COMPOSITE MODEL OF WEB PAGE PREDICTION USING PAGE RANKING ALGORITHM AND MARKOV MODEL

Priyanka Bhart[1], Diwakar Shukla[2], Dayal Sharan Saini[3], Aman Gaur[4]

[1]Asst. Prof., [2,3,4]Student, Dept. of Computer Applications,
National Institute of Technology Kurukshetra, India)

ABSTRACT

As the recent Advent of the internet and various technologies, most of the people are heading towards it and because of that the amount of data is also increased. The web mining technology provides the efficient way through which user can get the appropriate page that in less time. Prediction of the next page will save user's time and improve the quality of work on the web. In this paper webpage prediction concept has been implemented through Markov model and page ranking algorithm. The k means algorithm is one of the simplest and finest ways for solving the problems of clustering by dividing the various webpages into clusters. Page Rank algorithm is also used for assigning the probabilities to each page depend upon no. of time that page is been accessed. Markov model has played a key role on clusters for finding out the no of occurrences of the webpage under various web-sessions and also for predicting the next webpage through the current web page. Various type of matrix (such as transition, similarity etc.) is one of the roots of webpage prediction concept.

Keywords- *K-means clustering algorithm, Markov model, Page ranking Algorithm, Transition Probability, Web page prediction.*

I. INTRODUCTION

In today's world, the web is one of the important sources for our daily life. The Internet contains lots of information and data which can be used for the public as well as private. It connects through various public and private networks. The advantage of the internet is that it can be used from anywhere and anytime. Nowadays the Internet can be used for the variety of things such as Banking, Travelling, Communication, Search maps, sending and receiving emails etc. When any user accesses a webpage more than one time than for their convenience the data is saved in web log files. Now If next time the user accesses the same page then instead of loading the same page again the browser will open the page directly from the data that is saved in log files. This process is only for the betterment of the user by saving the time and also for an appropriate page at given time. This paper focuses on Page Ranking Algorithm and Markov Model through which Web page Prediction is done.

II. RELATED WORK

Web mining emphasizes on knowledge extraction from the web pages. Web usage mining is focused on gathering information from the web log files accessed by the users. Web content mining is defined as the examining and mining of text, graphs, and pictures from a Web page to find out the importance of the search query. Many of the authors usually proposed their work on identifying the user's web page prediction. Most of the models developed in the literature to understand user's behavior. Eirinaki et al. [1] propose a method that includes link analysis, such as the page rank measure, into a Markov model. Schechter et al. [2] employ a tree-based data structure that depicts the collection of paths inferred from the log data to predict the next page access. Fosler-Lussier et al. [3] described working on Hidden Markov model using transition matrix. Ahmad and Sultan [4] proposed a system model which involves web mining techniques with an e-commerce application.

III. APPROACH

The final output in predicting next webpage consisting of a set of steps. Each step has it is own significance. Preprocessing, clustering, Makov model, User session are major steps among them. The Approach is discussed below.

3.1. Preprocessing

Preprocessing is performed because the dataset have many special characters (like ‘@’, ‘.’, ‘;’ etc.), prepositions, articles etc. so that data will be reduced by removing these things. This data will be used for clustering. Emanate by preprocessing to outflow words (for example, “framing“ is converted “frame”). Given figure 1 is showing how preprocessing is performed.

3.2. K-Means Algorithm

K-Means algorithm is an advanced algorithm that gains its name from its method of operation. This data mining or machine learning algorithm is well-known to cluster observation into groups of related observation without any prior knowledge. It is used to divide a set observation into k groups of clusters, where k is an input parameter. Then it randomly selects k points as the centroid of clusters. Recalculate the centroid or mean by assigning observation to their closest cluster centroid by using Euclidean distance function.

This iterative approach of clustering algorithm improves the revised centroid of clusters. Iteration remains persistent until the clusters converge.

In this paper, web sessions are clustered using K-Means Clustering Algorithm, for creating a group of webpages in the form of cluster, which has been used for predicting next web page.

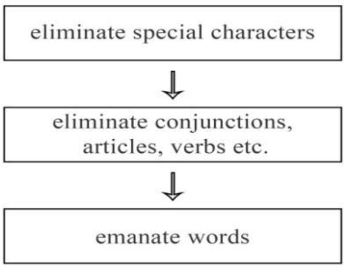
eliminate special characters
eliminate conjunctions, articles, verbs etc.
emanate words

Fig 1. preprocessing steps

3.2.1 Algorithmic steps of K-Means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and K = {k1,k2,…….,kc} be the set of centroids.

1. Randomly select ‘c’ cluster centroid.
2. Calculate the Euclidian distance between each data point and cluster centroid.
3. Assign the data point to the cluster centroid whose distance from the cluster centroid is a minimum of all the cluster centroid.
4. Recalculate the new cluster centroid.
5. Recalculate the distance between each data point and newly obtained cluster centroid.
6. If no data point reassigned then stop, otherwise repeat from step 3.

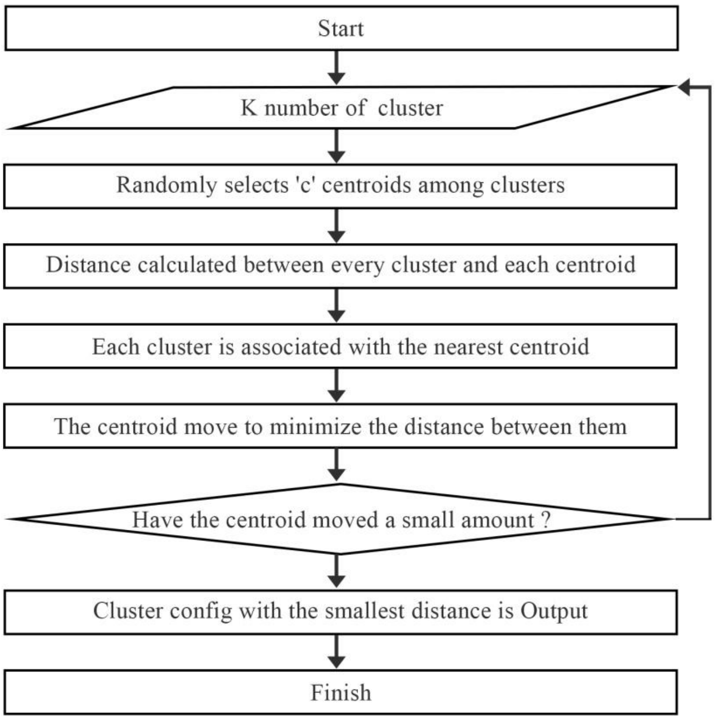
Start
K number of cluster
Randomly selects 'c' centroids among clusters
Distance calculated between every cluster and each centroid
Each cluster is associated with the nearest centroid
The centroid move to minimize the distance between them
Have the centroid moved a small amount ?
Cluster config with the smallest distance is Output
Finish

Fig 2. flowchart of k-means algorithm

3.3. Markov Model

Markov model is a stochastic model used to represent the probability distribution over a sequence of observation. It deals with temporal or sequential data i.e. data that are ordered. A Russian mathematician Andry Markov (1856-1922) provided a way to model the dependencies of current information with previous information. It is composed of states, transaction scheme between states and emission of outputs (discrete or continuous). Markov model is widely used for developing algorithms for predicting the next web page from user's web log record. Markov model can be categorized into four categories on the basis of the system state.

System state is fully observable	System state is partially observable
Markov chain	Hidden Markov model
Markov decision model	Partially observable Markov decision process

Table 1. markov models

Markov chains are the mathematical system that jumps from one state to another. The next state of the process only depends on the previous state and not the sequence of states. This principle makes the calculation of conditional probability easy and enables this algorithm to be applied in a number of scenarios.

Let's take an example, A student has two subjects - Science and Mathematics. He has to cover the syllabus of both the subjects in a limited time. Following are the conclusion drawn by him for his strategy to complete the syllabus:

P(S->S): Probability of a student staying with the Science = 0.7

P(S->M): Probability of a student switching from Science to Mathematics = 0.3

P(M->M): Probability of a student staying with Mathematics = 0.9

P(M->S): Probability of a student switching from Mathematics to Science = 0.1

Above four statements made by the student can be drawn in a simple transition diagram.

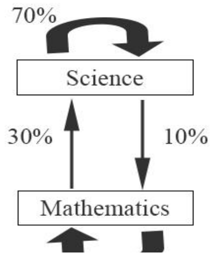

Fig. 3. transition diagram

3.4. User Session

The computer knows who you are and which application you have started and when you end. But the internet doesn’t know this whole information: the web server doesn’t know anything about the user because the HTTP address doesn't maintain state. Session variables solve this Condition by storing user details to be used across various pages (e.g. username, favorite color, etc.). By default, session variables last till user closes the browser. So session variables hold information about one single user, and are available to all pages in one application.

The user has the main advantage that he can check the last page that he has visited with the help of sessions. Now consider a user’s session and the sequence in which he has visited the web page each after another and with the help of these sessions we can generate 1st order and 2nd order Transition Probability Matrix. In the Given example below, we consider a user session(US1) in which user has visited various categories pages(laptop(L) ,camera(C), Sports(SP), Shoes(SH) US1:{L,C,SP,SH}. Now here we’ll apply Markov Model for predict next page from the previous movement executed by the user. For the 1st order Markov Model Page, SH corresponds to state S4. For the 2nd order, page SH, SP is used to predict next web page. Web prediction becomes easy only when the Transition Probability Matrix (TPM) is executed in proper manner. For perceiving the Highest Probability of a page, TPM will be used.

In the example, consider the 1st web session’s pages L, C, SP. Now SP is associated with the session US3. So after analyzing the 1st order Markov Model, Which is given in table 2, L is the Predicted Page that is used by the user.

In table 3 some process are executed by using 2nd order of Markov model taking two pages. Some of those are described in the table 2.

Consider some Web sessions

US1 :{L, C, SP}

US2 :{C, SP, L, SH}

US3 :{SP, L, C, SH}

US4 :{C, SH}

US5 :{SH, SP, L, C}

1st order	L	C	SP	SH
S1={L}	0	2	0	2
S2={C}	0	0	2	2
S3={SP}	3	0	0	0
S4={SH}	0	0	1	0

Table 2. 1st order transition probability matrix

2nd order	L	C	SP	SH
{L,C}	0	0	1	1
{L,SH}	0	0	0	0
{C,SP}	1	0	0	0
{C,SH}	0	0	0	0
{SH,L}	0	1	0	2
{SP,SH}	1	0	0	0

Table 3. 2nd order transition probability matrix

3.5. Page Rank

Page Rank algorithm used by Google search engine to determine the rank of web pages. It has been named after Larry Page, Google's co-founder, and president. Google uses Page Rank to adjust results so that sites that are having higher value will goes up in the results page of a user's search accordingly. Google's page rank is based on backlinks. The more links your page gets, the higher its page rank score will be. Page ranking has an important role in web searching because today the web has millions of web pages and to search a right page between them is not an easy task. The number of times the user visited the page, Page Rank of the given page will increase. This number is divided by the total number of pages the user has browsed. The web page is treated as a directed graph G = (V, E), where V is the set of vertices or nodes, i.e., The set of all pages, and E is the set of all pages in the graph, i.e. Hyperlinks [7]. The order to rank the pages Google work by Finding all pages matching the keywords of the search. Rank accordingly using the page factors such as keywords. Calculate in the inbound anchor text, Then Google adjusts the results by Page Rank scores.

3.6. Proposed Alogorithm

1. Gather all the web-sessions from weblog containing visited web pages.
2. Preprocess the data for pattern discovery and analysis.

3. Using K-Means algorithm, cluster the web-sessions for getting the group of webpages cluster wise.
4. Find the probability of accessing a web-page in a web-session by the following formula:-

   Let,

   X = Total No. of access of a web page within a session

   Y = Total No. of access of a web page within a session

   Now, Probability = X / Y
5. Evaluate the Page Rank (PR) for each web page by the rule given below:-

   $$PR = \mu * (A / B) + (1 - \mu) * C$$

   Where µ is the dumping factor (µ is very small, so we usually take it as 0.85)

   A = Probability of the current web page.

   B = Sum of total no. of outbound links.

   C = Probability of the next page
6. The mean is calculated:

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

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

   Minimum PR=Minimum web-page Ranking among the candidate webpage.

   Now remove the webpages whose page rank value is less than the Mean value.

IV. CONCLUSION

Web page prediction improves the user's experience on the web. It helps the user in finding the suitable web page in the least time and also optimize the search results. Markov Model and Page Ranking algorithm are considered here for Web Page Prediction. Here, in the Preprocessing, Special Characters are eliminated so that they can help in making clusters. After pre-processing the k means provides an efficient way for creating the clusters. With the help of Page Ranking Algorithm, Probability of each webpage can be defined easily. The mean value is considered to find the maximum visited page. Transition Probability Matrix is calculated between web-page to predict the most suitable Pages. However to work with higher versions of Markov model many features should be considered.

REFERENCES

[1] Eirinaki, M., Vazirgiannis, M., Kapogiannis, D. 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, 2005, pp. 2-9.

[2] S. Schechter, M. Krishnan, and M. Smith, Using Path Profiles to Predict HTTP Requests," Computer Networks and ISDN Systems, vol. 30, 1998, pp. 457-467.

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

[4] Ahmad Tasnim Siddiqui and Sultan Aljahdali. International Journal of Computer Applications. Vol.69, Issue.8. May 2013

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

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

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