WEB MINING USING K-MEANS CLUSTERING AND LATEST SUBSTRING ASSOCIATION RULE FOR E-COMMERCE

Rudra Prasad Chatterjee¹, Kaustuv Deb², Sonali Banerjee³,
Atanu Das⁴, Rajib Bag⁵

¹,²,³,⁵ Supreme Knowledge Foundation Group of Institutions, West Bengal, India.
⁴ Netaji Subhash Engineering College, Kolkata, India

¹rudrachatterjee.59@gmail.com, ²kaustuvdeb24@gmail.com,
³sonali.banerjee03@gmail.com, ⁴atanudas75@yahoo.co.in, ⁵rajib.bag@gmail.com

Corresponding Author: Atanu Das

https://doi.org/10.26782/jmcms.2019.12.00003

Abstract

User latency plays a significant role in e-commerce. This latency can be minimized by a priori predicting and fetching probable web pages for web users to run the e-commerce activities. Those prediction techniques are normally supported by clustering, classification and some association rules based on the data set of web logs of navigations, searching and attached web links with the e-commerce web pages. This paper proposes an integrated web page prediction technique by analyzing web users’ previous navigational behavior. K-means clustering and latest substring association rule are considered for developing the proposed method of e-commerce web page prediction. The proposed method is evaluated by analyzing the precisions values of the output clusters using the proposed prediction technique.

Keywords : Web page prediction, K-Means Clustering, Latest Substring Association Rule, Subsequence Association Rule, Substring Association Rule.

I. Introduction

In modern digital world, E-commerce web sites have become highly popular among web users. Web users rapidly visit many web pages of e-commerce sites in order to get much information about various products in miscellaneous domains and to buy products as per their needs. But, users find it very difficult to get fruitful and needed product information prone e-commerce web pages very often. Finding relevant and useful information regarding various products through the searching of huge number of e-commerce web pages not only demands enormous time but also appear as a tedious activity for the web users. Again, remembering the web addresses of the searched e-commerce web pages for future access also incurs great deal of problem for the users. In many cases, a visit to a single e-commerce web page brings
partial information regarding needed products for users. In these cases, an e-commerce web tour through a series of e-commerce web pages actually makes the entire product information available to the users. Also, retrieving of product information in a structural way often requires accessing of many e-commerce web pages in an ordered manner. Users sometimes eventually find a perfect orderly accessed e-commerce web tour. But they often find it difficult to trace the same tour path during future searching activities. Thus, a reference list of e-commerce web pages predicted to be visited by web users in future with proper access ordering could bring substantial benefits for web users. Web mining emphasizes on extracting knowledge from the web pages. Web mining can be classified into three distinct areas such as web structure mining, web content mining and web usage mining. Web structure mining aims at generating the structured summary about web sites in order to identify relevant documents. Web content mining involves efficiently extracting useful and relevant information from websites and databases. Web usage mining involves the analysis and discovery of users’ access patterns from web servers. Web page prediction is a key application of web usage mining where the next web page that may be visited by the web users is predicted. This prediction is done by observing web users’ previous web access patterns from the log records of the browsers through which web pages were accessed by the web users. A well developed sound web page prediction technique can build a reference list of e-commerce websites arranged in the order of their predicted future visits by web users. This list can help web users in finding their needed information with ease.

In this paper, an integrated web page prediction technique is developed by applying K-means clustering and Latest Substring Association (LSA) rule mining method. This paper also has portrayed the comparisons between three methods, such as subsequence association rule, substring association rule and LSA. The rest of the paper is organized as follows. The following section presents the background literature associated with the present work. The immediate next section presents the web mining methodology under consideration. The results of the empirical investigations are presented in the section next to the section portraying the web mining methodology. Finally, this paper ends with the conclusion and the future prospects of the present work.

II. BACKGROUND

Web page prediction has long history since the inception of www and designing required static web pages under a website beforehand can be treated as its first course example. This field gained its attractiveness with the evolution and popularity of e-commerce after 2000. A model has been proposed in [XIV] for secure web site navigability through web mining where classification techniques has been used for mining data. [X] proposes a novel Page Rank based clustering (PRC) algorithm which produces graph partitioning with high modularity and coverage and it uses the hypertext structure. [XII] suggests a new technique Public Key Infrastructure (PKI) which uses digital certificate to identify the validity of web users. [XV] used K-means clustering, Markov model and page ranking algorithm for webpage prediction where as [XIX] proposes a web page prediction model giving significant importance to the user’s interest using the apriori algorithm and the user’s navigational behavior
through latest substring association rule. A composite web page prediction model giving significant importance to the user’s interest has been implemented using the clustering technique and the Markov model has been introduced by [XVIII] where cellular automata has been used for storing the predicted web pages which makes this model more memory efficient. [IV] proposes a new methodology of process mining Map Miner Method (MMM) which can be used to determine user’s interest at specific web pages.

[XI] suggest the usage of web data in various E-commerce sites to predict web user’s interest on a specific web sites. [XVIII] works on server logs from the MSNBC data set and it aims at predicting subsequent web page using Apriori Prefix Tree (APT) algorithm. There are various web mining techniques which are used to identify the most frequent web data such as association rules, low order and higher order Markov models etc. A new web usage mining techniques has been proposed by [I] where pairwise nearest neighbor (PNN) based clustering and then sequential pattern mining are combined to identify patterns of next page accesses. Markov model is used for pattern mining and at the end of that paper comparison of traditional Markov model and proposed method has been depicted. A new privacy preservation association rule mining algorithm has been proposed by [VII] which has been implemented in secure web mining. [XV] discussed several methods of clustering and artificial neural network (ANN) which has been used for web access prediction model. [VI] proposes a privacy preserving association rule mining algorithm which provides a secure web mining and complexity of this algorithm has also been evaluated in the analysis part. [III] focused in an efficient outlier detection method on transactional datasets, namely, EFPOR, where outliers are detected based on frequent patterns of item set within transactions and the performance efficiency of the algorithm has been verified at the end of this paper which provides better accuracy. [V] proposes a model for an E-learning system which is based on semantic web and E-learning agents. Two Monte Carlo algorithms have been proposed by [II] which has been used to build new queries based on relevant documents of the previous query and the proposed algorithm has been compared with another approach based on cosine measure. [IX] characterized Markov model based prediction where the proposed system has been used to predict the next webpage access of a user and four different types of Markov models, namely, Support-Pruned Markov Model, Confidence-Pruned Markov Model, Frequency-Pruned Markov Model and Error-Pruned Markov Model have been compared. Several parameters such as accuracy, coverage, F1-Measure have been evaluated to understand the effectiveness of the proposed schemes. In the work of [XIII] for predicting user access behavior various web mining techniques have been used and also cloud computing has been employed with it.Two different clustering techniques, namely, fuzzy c-means (FCM) clustering and FLAME clustering algorithms, were explored for prediction task where the performance of FLAME clustering algorithm [XVII] was better than that of fuzzy C-means, fuzzy K-means algorithms and fuzzy self-organizing maps (SOM) and the user browsing time also improved without compromising prediction accuracy.[XXI] represents a technique for information similarity analysis using clustering and feature vector techniques has been used to identify the similarity of the functional text.
III. Methodology

Preprocessing

The first step for web page prediction is to clean the raw data of the web usage log (record or file). Hence, the documents that are not directly requested by users have been filtered manually. The steps of preprocessing of data are given below:

1. Concept dictionary generation: Initially, concept dictionary was implemented where each term has been specified under their corresponding domain and a unique term ID has been generated for each term.
2. Frequency matrix generation: In this step, the term counts were calculated for each web page by observing that how many times each term is accessed in that web page for a particular web session by the web users.
3. Occurrence matrix generation: In occurrence matrix all the terms under a particular domain for each web page has been observed which can be used as an input file for the clustering method.

K-Means clustering 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 observations into groups of related observations without any prior knowledge. It is used to divide a set of observations into k groups of clusters, where k is an input parameter. Then it randomly selects k points as the center of clusters. Recalculate the center or mean by assigning observation to their closest cluster center by using Euclidean distance function. This iterative approach of clustering algorithm improves the revised center of clusters. Iteration remains persistent until the clusters converge.

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 center.
  (ii) Repeat
        {
            Re-assign 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 by calculating the mean value of the new assigned objects for each cluster.
        }
     } until no change.
}
Subsequence Association Rule

A subsequence can be generated from a series of URL that appear in the similar sequential order as they were accessed in the web log. This type of association rule maintains only the order of the web session information. Hence this rule is very less stricter than the other association rule and as a result it considers too many rules when dataset becomes large. A simple example of subsequence association rule has been depicted in Table 1.

| Antecedent Window | Consequent Window | Association Rule |
|-------------------|-------------------|------------------|
| P,Q,R             | S                 | {Q,R}->{P}, {P,R}->{S}, |
| Q,P,R             | S                 | {P}->{Q}, {Q}->{R}->{S} |

In this example the existing web pages of the antecedent window for both web sessions violates the order in which they had visited by web users. Hence rule {P,Q}->{S} is rejected according to subsequence association rule.

Substring Association Rule

Substrings are any combination of adjacent URL’s of the antecedent window. This rule maintains the order and adjacency of the web session information. Hence this rule is little bit stricter than subsequence association rule. The left hand side of the rule must encode string information at the time of rule extraction from the log tables. A brief explanation about this rule has been reflected in Table 2.

| Antecedent Window | Consequent Window | Association Rule |
|-------------------|-------------------|------------------|
| P,Q,R             | S                 | {P}->{Q}, {Q}->{R}->{S} |
| Q,P,R             | S                 | {P}->{Q}, {Q}->{R}->{S} |

The generated rules from Table 2 maintains not only the order but also the adjacency of web pages. In first web session web pages Q and R were adjacent where as web pages P and R become adjacent in next web session. Hence the rules {Q,R}->{S} and {P,R}->{S} has been rejected in this type of association rule.

Latest Substring Association Rule (LSA)

Association rule mining [XIX] is a major pattern discovery technique. Association rule discovery on usage data results in finding group of items or pages that are commonly accessed together. The final step in the training process is to generate the association rules from the original data. Association rules are mainly defined by two
metrics: support and confidence. Where ‘support’ is the measurement that how many times a particular web page visited in a specific web session.

\[
\text{confidence} = \frac{\text{support}(\text{L.H.S} \cap \text{R.H.S})}{\text{support}(\text{L.H.S})}
\]  

(1)

If there is a rule containing L.H.S→ R.H.S then we can easily calculate the confidence threshold value using formula (1). After that we will concentrate on implementing the LSA which is given below.

**Table 3:** Example of LSA

| Antecedent Window | Consequent Window | Association Rule |
|-------------------|-------------------|------------------|
| P,Q,R             | S                 | \{R\}→S          |
| Q,P,R             | S                 |                  |

In LSA not only the order web two windows (Antecedent and Consequent) maintains but also adjacency and recency of the web pages of both windows has been maintained. In our example three web pages (P,Q and R) has been visited in both web session by web user. The order of the web page P and Q visited in two web session changes but web page R has visited as a latest page in both web session. Hence web page R will be selected as a L.H.S of the association rule. Similarly in both web session web page S exist in Consequent Window as a result web page S will be the R.H.S of the rule. Therefore only one rule can be considered according to LSA which has been illustrated in Table 3.Hence it can be conclude that LSA is more stricter than Subsequence and Substring Association rule.

**Proposed K-Means Latest Substring Association Algorithm**

The proposed algorithm is developed by amalgamating K-Means clustering algorithm and Latest Substring Association Rule for the purpose of web page prediction for E-commerce web sites. Steps of the proposed algorithm are stated below-

1. At first a certain amount of web sessions from the web log containing previously visited web pages are collected.

2. After that the web sessions are preprocessed i.e. in this stage concept dictionary, frequency matrix and occurrence matrix has been generated.

3. At the next stage web pages are clustered using K-means clustering algorithm to get the web pages cluster wise.

4. Calculate support of each web page from the web session data i.e. how many times that particular web page has been visited by the web users.

5. After that calculate confidence value of each rules generated for each cluster individually where

\[
\text{confidence} = \frac{\text{sup}(\text{L.H.S} \cap \text{R.H.S})}{\text{sup}(\text{L.H.S})}
\]

(2)
Step 6. Identify the rules which contains minimum confidence threshold and add them to the set of strong rules.

Step 7. Repeat step 4 to 6 until all the cluster members are completely traversed.

Step 8. Finally calculate Precision value for each cluster.

\[
\text{Precision} = \frac{C}{N}
\]

Where \( C = \) Number of correct rules, \( N = \) Total number of generated rules in a particular cluster.

Step 9. End.

In this paper, the final steps which provides the predicted pages as a result consists of some steps where each step has its own vitality for manipulating that particular page. Among them, the major steps are- K-means clustering and latest substring association rule mining.

IV. Experimental Analysis

In experimental analysis, a sample server preprocessed log file consisting of a sequence of web pages visited by web users is taken as input data. Based on this input file, a unique id has been created for each web page URL in the following way-

\begin{align*}
\text{WS1} & \rightarrow \{\text{W1, W6, W8}\} \\
\text{WS2} & \rightarrow \{\text{W4, W7, W8, W10}\} \\
\text{WS3} & \rightarrow \{\text{W5, W8, W2}\} \\
\text{WS4} & \rightarrow \{\text{W1, W3, W9}\} \\
\text{WS5} & \rightarrow \{\text{W2, W5, W8, W10}\} \\
\text{WS6} & \rightarrow \{\text{W4, W6, W9}\} \\
\text{WS7} & \rightarrow \{\text{W1, W2, W3, W10}\} \\
\text{WS8} & \rightarrow \{\text{W4, W7, W2, W5}\}
\end{align*}

It can be seen from the above arrangement that 8 web sessions (WS1, WS2, ..., WS8) have been taken as a sample input data where the web pages visited (having unique ids like: W1, W2, ..., W10) in a particular web session have also been depicted. As an example, in the web session WS1, three web pages have been visited in the order W1, W6 and W8. After completing the preprocessing of web log data, only some terms such as software, database under the domain of CSE are considered in distinct web pages and concept dictionary has been developed. Table 4 shows the concept dictionary.
By using concept dictionary we can easily identify each term with their associated term ID and now to implement the frequency matrix all terms are traversed of a particular web page. The term count for each web page was calculated by observing that how many times each term is accessed in that web page for a particular web session depicted in Table 5.

By using concept dictionary we can easily identify each term with their associated term ID and now to implement the frequency matrix all terms are traversed of a particular web page. The term count for each web page was calculated by observing that how many times each term is accessed in that web page for a particular web session depicted in Table 5.

### Table 4: Concept Dictionary

| Term ID | Domain1: CSE | Term ID | Domain2: ME | Term ID | Domain 3: EE |
|---------|--------------|---------|-------------|---------|-------------|
| T1      | OS           | T6      | Spring      | T11     | Current     |
| T2      | Software     | T7      | Stress      | T12     | Circuit     |
| T3      | JAVA         | T8      | Torque      | T13     | Amplifier   |
| T4      | C            | T9      | Construction| T14     | Modulation  |
| T5      | Database     | T10     | Hydraulics  | T15     | Controller  |

### Table 5: Frequency Matrix

|       | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 | T11 | T12 | T13 | T14 | T15 |
|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|
| W1    | 2  | 1  | 2  |    |    | 1  |    |    |    |     |     |     |     |     |     |
| W2    |    |    | 1  | 1  |    |    |    |    |    |     |     |     |     |     |     |
| W3    | 2  | 2  |    |    |    |    |    |    |    |     |     |     |     |     |     |
| W4    | 3  | 1  |    |    |    |    | 2  | 3  | 1  |     |     |     |     |     |     |
| W5    |    |    |    |    |    |    |    |    | 2  | 1  | 1  | 1  |     |     |     |
| W6    | 2  | 1  | 2  | 1  |    |    |    |    |    |     |     |     |     |     |     |
| W7    |    | 1  | 2  | 2  | 1  |    |    |    |    |     |     |     |     |     |     |
| W8    |    |    |    |    |    |    |    | 2  | 2  |    |    | 2  | 1  |     |     |
| W9    | 2  | 1  | 2  |    |    |    |    |    |    | 1   |     |     |     |     |     |
| W10   | 1  | 1  |    | 1  | 1  | 1  |    | 4  | 2  |     | 1   |     |     |     |     |

After implementing the frequency matrix we are concentrating to design the occurrence matrix where all the terms under a particular domain for each web page has been observed such as web page W1 provides 5 terms related to CSE domain, 3 terms related to ME domain but terms related to EE is null. Thus the entire occurrence matrix has been generated which shows the web pages provide number of terms with their associated domain. The generated occurrence matrix are presented in the following Table 6.
Table 6: Occurrence Matrix

| Webpage | Term-CSE | Term-ME | Term-EE |
|---------|----------|---------|---------|
| W1      | 5        | 3       | 0       |
| W2      | 2        | 0       | 7       |
| W3      | 4        | 0       | 0       |
| W4      | 4        | 6       | 0       |
| W5      | 0        | 0       | 5       |
| W6      | 6        | 0       | 0       |
| W7      | 0        | 6       | 0       |
| W8      | 0        | 4       | 3       |
| W9      | 5        | 0       | 2       |
| W10     | 2        | 3       | 7       |

Table 7: Case (Webpage) Wise Cluster Membership

| Case Number | Cluster | Case Number | Cluster |
|-------------|---------|-------------|---------|
| 1           | 3       | 6           | 3       |
| 2           | 2       | 7           | 1       |
| 3           | 3       | 8           | 1       |
| 4           | 1       | 9           | 3       |
| 5           | 2       | 10          | 2       |

Table 8: Clusters with Web Page Membership Information

| Cluster No. | Web Pages   |
|-------------|-------------|
| C1          | W4,W7,W8    |
| C2          | W2,W5,W10   |
| C3          | W1,W3,W6,W9 |
| Sl No. | Cluster No. | Rule  | Confidence |
|--------|-------------|-------|------------|
| 1      | C1          | W4→W7 | 66.66%     |
| 2      | C1          | W4→W8 | 33.33%     |
| 3      | C1          | W7→W4 | 66.66%     |
| 4      | C1          | W7→W8 | 33.33%     |
| 5      | C1          | W8→W4 | 25%        |
| 6      | C1          | W8→W7 | 25%        |
| 7      | C1          | W4, W7→W8 | 50% |
| 8      | C2          | W2→W5 | 75%        |
| 9      | C2          | W2→W10 | 50%   |
| 10     | C2          | W5→W2 | 100%       |
| 11     | C2          | W5→W10 | 33.33% |
| 12     | C2          | W10→W2 | 66.66% |
| 13     | C2          | W10→W5 | 33.33% |
| 14     | C2          | W2, W5→W10 | 33.33% |
| 15     | C3          | W1→W3 | 66.66%     |
| 16     | C3          | W1→W6 | 33.33%     |
| 17     | C3          | W1→W9 | 33.33%     |
| 18     | C3          | W3→W1 | 100%       |
| 19     | C3          | W3→W6 | 0%         |
| 20     | C3          | W3→W9 | 50%        |
| 21     | C3          | W6→W1 | 50%        |
| 22     | C3          | W6→W3 | 0%         |
| 23     | C3          | W6→W9 | 50%        |
| 24     | C3          | W9→W1 | 50%        |
| 25     | C3          | W9→W3 | 50%        |
| 26     | C3          | W9→W6 | 50%        |
| 27     | C3          | W1, W3→W9 | 50%   |
In this paper after generation of occurrence matrix web pages are clustered using K-means clustering algorithms, for getting group of web-pages cluster wise, which have been used for next web page prediction purpose which is depicted in the Table 7.

Here case numbers are web pages and cluster represents which web page belongs to which cluster number. For example, web pages \{W1, W3, W6, W9\} belongs to cluster C3 where as web pages \{W2, W5, W10\} is in cluster C2 and so on. The final clustering result is shown in Table 8.

From the clustered results, we can easily distinguish between the web pages. In this paper, after performing K-means clustering several association rules has been compared according to their performances. Experiments were performed using subsequence association rule, substring association rule and LSA. From all the generated rules, those rules which satisfy the minimum confidence threshold (in our example which is 60%) are considered as strong rules. By using those strong rules web page predictions have been performed such that the L.H.S of the rule has been considered as visited web page where as R.H.S of the rule as next web page.

The result of generated association rules along with their corresponding confidence values are illustrated in the Table 9,Table 10 and Table 11 respectively.

**Table 10: Rule Generation Using Substring Association Rule**

| Sl No. | Cluster No. | Rule          | Confidence |
|-------|-------------|---------------|------------|
| 1     | C1          | W4 -> W7     | 66.66%     |
| 2     | C1          | W7 -> W8     | 33.33%     |
| 3     | C2          | W2 -> W5     | 75%        |
| 4     | C2          | W5 -> W10    | 33.33%     |
| 5     | C3          | W1 -> W3     | 66.66%     |
| 6     | C3          | W3 -> W9     | 50%        |

**Table 11: Rule Generation Using Latest Substring Association Rule (LSA)**

| Sl No. | Cluster No. | Rule   | Confidence |
|--------|-------------|--------|------------|
| 1      | C1          | W4 -> W7 | 66.66%    |
| 2      | C2          | W2 -> W5 | 75%       |
| 3      | C3          | W1 -> W3 | 66.66%    |
The result of generated strong rules (which contain minimum confidence threshold equal to 60%) of all above mentioned association rule has been depicted in Table 12, Table 13 and Table 14 respectively.

**Table 12:** Web Page Prediction Using Subsequence Association Rule

| Current Web Page | Next Web Page | Probability |
|------------------|---------------|-------------|
| 4                | 7             | 66.66%      |
| 7                | 4             | 66.66%      |
| 2                | 5             | 75%         |
| 5                | 2             | 100%        |
| 10               | 2             | 66.66%      |
| 1                | 3             | 66.66%      |
| 3                | 1             | 100%        |

**Table 13:** Web Page Prediction Using Substring Association Rule

| Current Web Page | Next Web Page | Probability |
|------------------|---------------|-------------|
| 4                | 7             | 66.66%      |
| 2                | 5             | 75%         |
| 1                | 3             | 66.66%      |

**Table 14:** Web Page Prediction Using LSA

| Current Web Page | Next Web Page | Probability |
|------------------|---------------|-------------|
| 4                | 7             | 66.66%      |
| 2                | 5             | 75%         |
| 1                | 3             | 66.66%      |

From the above table, it can easily be observed that the number of generated association rules in case of Subsequence and Substring association rule are too large than LSA but the number of strong association rules are very less in case of both substring and LSA rather than Subsequence Association rule mining methods. Hence, we have calculated the accuracy value for all the association rules as per equation (4)-
Accuracy value for all above mentioned association rules has been evaluated in Table 15 using (4).

Table15: Accuracy Value of Association Rule

| Association rule | Total number of association rules | Total number of strong association rules | Accuracy     |
|------------------|----------------------------------|-----------------------------------------|--------------|
| SUBSEQUENCE      | 27                               | 7                                       | (7/27)*100% = 25.92% |
| SUBSTRING        | 6                                | 3                                       | (3/6)*100% = 50%     |
| LSA              | 3                                | 3                                       | (3/3)*100% = 100%    |

Fig 1. Accuracy Value of the Association Rule

Table 16: Cluster Wise Precision Value

| Cluster | Precision |
|---------|-----------|
| C1      | 33.33%    |
| C2      | 50%       |
| C3      | 16.66%    |

Copyright reserved © J. Mech. Cont.& Math. Sci.  
Rudra Prasad Chatterjee et al
From Fig 1 it is proved that accuracy value of LSA is quite better than the other two association rules because it not only consider order and adjacency but also recency of the web session information. From Table 15 we can reach in a decision that LSA produces much better than the others and as a result for prediction task LSA has been combined with K-means in our research work. After that the precision value corresponding to their cluster number has been evaluated according to LSA shown in the Table 16 using (3). The experimental result of Table 16 shows that the precision value is highest for cluster number C2 which consist of web pages {W2, W5, W10}. From that result we can easily determined that all the web pages under cluster C2 should get the highest rank among all the web pages in our experiment and that web pages have to be synchronized according to their rules. The experimental results are illustrated in Fig 2. Precision value with their corresponding cluster numbers are depicted in the Fig 2.

![Fig 2. Precision Value With Respect To Cluster Numbers](image)

From the above Fig 2, important clusters can be identified based on their precision value which is cluster C2 in our experiment. The precision of cluster C3 is 50% which suggests that the web page under cluster C2 is more important for the requirement of next web users.

V. Conclusion

This paper has portrayed a new method of predicting e-commerce web pages by integrating two powerful and well known techniques such as LSA rule mining method and K-means clustering algorithm. This line of integration may open a new direction in e-commerce web prediction to reduce users’ latency. It may be noted that the web usage log records are required to be refined manually or filtered before actual application of the proposed method. E-commerce users’ access patterns are explored by the proposed method for predicting the e-commerce web pages with best future visit possibility. Higher precision accuracies presented in the investigation results demonstrated the effectiveness of the proposed method in the present purpose.
Results produced by K-means clustering have shown justified belongingness of all e-commerce web pages taken in experiment in three clusters. This clustering activity is performed based on the distance grouping concept. No discrepancy is found in the clustering process. In the next step, subsequence association rule mining method is applied on the clusters produced by the K-means clustering method. Subsequence association rule mining method has produced seven strong association rules for web page prediction. These seven rules have high (greater than 60%) probability of occurrence out of twenty seven association rules with an accuracy of 25.92%. In comparison to this, subsequence and substring association rule mining method has produced three strong association rules. These three rules have high (greater than 60%) probability of occurrence out of six association rules with a better accuracy of 50%. Finally, the LSA rule mining method has successfully generated three strong association rules having high (greater than 60%) probability of occurrence out of three association rules. Thus, LSA method has achieved the highest accuracy. These prediction results will help the e-commerce web users to find their required and most frequently visited e-commerce web pages very quickly. This outcome will eventually help them to get legitimate and required product and transaction information more easily. This in turn enhances all the e-commerce activities. Proposed method is tested on the log records of ten e-commerce web pages and has produced results significant to be taken in to considerations. Use of more sophisticated machine learning techniques with the proposed method and testing on more log records of many other types of web pages may produce few other insights of the proposed method.

References

I. A. Anitha , “A New Web Usage Mining Approach for Next Page Access Prediction”, Int. Journal of Computer Applications, 8 (11), 7-10, 2010.

II. C. Gutierrez-Soto, “On The Reuse of Past Searches in Information Retrieval: Study of two Probabilistic Algorithm”, Int. Journal of Information System Modeling and Design. 6 (2). 72-92, 2015.

III. C. Ray, and P. Das, “An Outlier Detection Based on Frequent Pattern”, Int. Journal of Engineering Research & Technology (IJERT), 2(11), 2810-2814, 2013.

IV. G. Khodabandelou, C. Hug, and C. Salinesi, “Mining User’s Intents from logs”, Int. Journal of Information System Modeling and Design, 6(2), 43-71, 2015.
V. G. Shrivastava, K. Sharma, and A. Bawankan, “A new framework semantic web technology based e-learning”, In Environment and Electrical Engineering (EEEIC), 2012 11th International Conference on (pp. 1017-1021).

VI. G. Shrivastava, and V. Bhatnagar, “Secure Association Rule Mining for Distributed Level Hierarchy in Web”, International Journal on Computer Science and Engineering, 3(6), 2240-2244, 2011.

VII. G. Shrivastava, and V. Bhatnagar, “Analyses of Algorithms and Complexity for Secure Association Rule Mining of Distributed Level Hierarchy in Web”, International Journal of Advanced Research in Computer Science, 2(4), 2011.

VIII. I. Chhabra, G. Suri, “Data Science and Knowledge Discovery through Data Mining Paradigms”, Journal of Mechanics of Continua and Mathematical Sciences, 14(2), 167-173, 2019

IX. J. Chembath, and T. Fredrik, “Ensemble System Using Hybrid Next Web Page Prediction Algorithm Combining Clustering, Markov Model With Associative and Longest Common Subsequence Classifiers”, Int. Journal of Pure and Applied Mathematics, 119(12), 5329-15339, 2018.

X. K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S. K. Pham and E. Smirnova, “PageRank based clustering of hypertext document collections”, In Proc. of the 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 873-874, ACM, 2008.

XI. K. Sharma, G. Shrivastava, and D. Singh, “Risk Impact of Electronics-Commerce Mining: A Technical Overview”, International Conference on Computing Communication and Information Technology, pp 96-100, 2012.

XII. K. Sharma and G. Shrivastava, “Public Key Infrastructure and Trust of Web Based Knowledge Discovery”, International Journal of Engineering, Sciences and Management, 4(1), 56-60, 2011.

XIII. K. Sharma, G. Shrivastava, and V. Kumar, “Web mining: Today and tomorrow”. Electronics Computer Technology (ICECT), 2011 3rd International Conference on (Vol. 1, pp. 399-403).

XIV. K. Sharma, and V. Bhatnagar, “Private and Secure Hyperlink Navigability Assessment in Web Mining Information System”, International Journal on Computer Science and Engineering, 3(6), 2245-2250, 2011.

XV. O. P. Mandal, and H. K. Azad, “Web Access Prediction Model Using Clustering and Artificial Neural Network”, Int. Journal of Engineering Research and Technology, 3(9), 195-199, 2014.

XVI. P. Bhart, D. Shukla, D. S. Saini and A. Gaur, “Composite Model of Web Page Prediction Using Page Ranking Algorithm and Markov Model”, 4th Int. Conf. on Recent development in Engineering Science, 125-13, 2017.
XVII. P. Sampath and M. Prabhavathy, “Web Page Access Prediction using Fuzzy Clustering by Local Approximation Memberships (FLAME) Algorithm”, ARPN Journal of Engineering and Applied Sciences, 10(7), 3217-3220, 2015.

XVIII. R. Geetharamani, P. Revathy and S. G. Jacob, “Prediction of Users Webpage Access Behaviour Using Association Rule Mining”, Sadhana, 40(8), 2353-2365, 2015.

XIX. R. P. Chatterjee, M. Ghosh, M. K. Das, and R. Bag, “Webpage Prediction Using Latest Substring Association Rule Mining”, In Foundations and Frontiers in Computer, Communication and Electrical Engineering: Proc. of the 3rd Int. Conf. C2E2, Mankundu, West Bengal, India, 15th-16th January, 2016. (p. 157), CRC Press.

XX. Sunita, V. Rana, “Improved probable clustering based on data dissemination for retrieval of web URLs”, Journal of Mechanics of Continua and Mathematical Sciences, 14(5), 285-294, 2019.