ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Improved K-NN Approach for Automated Web Usage Mining and Recommendation System

Himanshi Kirar
Department of CSE&IT
MITS, Gwalior, India

Punit Kumar Johari
Department of CSE&IT
MITS, Gwalior, India

Abstract: Web Usage Mining (WUM) is to discover interesting patterns of usage from web based data to understand it and to serve the needs of the web based applications in a considerable superior manner. Here in WUM technique, there is a use of Automated web usage data mining and the recommendation system (which works on analyzing the behavior of frequent user) using an improved K-Nearest Neighbor (K-NN) classification method (where it is essential to calculate distance, assigns pattern matching value and weight to calculate the contribution of the neighbors). In this propose a paper Manhattan distance (MD) based nearest neighbor approach is used for the class which is undefined. And also calculate the weights of matched fields with those of the pattern matching value and assign majority or maximum weighted neighbor as an undefined class label. Using With this distance based learning is becoming clear, decrease in computation time, it is easier to know which attribute can produce a better result, and so on.

Keywords: Web Usage Mining (WUM), K-NN, Recommendation system (RS), Manhattan distance (MD) etc.

I. INTRODUCTION

WUM is the third and last category of web mining. It provides path leading to access web pages. The information required is collected automatically into access logs via the web server. CGI scripts provide useful knowledge such as reference logs, user subscription information and survey logs. WUM is very useful for companies and their intranet/internet applications and information processing. It is also beneficial in E-commerce and product oriented user services. WUM can be studied in two different generally utilized methodologies. One is to draft the usage data on the web server into relational tables before any data mining strategies is performed. Another method uses the log data directly by using pre-processing technique. Web usage data is can also be represented in graph form. Web usage can be classified in two ways:

i) Knowledge about user profile
ii) Knowledge about client navigation pattern. An information provider is interested in strategies that could build the adequacy of the information on their websites or we can say they are interested in navigation patterns. A tool such as system enhancement, personalization, site improvement, usage characterization and business intelligence improve the navigation patterns. Web usage data includes activities like browser logs, a web server access logs, and registration data. Proxy server data/logs, User profiles, user queries, cookies, registration data, transactions and mouse clicks etc. It also uses links, text and profiles like transaction records and business records etc. that are concluded by the user [1].

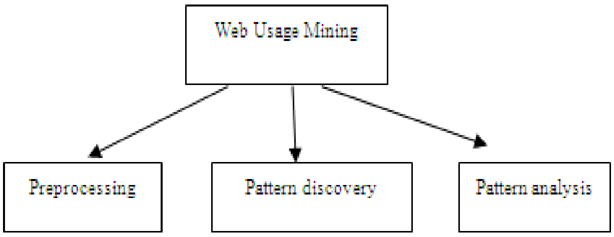


Fig 1. Web usage mining process

II. K- NEAREST NEIGHBORS ALGORITHM (K-NN)

In pattern discovery, this algorithm is a nonparametric technique used for classification. The input consists of k nearest training examples in the feature space. The class membership is an output, in a K-NN classification method. An entity is classified or predicated on the basis of maximum vote of its neighbors and majority of the entities, with the entity being assigned to the class that is most common amongst its k nearest neighbors (k is a positive integer, typically small). In K-NN classification, if the value of k = 1, then the object is assigned to the class of that single nearest neighbor, where k denotes no. of neighbor. This value is average values of its k nearest neighbors. The K-NN algorithm is one of the simplest and easiest, straightforward method of all machine learning algorithms. For classification, it is valuable or more efficient to assign weights to the contribution of the neighbors, so the nearest neighbors can contribute more to the average than the distance ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d represents the distance to the neighbor. The neighbors are taken from a collection of entities for which the class (for K-NN classification) or the entity property value is known. This could be understood as the training set for the algorithm. A defect of the K-NN algorithm is to find a way to predict the unknown class label with the use of various distance methods [2].

III. RECOMMENDATION SYSTEM

RS is based on user behavior patterns and products or items can be recommended that is based on an overall sellers on a site, or on an analysis of the previous purchasing behavior of the purchaser as a prediction of future purchasing behavior. The types of recommendation include suggesting products or items to the consumer, providing personalized items information, summarizing community opinion, and providing community critiques. Comprehensively, these recommendation techniques are part of personalization on a site because they help the site adapt itself to each user. Text summarization is a chain of a reducing a given document into a shortened, modified by extricating the most imperative information from it [3].

RS can be built with many approaches. Below are some of them:

- Random prediction algorithm is an algorithm in which we randomly select an item from the set of available items and recommends them to the user. Since the item's selection is done randomly, the accuracy of the algorithm is based on luck; the greater the number of items is the chance of good selection lowers. Random prediction has a greater probability of failure. Thus, it has never been taken seriously by any researcher or vendor and only serves as reference point1, helping to compare the quality of the results obtained by the utilization of a more sophisticated algorithm.
- Frequent sequences can help build recommender systems. For example, if a customer frequently rates items we can use the frequent pattern to recommend other items to him. The only problem is that this method will only be efficient after the customer makes minimum purchases.
- Collaborative filtering algorithm (CF) is an algorithm that requires the recommendation seekers to express their preferences by rating items. In this algorithm, the roles of recommendation seeker (a user) and preferred provider are merged; the more users rate items (or categories), the more accurate the recommendation becomes.
- Content based algorithms are algorithms that attempt to recommend items that are similar to items the user liked in the past. They treat the recommendation's problem as a search for related items. Information about each item is stored and used for the recommendations. Items selected for recommendation are items that content correlates the most with the user's preferences. For example, whenever a user rated an item, the algorithm constructs a search query to and other popular items by the same author, artist, or director, or with similar keywords or subject. Content based algorithms analysis item descriptions to identify items that are of particular interest to the user [4].

IV. LITERATURE SURVEY

Abhirami.K et al. [2016] this paper describes a solution using genetic algorithm that attempts to discover the rules occurring at the junction of fuzzy set boundaries. Recognizing the relationship of site pages is imperative to improve the client encounter on web route by giving exclusively characterized administrations. In view of the conduct of client on the web, sections are made of logs of web servers and these passages are dug for displaying. Because of dynamic conduct of clients, in the Web use mining process, fluffy affiliation decides that have a fleeting property extricates helpful information when affiliations happen. Notwithstanding, there is an issue with conventional fleeting fluffy affiliation manage mining calculations

Suharjito, et al. [2016] in this paper, we propose to use classification technique with K-NN algorithm implemented with Euclidean distance to classifying and identifying frequent access pattern. The result shows that the K-NN algorithm can be implemented in WUM and can help company to find interesting knowledge in web server log [6].

P. Sukumar, et al. [2016] this paper is mainly related to WUM. The contribution of this paper is based on the investigation of data preprocessing, and is used to determine the effectiveness of the algorithms, its limitations, and their stands are verified. Various preprocessing algorithms and its heuristics are applied and examined by implemented using programming languages. Data preprocessing algorithms are used to parse the raw log files that involve splitting of the log files and then cleansed to obtain superior quality of data. Based on this data, the unique users are identified which in turn helps to identify user sessions [7].

V.Anitha, et al. [2016] this paper gives a consideration on WUM to foresee the conduct of web clients in view of web server log records. Clients utilizing site pages, a successive get to ways and regular get to pages, connections are put away in web server log records. A Web log alongside the uniqueness of the client catches their perusing conduct on a site and talking about with respect to the conduct from investigation of various calculations and distinctive techniques [8].

Doddegowda B J, et al. [2016] in the paper, Data on World Wide Web has been growing in an exponential manner. This raises a severe concern on information overload challenges for the users. Retrieving the most relevant information from the web as per the user requirement has become harder because of the large collection of heterogeneous documents. One approach to overcome this is to personalize the information available on the Web according to user requirements. This is called Web Personalization process that adjusts information/services delivered by a Web to the needs of each user or group of users, taking their behavioral patterns. Frequent Sequential Patterns (FSPs) that are extracted from Web Usage Data (WUD) are very important for analyzing and understanding users’ behavior to improve the quality of services offered by the World Wide Web (WWW). User behavioral patterns are required to build profiles for each user, using which Personalization of a website is made [9].

Changqing Ji, et al. [2016] in the paper, a distributed method of K-NN queries applying Map Reduce program model will be introduced. In the very beginning, we propose distributed methods which set up a novel distributed spatial data index: Inverted Voronoi Index that combines both inverted index and Voronoi diagrams. Next, we propose a K-NN queries, processing algorithm, it is very efficient because it is based on Voronoi and uses Map Reduce. Last but not least, we present the outcomes of extensive experiment that are gained by both real and simulated data sets which indicate efficiency and scalability of the proposed approach [10].

D.A. Adeniyi et al. [2016] in this work, we introduce an investigation of programmed web utilization information mining and proposal framework in view of current client conduct through his/her snap stream information on the recently grew Really Simple Syndication (RSS) pursuer site, so as to give applicable data to the person without unequivocally requesting it. The K-NN order strategy has been prepared to be utilized on-line and in Real-Time to distinguish customers/guests click stream information, coordinating it to a specific client gathering and suggest a customized perusing alternative that address the issue of the particular client at a specific time [11].

V. PROPOSED WORK

a) The problem occurs in existing work:

WUM is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Web usage data capture the identity or origin of Web users along with their browsing behavior on a Web site. In existing technique author apply automated web usage data mining and recommendation system using K-Nearest Neighbor (K-NN) classification method, in this approach value of K is undefined, so the basic problem of this is

- Need to determine the value of parameter k.
- Which attributes are better to use producing the best results.
- Need large computation to calculate distance between test tuple and training tuple. If we want to calculate missing class label, then we have to calculate the distance between missing class and each tuple of the dataset.
- K-NN classification includes irrelevant data.

Solve this problem we proposed a pattern based Manhattan distance based nearest neighbor technique for the web usage recommendation system.

b) Methodlogy

This section present detail description of the implementation of WUM. The presentation of the application of the proposed methodology for the analysis of user's RSS reader website. We have developed a real time, online RS that can assist the administrator to improve the content, presentation of their websites by recommending a set of classes or choice that satisfies the need of active user based on the user's current click stream.

- ***Preprocessing***

In the original database file extracted, not all the information are valid for web usage data mining, we only need entries that contain relevant information. The original file is usually made up of text files that contains large volume of information concerning queries made to the web server in which in most instance contains irrelevant, incomplete and misleading information for mining purpose.

- ***Pattern discovery:***

Pattern discovery is the procedure of web mining, which incorporates grouping of users based on similarities in their profile and search behavior. There are various web usage data mining techniques or methods and algorithms that can be received for pattern discovery and recommendation, which include, path analysis, clustering, and associate rule. In this paper, we have experimented with the Improved K-NN classification technique in order to observe and analyze user behavior pattern and click stream from the pre-process to web log stage and to recommend a unique set of object that satisfies the need of an active user, based on the users' current click stream.

- ***Pattern analysis:***

Pattern analysis is the final stage in WUM which is aimed at extracting interesting rules, pattern or statistics from the result of pattern discovery phase, by eliminating irrelevant rules or statistics. The pattern analysis stage provides the tool for the transformation of information into knowledge.

c) Our approach improved k-nn classification

In the existing work large computation is required for calculating the distance to each tuple and in addition also calculating distance of the irrelevant tuples of the dataset. So, for improving the efficiency, performance of the K-NN classification we proposed a new approach **Improved K-NN classification.** The Improved K-NN enhance the efficiency and provide more relevant and accurate results and provide less computation. Proposed work parameters are Pattern matching values, distance and weights. In proposed work, missing class label finding by calculating the MD and calculating the weights. Basically IK-NN based on Pattern matching value where pattern matching values are used to find out the relevant tuples firstly to find the relevant tuples in the dataset we calculate the pattern matching value through finding the similarity between test tuples and training tuples then calculate the MD. Then on the bases of pattern matching value and MD we will calculate the weight of the tuples.

After calculating the weight majority and maximum weights assign the own class labels as a missing class label and recommended all the web data related to the class. Recommendation system for WUM is one of interesting task because of this we can recommend better class for users, proposed approach first find out pattern of our web after that pattern based MD calculate for undefined class type.

Algorithm:

```
Step 1: Initialize the values
            i, j, k
           R= length of rows

Step 2:  for i=1toR
           If (class== NULL)
            {
              Classless
             }
            Else
             {
            Class full
             }
              i= i+1
```

Step 3: Match the pattern between Test tuple & Training tuple

```
        Test tuple = Classless
        Training Tuple = Class Full
         Pat _Match = 0
         j=1
         If ($X_{k, 1}$==$X_{j,1}$ )
         {
       Then patt_match$_1$ = 1
         }
         Else
         {
         Patt_match$_1$ = 0
          }
          If ($X_{k,2}$ ==$X_{j,2}$)
         {
        then  Patt_match$_2$= 1
         }
        Else
        {
          Pat_match = 0
      }
      If ($X_{k,3}$==$X_{J,3}$)
     {
      Then pat_match$_3$ =1
       }
       Else
      {
       Pat_match$_3$ = 0
       }
       P_val= Pat_match$_1$+Pat_match$_2$+Pat_match$_3$
       j=j+1
```

```
Step 4:  if {
         P_value== 1||2
         Then collect all the rows/tuple whose Pat_match
         Value is 1or2.
         Calculate Manhattan distance
          $D_M$= |$X_i$-$Y_i$|
          Where, $X_i$ = Test tuple
                 $Y_i$ = Training tuples
          Then calculate Weight

         W=  P_val*1/$D_M$*0.5

           P_val= Patern matching values
           $D_M$=     Mahanttan distance
         After calculating the weights of each matched or
          Relevant row assign majority or maximum weight's
         Class as a missing class label.
          }
          Else
         {
         Discard all the tuples whose Patt_Match value is 0
         }

Step 5:  Exit
```

VI. WORKING OF IMPROVED- KNN CLASSIFIER

We can explain the working of Improved-KNN through example.

Let us consider table 1, where is RSS reader site's client stream as a vector with 3 attributes: Daily name News category and Added required feed type, with users represented by X_1, X_2, X_3, X_4, . . ., X_{11} as the class labels as shown in Table 1. Assuming the class of user X_3 is unknown.to find the class of X_3 firstly we divides table1 into 2 sub tables. One table contain all the training tuples and 2nd table contain all the test tuples.

To determine the class of user X_3, we have to compute the find the pattern matching value of each tuples and then calculate the MD only for those tuples who have pattern matching value 1 and 2.

Table 1- The RSS reader's data mart class labels training tuple

User	Daily's name	News Category	Added required feed type	Class
X_1	CNN news	World	www.*world	World
X_2	China daily	Business	www.*business	Business
X_4	CNN news	Politics	www.*politics	Politics
X_5	Punch ng	Entertainment	www.*entertainmnt	Entertainment
X_6	Thisday news	Politics	www.*politics	Politics
X_7	Vanguard news	Sports	www.*sports	Sports
X_8	Complete football	Sports	www.*sports	Sports
X_9	Vanguard news	Politics	www.*politics	Politics
X_{10}	China daily	Politics	www.*politics	Politics
X_{11}	Thisday news	World	www.*world	World

Now for Pattern matching we have to match each attribute of test tuple with each attribute of training tuple. We can explain through this pattern matching table-

Table 2 – Training tuples

user	Daily's name	News category	Added required feed type	Pattern matching values
X_1	CNN news	World	www.*world	0
X_2	China daily	Business	www.*business	0
X_4	CNN news	Politics	www.*politics	2
X_5	Punch ng	Entertainment	www.*entertainmnt	1
X_6	Thisday news	Politics	www.*politics	2

X_7	Vanguard news	Sports	www.*sports	0
X_8	Complete football	Sports	www.*sports	0
X_9	Vanguard news	Politics	www.*politics	2
X_{10}	China daily	Politics	www.*politics	2
X_{11}	Thisday news	World	www.*world	0

Table 3 – Test tuples

User	Daily's name	News Category	Added required feed type	Class
X3	punch ng	Politics	www.*politics	?

After matching tuples we have a pattern matching values 0 ,1,2. Whereas a tuples of value 1 and 2 are relevant and tuples of value 0 are irrelevant. We will take all the tuples who have values 1 and 2 and discard all the tuples who have value 0.

Table 4- Values of pattern matching values

User	Daily's name	News category	Added required feed type	Pattern matching values
X_4	CNN news	Politics	www.*politics	2
X_5	Punch ng	Entertainment	www.*entertainmnt	1
X_6	Thisday news	Politics	www.*politics	2
X_{10}	China daily	Politics	www.*politics	2

After finding the pattern matching value we will calculate MD for each tuple in table 3. In table 3 we have only relevant tuples that are actually useful for finding the unknown class label.

Manhattan distance D_M= $|X_i\text{-}Y_i|$

Where,

X_i = Test tuple

Y_i = Training tuples

We calculate MD for user (X_4,X_3)

The MD between two tuples for instance training tuple X4 and test tuple X_3 ie.

X_4 = (x_{41}, x_{42}, x_{43}) and X_3 = (x_{31}, x_{32}, x_{33}) each with the following attributes as in Table 3.

X_4= (CNN news, Politics, www.*politics) and X_3 =(Punch ng, politics, www.*politics) will be:

If the values are the same then the difference is taken to be zero(0), otherwise, the difference is taken to be one(1). So, for ($x_{1,1}$ and $x_{4,1}$) ie. (CNN news and Punch ng), the difference is 1, for (x_{12} and x_{32}) ie. (Politics and Politics) the difference is 0, likewise for (x_{13} and x_{33}) ie., (www.*politics and www.*politics) the difference is 0 as well, therefore,

$$D_M(X_4,X_3) = |(X_{4,1}\text{-}X_{3,1})+(X_{4,2}\text{-}X_{3,2})+(X_{4,3}\text{-}X_{3,3})|$$
$$D_M(X_4,X_3) = |1+0+0|= 1$$

Repeating the whole process for all the available users produced a stream of data as in Table 5.

Table 5- Data showing distance to user X_3

User	Daily's name	News category	Added required feed type	Mahanttan distance D_M
X_4	CNN news	Politics	www.*politics	1
X_5	Punch ng	Entertainment	www.*entertainmnt	2
X_6	Thisday news	Politics	www.*politics	1
X_9	Vanguard news	Politics	www.*politics	1
X_{10}	China daily	Politics	www.*politics	1

Now, after calculating the Manhattan distance we calculate weight for the table 3 with the pattern matching value and Manhattan distance.

Weight= pval*1/D_m*0.5

Where,

pval= pattern matching value

D_m = mantahaantan distance

Weight(X4, X3) = 2*1/1*0.5

= 1.3333

Repeating the whole process for all the available users produced a stream of data as in Table 6.

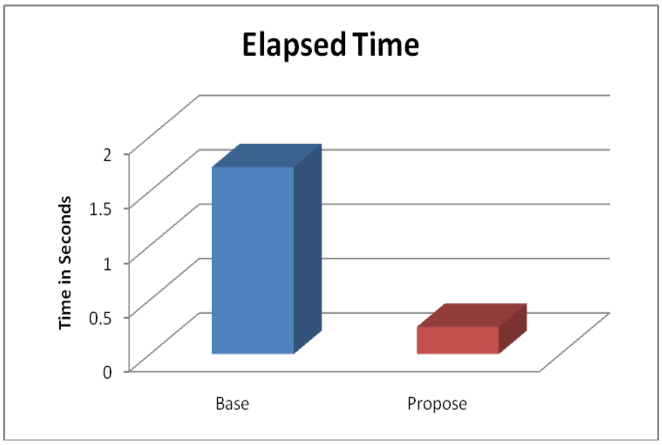
Table 6- Data showing weights to user X_3

User	Daily's name	News category	Added required feed type	weight
X4	CNN news	Politics	www.*politics	1.3333
X5	Punch ng	Entertainment	www.*entertainmnt	0.400
X6	Thisday news	Politics	www.*politics	1.3333
X9	Vanguard news	Politics	www.*politics	1.3333
X10	China daily	Politics	www.*politics	1.3333

After calculating weight we pick maximum weight or majority weight and assign that weight class as an unknown or missing class label. In this table we will pick "politics" class.

V. RESULT

The result analysis is the comparison between existing work and proposed work. Proposed work gives better, accurate class label results and it takes less computation time to calculate distance and predict unknown class label. The basic difference between existing and proposed work is Time. The elapsed time is 0.250038 seconds in IK-NN method. This is the total calculation time of the predicating unknown class label where is in previous work it takes 1.718486 seconds.



Graph 1- Comparison between proposed and based results in term of time

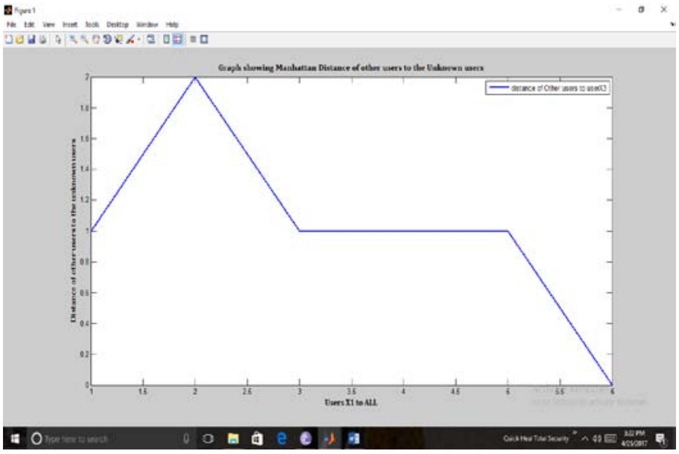


Figure 1
File Edit View Insert Tools Desktop Window Help
Graph showing Manhattan Distance of other users to the Unknown users
distance of Other users to userX3
Distance of other users to the unknown users
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Users X1 to ALL
Type here to search
Quick Heal Total Security
3:22 PM
4/25/2017

Graph 2- Graph showing MD from the other User/Neighbor to user X3.

VI. CONCLUSION AND FUTURE RECOMMENDATIONS

In this work, provides a basis for automatic Real-Time recommendation system. The system performs classification of users on the simulated active sessions extracted from testing sessions by collecting active users' click stream and matches this with similar class in the data, so as to produce a set of recommendations to the users in a Real-Time. In this work WUM is used to classify and discover the useful pattern for the relevant data and basically this approach uses MD which is used in K-NN algorithm that is used to compute the weight of the field which is matched with those of pattern matching fields. Here the value of pattern field distance based nearest neighbor approach is used. And at last we will calculate the weight of pattern matching field. Many research also need to be carried out on many other data mining techniques, comparing the result with this model, so as to determine the most effective model in handling a problem of this nature in the nearest future.

VII. REFERENCES

[1] Kratika Srivastava "Comparative Analysis of Web Mining Techniques: Survey" International journa l for res earch in ap pl i ed sc ienc e and engineering technolo gy (ijras et), Vol. 2 Issue V, May 2014 ISSN: 2321-9653.

[2] Dr.P.Tamijeselvy, Sangavi. S, Suvetha. T, Umashankari. T "Web usage mining using Improved KNN Algorithm "International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869 (O) 2454-4698 (P), Volume-4, Issue-2, February 2016".

[3] Harshit Patel, Prof. Pooja Jardosh "A Survey on Recommendation System Using Web Usage Mining" International Institution for Technological Research and Development Volume 1, Issue 1, 2015.

[4] Reema Sikka Amita Dhankhar Chaavi Rana "A Survey Paper on E-Learning Recommender System" International Journal of Computer Applications (0975 – 888) Volume 47– No.9, June 2012.

[5] Abhirami.K "Web Usage Mining using Fuzzy Association Rule" 978-1-4673-6725-7/16/$31.00 ©2016 IEEE.

[6] Suharjito, Diana and Herianto "Implementation of Classification Technique in Web Usage Mining of Banking Company" 2016 International Seminar on Intelligent Technology and Its Application, 978-1-5090-1709-6/16/$31.00 ©2016 IEEE.

[7] P. Sukumar, L. Robert and S. Yuvaraj "Review on Modern Data Preprocessing Techniques in Web Usage Mining (WUM)" 2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions, 978-1-5090-1022-6/16/$31.00 ©2016 IEEE.

[8] V.Anitha, Dr.P.Isakki @ Devi "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining" 978-1-4673-8437-7/16/$31.00 ©2016 IEEE.

[9] Doddegowda B J, G T Raju, Sunil Kumar S Manvi "Extraction of Behavioral Patterns from Preprocessed Web Usage Data for Web Personalization" IEEE 2016 International Conference On Recent Trends In Electronics Information Com

[10] Changqing Ji , Baofeng Wang , Shuai Tao , Junfeng Wu3 , Zumin Wang2 , Long Tang4 , Tiange Zu1 , Gui Zhao1 "Inverted Voronoi-based kNN Query Processing with MapReduce" 2324-9013/16 $31.00 © 2016 IEEE DOI 10.1109/TrustCom/BigDataSE/.

[11] D.A. Adeniyi, Z. Wei, Y. Yongquan "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method" Applied Computing and Informatics (2016) 12, 90–108.

.