher. Megha P. Jarkad et al. [2] proposed how user future request can be predicted in smaller time using clustering, classification. B. Nageswara et al. [3] presented a new document clustering method based on correlation implementing indexing. It simultaneously maximizes the correlation between the documents inside cluster and minimizes the correlation between the documents outside the cluster. Sunil kumar, Ms. Mala Kalra et al. [4] presented a survey of web page prediction technique by using Markov model. P. Parthasarathi et al. [5] presented a clustering method based on correlation preserving indexing and the use of natural language processing tool. Meenu Brala et al. [6] explored various applications of web usage mining. They also analyzed its problems and challenges.

3. METHODOLOGY
The final step which results in the predicted page or the next page consists of a series of steps. Each step has its own unique importance in calculating that particular page. Preprocessing, Clustering, Page Rank algorithm, Markov model are the major steps among them. The methodology is discussed below.

3.1 Preprocessing

Preprocessing is done to represent the data in form that can be used for clustering. The process of preprocessing is described below and shown by figure 1.

1) Remove all stop words (like – '.', '?' etc).
2) Then prepositions, conjunctions and articles are removed.
3) Stemming is done to stop or restrict the flow of word. For example, the Word 'seriously' becomes 'serious' after the stemming operation.

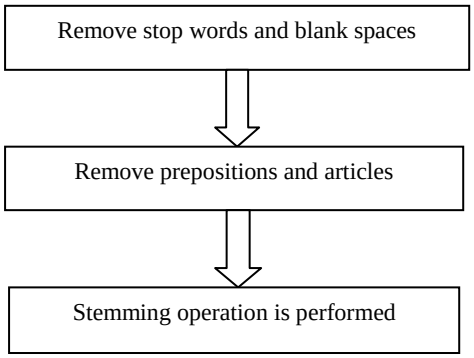

Fig 1: Flowchart of preprocessing

3.2 K-Means Algorithm

K-Means clustering technique is a part of static clustering technique. The k-means clustering algorithm is well-known for clustering large data sets. This clustering algorithm is one of the simplest unsupervised learning algorithms that solves clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes, into k clusters. The main objective is to find out the cluster centroid for each cluster. The centroid of a cluster is formed using the concepts of similarity function and euclidean distance to all objects in that cluster.

The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in previous iteration. The iteration continues until the assignment is stable, that is the cluster formed in the current round are the same as those formed in the previous round.

The flowchart and examples of k-means is described and shown in figure 2 and figure 3 respectively.

In this paper web-sessions are clustered using K-Means Clustering Algorithm, for getting group of web-pages cluster wise, which have been used for next web page prediction purpose.

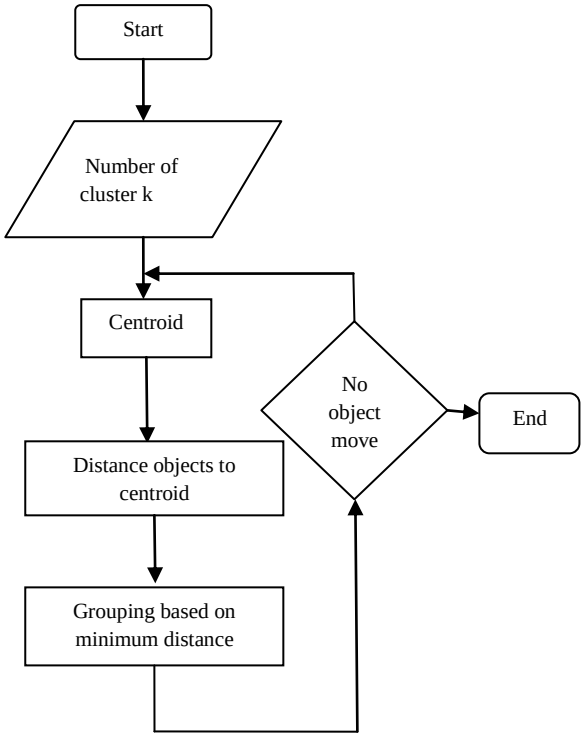

Fig 2: Flowchart of K-Means clustering technique

3.2.1 Steps of K-Means Algorithm

Input:

k: the number of clusters

D: data set containing n objects

Output: a set of k clusters

Method:

{

(i) Arbitrarily choose k objects from D as initial cluster centre.

(ii) repeat

{

reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

update the cluster means, that is, calculate the mean value of the objects for each cluster.

} until no change.

}

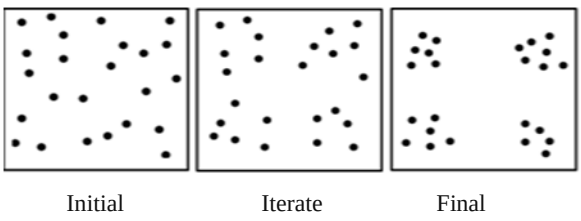

Fig 3: Illustration representing k-means partition based algorithm

3.3 Web Sessions

Session is the term used to refer to a user's time browsing a website. It is meant to represent the time between their first arrivals at a page in the site until the time they stop using the site. Session object can be used to store variables specific to a particular user and web server will maintain these variables when the client moves across pages within website. In most servers there is a timeout that automatically ends a session unless another page is requested by the same user.

There are some web sessions which are shown below in the sequence visited web pages each after another and from these sessions the first order and second order transition probability matrices are occurred. In the example given below, consider the web session WS5: {P4; P3; P1; P4}. Here markov model is used to predict next action from previous actions performed by the user. For the 1st order Markov model page P4 corresponds to state S4. For 2nd order markov model, page P4, P3 is used to predict the next web page. Once, the Transition Probability matrix (TPM) is fully updated, web-page prediction becomes easy. TPM is used to find the page with highest probability so that we can predict it.

In this example, consider the first web session's pages P1, P2, P3. Now P3 is associated with the session WS3. So, after calculating the 1st order markov model, which is shown in Table 1, P1 is the predicted web-page that is accessed by the user next.

In table 2 same processes is performed by using 2nd order of markov model taking two pages. Some of those are shown in the table 2.

Let the web Sessions are as follows:
WS1: {P1; P2; P3}
WS2: {P2; P3; P1; P4}
WS3: {P3; P1; P2; P4}
WS4: {P2; P4}
WS5: {P4; P3; P1; P4}

Table 1: First order Transition Probability Matrix of the sample web sessions

1st Order	P1	P2	P3	P4
S1={P1}	0	2	0	2
S2={P2}	0	0	2	2
S3={P3}	3	0	0	0
S4={P4}	0	0	1	0

Table 2: Second order Transition Probability Matrix of the sample web sessions

2nd Order	P1	P2	P3	P4
{P1,P2}	0	0	1	1
{P1,P4}	0	0	0	0
{P2,P3}	1	0	0	0
{P2,P4}	0	0	0	0
{P3,P1]	0	1	0	2
{P4,P3}	1	0	0	0

3.4 Page Rank Algorithm

PageRank has been introduced by Google and is known after Larry Page, Google's co-founder and president. The main part of any information retrieval system is ranking. Nowadays search engines may return million of pages for a certain query. It becomes hectic for a user to preview all the returned results. So, page ranking is useful in web searching. Page Rank algorithm is the link-analysis algorithm used to assign numerical weightage to web-pages which determines the relative importance of web-pages. For computing a ranking for each and every web page which is based on the graph of the web page rank is also helpful. PageRank is designed to simulate the behavior of a web surfer who navigate a web by randomly selected some links. It is used by Google Search Engine to rank web-pages and web-documents in the search engine results. More the number of times a web-page is visited more is its rank and it can be retrieved from the search engine results earlier. Page Rank of a given page is this number divided by the total number of pages the surfer has browsed. In the paper "Proposed Approach for Web Page Access Prediction Using Popularity and Similarity Based Page Rank Algorithm" [7] Phyu Thwe suggested about the justification of using page rank for ranking web pages comes from the random surfer model. The number of times the surfer has visited each page is counted for calculating page rank. Page Rank is used to measure the importance and behavior of web pages. It has applications in search, browsing, and traffic estimation.

3.5 Markov Model

Andrey Andreyevich Markov (June 14, 1856 – July 20, 1922) was a Russian mathematician. He became famous for his work on the theory of stochastic Markov processes. In future his research work came to be known as Markov process and Markov chains. He founded the Markov chains in 1906 when he produced the first theoretical results for stochastic processes by using the term "chain" for the first time. Markov model is a mathematical model that makes it possible to study complex system by establishing a state of the system and then effecting a transition to a new state. There are four common Markov models such as Markov Chain, Hidden Markov Model, Markov Decision Process, and Partially Observable Markov Decision Process. It has been widely used for predicting next web page from user's web log record. The probability of visiting a web page does not depend on all the pages in the web session. The state-space of a markov model depends on the number of previously performed web pages that are used in predicting the next web pages. In the literature, different kinds of Markov process are designated as "Markov chains". In First Order Markov Model the next page is predicted by only looking at the last web page performed by the user but in case of second order of Markov Model, the last two web pages visited by the user is looked upon and in case of K-order Markov Model, the next page is predicted by looking at last K number of web pages. It is widely used in physics, chemistry, testing, speech recognition, information sciences, queuing theory, internet applications, statistics, economics and finance, social sciences, genetics, games, music, Markov text generators and bio-informatics. The first-order Markov models have got numerous successes in many sequence modeling and in many control tasks.

In this paper, web-page prediction has been improved by combining clustering of web sessions and page rank algorithm using Markov Rule.

3.6 Proposed Approach

1. At first collect all the web-sessions from the web-log containing visited web-pages.
2. Now preprocess the web-sessions for clustering purpose.
3. Now web-sessions are clustered using K-Means Clustering Algorithm, for getting group of web-pages cluster wise.
4. Evaluate probability(P) of accessing of a web-page in a web session by the rule given below:-

Let,

X=Total no. of access of a web page within a session.
Y= Total no. of access of a web page in all session.

Now, Probability (P) = $\frac{X}{Y}$,

5. Finding the Page Rank (PR) for each web page by the formula given below:-

$$PR=\mu*\frac{A}{B}+(1-\mu)*C.$$

Where μ is the damping factor (μ is very small number, experimentally found to be 0.85).

A= Probability of the current web page.

B=∑Total number of outbound links, where outbound links is the no. of outward links from a current web-page to another web page.

C= Probability of the next web page.

6. The Mean value is calculated:

$$Mean=\frac{\sqrt{(\text{Maximum PR})^2+(\text{Minimum PR})^2}}{2}$$

Maximum PR=Maximum web-page Ranking among the candidate web-pages from the current web-page.

Minimum PR=Minimum web-page Ranking among the candidate web-pages from the current web-page.

Now remove the web-pages whose Page rank value is less than the Mean value.

7. Transition Probability for 1st Order Markov Model is calculated for web-pages having values greater than mean value by the formula given below:-

$$TP_i^j=\frac{XY_i^j}{\sum_1^L XY_i^j}$$

Where XY_i^j=No. of times the access is made where i is the current web-page and j is the next web-page.

L=No. of outbound links.

TP_i^j=Transition Probability from i^{th} page to j^{th} page.

8. The next web-page is predicted from the highest value among all the transition probabilities.

4. EXPERIMENTAL RESULTS

Web-sessions are collected from the web-log containing 20 pages like www.flipkart.com, www.amazon.in, www.snapdeal.com, etc. After preprocessing of the web-sessions, they are clustered using K-Means to form 7 clusters. The result of clustering has been shown in Table 3. Probability of each web page in a session is calculated and using that, PageRank for each page is calculated and the Mean value is also calculated. Then, the pages having values less than the Mean are removed which ultimately results to pages 2, 4, 9, 15, shown in table 3. Finally, Transition Probability is calculated and depending on that value, the next web-page is predicted which is shown in table 4.

Table 3: K-Means Clustering result

Cluster No.	Web pages			
C1	2	10	5	16
C2	4	7		
C3	6	9	12	
C4	13	17	20	
C5	15	18		
C6	3	11	14	
C7	1	8	19	

Table 4: Result table

Current Web-Page	Predicted next web-page	Transition Probability
2	10	0.68
2	**5**	**0.75**
2	16	0.72
4	**7**	**0.65**
9	6	0.71
9	**12**	**0.77**
15	**18**	**0.73**

The above table shows that the transition probability from current page 2 to next web page is page 5 because of highest probability 0.75, whereas predicted next page from page 4 is page 7. Predicted web page from page 9 is page 12 with highest probability 0.77, whereas predicted web page from page 15 is page 18.

5. CONCLUSION

Web-page prediction improves the experience of the users on the web. It helps to optimize the search results and the user can effectively retrieve the most appropriate results. Here, Page Ranking and 1st order Markov Model is considered for Web Page Prediction. For PageRank Algorithm, the probability of each web-page is used to determine the PageRank. The Mean value is considered to find the maximum visited pages. This method of approach considers both Transition Probability and PageRank Algorithm to predict the next web-page. In the proposed paper, after preprocessing of the web-sessions, web-sessions are clustered using k-means. In the next step, probability of each web page is calculated which determines the PageRank of web-pages. In the final step, Transition Probability is calculated between web-pages to predict the most suitable pages. . However to work with the higher order Markov Model some more features should also be considered.

6. REFERENCES

[1] Dutta, R., Kundu, A., Dattagupta, R., Mukhopadhyay, D. 2009. An Approach to Web Page Prediction Using Markov Model and Web Page Ranking. Journal of Convergence Information Technology.

[2] Jarkad, M.P., Bhonsle, M. 2015. Improved Web Prediction Algorithm Using Web Log Data. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5.

[3] Rao, B.N., Keerthi, P., Sree Ramya, V.T., Kumar, S.S., Monish, T. 2014. Implementation on Document Clustering using Correlation Preserving Index. International Journal of Computer Science and Information Technologies (IJCSIT), pp. 774-777, Vol.5, issue 1.

[4] Kumar, S., Kalra, M. 2013. Web Page Prediction Techniques: A Review. International Journal of Computer Trends and Technology (IJCTT). Vol.4, Issue 7.

[5] Meena, S. U., Parthasarathi, P. 2013. Correlation Preserved Indexing Based Approach For Document Clustering. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol.2, Issue 2.

[6] Brala, M., Dhanda, M. An Improved Markov Model Approach to Predict Web Page Caching. International Journal of Computer Science & Communication Networks, Vol. 2(3), 393-399.

[7] Phyu, T. 2013. Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm, International Journal of Scientific & Technology Research, Vol.2, Issue 3.

[8] Eirinaki, M., Vazirgiannis, M., Kapogiannis, D. 2005. Web Path Recommendations based on Page Ranking and Markov Models. Proceeding Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 2-9.

[9] Fosler-Lussier, E. 1998. Markov Models and Hidden Markov Models-A Brief Tutorial. International Computer Science Institute.

[10] Swarnakar, S. 2012. Ontology-based context dependent document clustering method. Int. J. Knowledge Engineering and Data Mining, Vol.2, No.1.

[11] Deshpande, M., Karypis, G. 2004. Selective Markov Models for Predicting Web Page Accesses. ACM transactions on Internet Technology, Vol.4, No.2, pp.163-184.

[12] Anitha Elavarasi, S and Akilandeswari, J. 2014. Survey on clustering algorithm and similarity measure for categorical data. International Journal On Soft Computing, Vol.4, Isuue 02.

[13] Han, J. and Kamber, M. 2001. Data mining-concepts and techniques.