Novel Recommender System Design using Supervised and Unsupervised Techniques

Sherica Lavinia Menezes
Goa University
Department of Computer Engineering
Goa College of Engineering

Geeta Varkey
Goa University
Department of Computer Engineering
Goa College of Engineering

ABSTRACT
Recommender systems have been designed using association rule mining. However the rule generation complexity of ARM proves to be disadvantageous when dealing with huge amounts of data. Taking this disadvantage into consideration this paper proposes predicting missing items using associative classification techniques. To accomplish this task either a classifier or a clustering approach is chosen. This paper proposes classifying the items prior to prediction process using Naïve Bayes Classifier or hierarchical clustering approach. The advantages of the proffered approach are that the complexity of rule generation is lowered to a great extent and the prediction is done at a higher level of abstraction. The prediction algorithm chosen is associative classification mining using ComboMatrix. The classifier or the clustering mechanism maps huge datasets to a set of classes the size of which in most classes is smaller than the size of the dataset. Therefore this approach greatly reduces the size of the dataset and the overall complexity. This paper lists out the literature survey carried out in the field and the design of the proposed system. The experiment carried out shows that the performance and memory requirements of the proposed approach are more efficient than the method using only associative classification mining.

General Terms
Association Rule Mining, Machine Learning, Web Mining.

Keywords
ComboMatrix, Graph based prediction, Hierarchical Clustering, Naïve Bayes classifier, Recommender systems.

1. INTRODUCTION
The The World Wide Web serves as a huge repository of data and information. The process of mining this data to extract some important and relevant features or patterns is known as web mining. One of the areas in web mining is establishing association rules which help in predicting nature of transaction based on the information available in the history of the transactions carried out. The major challenge is given history of transactions; the system predicts any item that a user is missing in the current transaction. Such systems are known as recommender systems and they find relevant application in the field of market basket analysis.

Association rule mining finds its application in many areas one of them is the market basket analysis. While shopping, if users have bought bread, butter, milk and cheese together many a times then such analysis indicate that if the shopping mall keeps these items together in the display then it would simplify the user's purchase and also help the shopping mall. This idea can be extended to a recommender system; wherein given the history and nature of transactions the system predicts if the user is missing any item in the current transaction and suggests that item to the user. Such system is of great help to the user since it enlists the missing items and thereby helps the user shop effectively. The system also helps the vendors by listing the missing items and encouraging the user to purchase certain items that were overlooked by the user.

The major drawback of traditional association rule mining is the complexity of rule generation. Some of the major algorithms developed in the field of association rule mining are Apriori algorithm, FP tree and Itemset Tree. All these algorithms work at the level of individual items and the number of rules generated are very high. Application of these techniques on the data available on the Web causes a huge set of rules to be generated. Another aspect is deciding which items have to be considered thus listing a need of a threshold value. It is preferred that the items appearing most number of times should be considered. This threshold value is called the support of the itemset. In addition to listing which items are to be considered, it is necessary to list only the rule which provide relevant amount of information, for which a threshold value called the confidence level is used.

This paper proposes a different approach to the design of a recommender system. This proposed approach first aims at classifying the individual items in a transaction, perform prediction on the classified set of items and suggest a category of items to the users. To accomplish this, the paper proposes either Naïve Bayes classifier which will classify the items into distinct classes or Hierarchical Clustering mechanism can be employed to cluster the items for which no labels can be assigned. Once the items are classified or clustered the proposed approach constructs a graph from the classes/clusters based on the appearance of the classes/clusters in the transaction history. The graph is used for predicting the missing items from the user's current transaction. This approach proves helpful since it reduces the complexity of rule generation. Another significant advantage is that data is mined at a higher level of abstraction.

In the following sections this paper lists out the literature survey carried out, the problem statement and the existing approaches available in the selected research area. Later the design of the proposed system is elaborated and discussed.

2. LITERATURE SURVEY
In the field of Web mining, a lot of research is currently being carried out. This research is further divided in the various sub fields of Web mining. [3] discusses the current research trends in the field of web content mining. According to this survey there are 2 main approaches towards web content mining: agent – based approach and database based approach. The three types of agents that are commonly used in this area are: intelligent search agents, information filtering/categorizing agents and personalized web agents. Intelligent search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents use a number of techniques to filter data according to the predefine instructions.

Personalized web agents learn user preferences and discovers documents related to those user profiles. In the database approach it consists of well formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi-structured and multimedia data.

[1] discusses predicting missing items in shopping cart using itemset trees. In this approach first an itemset tree is constructed from the data available and then the uncertainty processing is done using Bayesian approaches and an algorithm based on the Dempster – Shafer theory. This approach is proved to be faster than the approaches used prior to it. [2] further improves predicting missing items using fast algorithms. The paper uses a matrix based fast algorithm for association rule mining. The paper compares the approach using fast algorithms and approach using itemset trees and proves that fast algorithms work better. However the use of a 2 dimensional array affects the scalability of the algorithm. As the data increases exponentially the complexity of the algorithm also increases remarkably.

Another approach is recommender systems in web mining is the approach based on collaborative filtering [10]. In the approach of collaborative filtering, the filtering and evaluating of items is done based on the opinions of other people. This technology involves the opinions generated by the huge communities present on the Web and their opinions are used to filter and process the data. Among the various approaches developed to provide personalized recommendations to users, ratings-based collaborative filtering recommenders constitute an important category. This popular technique however is dependent on the explicit ratings provided by users to various items. If this user-item rating matrix is sparse, picking out mentors for a target user becomes difficult. As a result the quality of recommendation is affected. [11] proposes an entropy based approach for designing recommender system using collaborative filtering.

In the area of web mining, one important issue is regarding the extraction of data from the websites. Since the data available in the websites does not follow any particular structure, the extraction of such unstructured data poses another challenge. [4], [5], [6] and [7] discuss the various techniques used in extracting data from websites. The data is extracted using approaches like the schema guided approach [4]. Genetic algorithms and regular expressions have also been used to extract data from the web [5]. Another approach used is the partial tree alignment approach [6]. Other techniques include, addressing elements in the document tree, tree edit distance matching algorithms, logic based approach, machine learning approaches, hybrid systems and template based matching [7]. In addition to these techniques there are tools available which automatically extract data from the websites. [8] presents a survey on the various categories that the web data extraction tools can be classified as and also mentions some tools in each of these categories.

3. PROBLEM STATEMENT

The main objective of this paper is to predict missing samples in a given transaction and suggest missing items to the user. The problem of designing the prognostic system can be defined as follows: Given any website, and a set of user transactions, the site suggests any item that would prove to be interesting to the user, based on previous user transaction data. This task is done at a higher level of abstraction. To achieve this, Naïve Bayes text classifier is employed to classify incoming transaction items. The Naïve Bayes text classifier proves to be simple yet efficient. Once the classifier accomplishes its task, a graph is constructed. The classes form the nodes of the graph and the edge values of the graph are updated using information from transaction history. The data structure chosen to store the graph is a ComboMatrix. ComboMatrix is a variation of an adjacency matrix. ComboMatrix is used for associative classification mining as proposed in [13].

The problem statement can be formally formulated as follows: Given any transaction T, consisting items {I1,I2,….In}; T = {I1, I2,……., In}, and a transaction history, following steps are taken:

a. For each Ij belonging to T, classify Ij into its respective class Ck, where Ck is one of the classes in the training set. Let this transaction be called as a classified transaction Tc= {C1,C2…..Cm}.

b. Predict the missing class in the classified transaction Tc using the graph based approach.

For data items that cannot be assigned a predefined label, hierarchical document clustering approach is suggested. The prediction should be of high quality and the system should be able to deal with a huge amount of data.

4. EXISTING APPROACHES

Existing approaches in this area use itemset trees and fast algorithms. These approaches employ association rule mining techniques. The first approach uses Itemset Trees [1] to establish the association rules between the items. The uncertainty in the occurrence is measured using Bayesian techniques and Dempster-Shafer theory. In this method the authors generate all high support and high confidence rules using itemset trees. Then consequents of all these rules are combined to give an estimated completion of shopping cart. This technique proves to appear better than the traditional techniques in association rule mining. However the disadvantage of this approach is that the rule generation complexity increases greatly with the increase of the average length of the transaction and with the number of distinct items.

Yet another method to predict missing items uses Boolean vector and the relational AND operation to discover frequent itemsets [2] without generating candidate items and generate the association rule. Association rules are used to identify relationships among a set of items in database. Initially Boolean Matrix is generated by transforming the database into Boolean values. The frequent itemsets are generated from the Boolean matrix. The association rules generated form the basis for prediction. The incoming itemset i.e the content of incoming shopping cart will also be represented by a Boolean vector and AND operation is performed with each transaction vector to generate the association rules. Finally the rules are combined to get the predictions. The advantages of this technique are that it doesn't generate candidate itemsets, it uses only a single pass over the database, the memory consumption is low and the processing speed is more as compared to the previous technique. The disadvantage of this approach is that the use of a Boolean matrix cannot handle huge amount of data. This disadvantage restricts the use of this technique in online applications where huge amount of data is generated.

The major drawback of the above mentioned technique is the rule generation complexity. Generating rules from a huge amount of data involves a lot of high memory and time complexity. In DS – ARM technique the rule generation

complexity increases by huge amount as the average transaction length increases and in the technique using fast algorithms though the rule generation complexity is lower the data structures used are not capable of handling huge amounts of data.

In addition to these techniques another paper [13] discusses a graph based approach towards association rule mining. This paper suggests an algorithm called ComboMatrix algorithm which predicts missing items using associative classification mining. The data structure used to store a graph is an adjacency matrix with a slight modification wherein the diagonal elements contain the list of vertices the adjacent to the given vertex. The advantages of this approach are reduced rule generation complexity however; the data structure used and the algorithm proposed do not perform well for large data sets.

The method proposed in this paper aims at reducing the rule generation complexity by classifying the items and then constructing a graph from it and using the graph for prediction purpose.

5. PROPOSED APPROACH

The objective of this paper is to design a prognostic system that will give operate efficiently for larger data sets. The technique classifies the data items prior to the prediction process. The classifier chosen is the Naïve Bayes text classifier. The Naïve Bayes text classifier works well for large datasets and is relatively simple to implement. It also gives good classification results. If the items in the dataset cannot be labeled then hierarchical clustering technique is suggested to cluster the data.

Once the classifier classifies the items a graph is constructed from the classified transactions. The classes form the nodes of the graph. The data structure used to store the graph is ComboMatrix [13]. Associative classification techniques are applied on the constructed graph using the algorithms proposed in [13].

The entire system can be divided into following blocks as shown in Fig.1:

- Data Extraction Phase
- Classification or clustering
- Graph construction from the classes
- Prediction based on the graph
- Output to the user

A graph is constructed from the classified transaction and graph is the basis on which the prediction algorithm is carried out. The output of the prediction algorithm is then listed to the users.

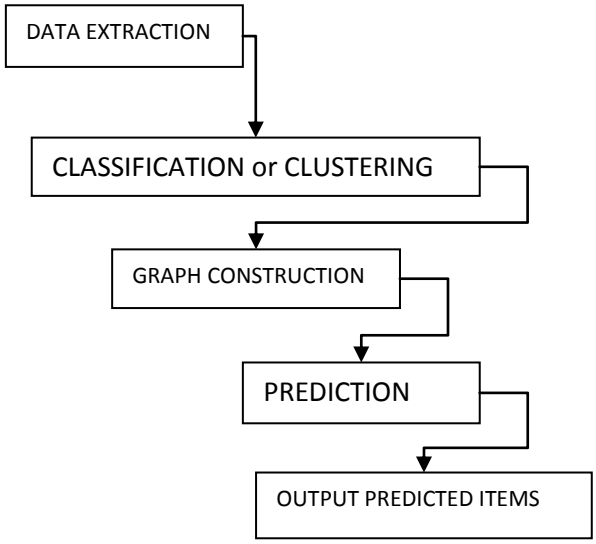

Fig 1: General Steps in Proposed Approach

Each of these steps is explained in detail in the following subsections.

5.1 Data Extraction

Typical unstructured data sources include web pages, emails, documents, PDFs, scanned text, mainframe reports, spool files etc. Extracting data from these unstructured sources has grown into a considerable technical challenge where as historically data extraction has had to deal with changes in physical hardware formats, the majority of current data extraction deals with extracting data from these unstructured data sources, and from different software formats. This growing process of data extraction from the web is referred to as Web scraping. For this phase an automated data extraction tool will be used to extract the data from the web sites. There are many tools available commercially and as open source.

The data is extracted from the web using a web data extractor called Web Harvest. Web Harvest is an open source Java based tool which helps in web data extraction. It offers a method to extract important and useful information from a set of collected web pages. To do this it leverages techniques such as XSLT, XQuery and Regular Expressions. Web-Harvest focuses on HTML/XML based web sites and is useful, since such pages still make a vast majority of the Web content. This tool can be easily supplemented by custom Java libraries to augment its extraction capabilities. Web-Harvest supports a set of useful processors for variable manipulation, conditional branching, looping, functions, file operations, HTML and XML processing, exception handling. The extracted information is stored in xml files.

5.2 Classification or Clustering

5.2.1 Naïve Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. The naive Bayes classifier greatly simplify learning by assuming that features are independent given class. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the

effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence.

Naïve Bayes classifier work well as text classifier. Text classification approach is used to classify products on shopping websites. Algorithms below illustrate the steps used to train and test Naïve Bayes text classifier.

The training algorithm takes as input a set of documents D and set of classes C. this algorithm first extracts the vocabulary i.e. distinct words from the training documents. N is the number of documents present in the training set. Steps 3 through 10 calculate the prior and conditional probabilities of words in each of their respective classes. The algorithm returns the extracted vocabulary and the prior and conditional probabilities.

Algorithm TrainNB(C, D)

1. $V \leftarrow$ ExtractVocabulary(D)
2. $N \leftarrow$ CountDocs(D)
3. for each $c \in C$
4. do $N_c \leftarrow CountDocsInClass(D, c)$
5. prior[c] $\leftarrow N_c / N$
6. $text_c \leftarrow ConcatTextOfAllDocsInClass(D, c)$
7. for each $t \in V$
8. do $T_{ct} \leftarrow CountTokensOfTerm(text_c, t)$
9. for each $t \in V$
10. do condprob[t][c] $\leftarrow \dfrac{T_{ct} + 1}{\sum_{t'}(T_{ct'} + 1)}$
11. return V, prior, condprob

Algorithm 1: Training Phase of Naïve Bayes Classifier

The test algorithm takes a document d as input and the other variables that are pre-calculated by the training phase. The algorithm then calculates the frequencies of the words in the document and returns the class where the score of the word is the highest. [14]

Algorithm TestNB(C, V, prior, condprob, d)

1. $W \leftarrow$ ExtractTokensFromDoc(V, d)
2. for each $c \in C$
3. do score[c] $\leftarrow$ log prior[c]
4. for each $t \in W$
5. do score[c] += log condprob[t][c]
6. return $\arg\max_{c \in C} score[c]$

Algorithm 2: Test Phase of Naïve Bayes Classifier

5.2.2 Hierarchical Document Clustering

Document clustering is the grouping of similar documents into clusters. The documents belonging to any cluster share higher similarity values with each other than with the documents belonging to other clusters. In hierarchical document clustering the documents are arranged in a hierarchical fashion, with the documents having parent-child relationship. Some of the special challenges faced by document clustering are high dimensionality, high volume of data, ease of browsing and establishing meaningful cluster labels.

Hierarchical document clustering can be performed in two ways viz. using the top down approach or more commonly known as the divisive approach and the bottom-up approach more popularly known as the agglomerative approach. In agglomerative hierarchical clustering the similarity is calculated between each pair of the clusters and clusters having highest similarity are merged. The divisive approach starts with all the data objects in one cluster and then iteratively splits a cluster into smaller clusters until some termination condition is satisfied. [12].

5.3 Graph Construction

The data structure chosen for the graph is the ComboMatrix. ComboMatrix is an adjacency matrix whose diagonal values contain the list of vertices that are adjacent to the corresponding vertex. For example, consider the following weighted graph showed in Fig 2.

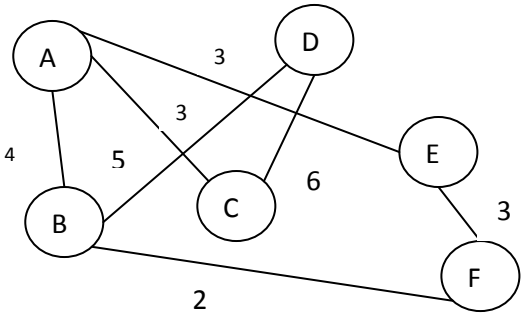

Fig 2: Weighted Graph

The ComboMatrix for the above graph is shown in Fig 3. The ComoboMatrix, as shown contains the list of adjacent vertices to a given vertex in its diagonal elements. The non diagonal elements contain the weights of the corresponding edges. For example, vertex A is adjacent to vertices B, C and E. therefore the diagonal entry for vertex A contains the list B, C and E. The entry for the cells, (A,B) contain the weight of the edge A-B which is 4 and so on. As seen in the example, the comboMatrix contains 36 entries for a graph with 6 nodes. Therefore the memory complexity of a comboMatrix is V^2 where V is the number of vertices in the graph. This huge amount of memory requirement makes the use of a ComboMatrix for huge datasets difficult.

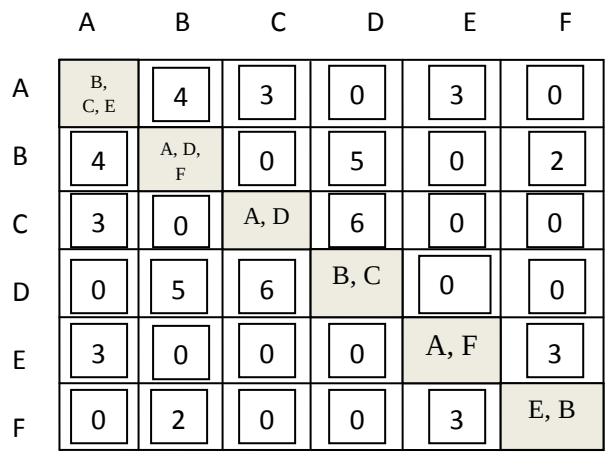

Fig 3: ComboMatrix for the given graph

One method to make this data structure and the associated prediction algorithm work for large datasets, especially on online shopping websites, is by employing a mechanism to map the large dataset to a set of smaller size. Classification and clustering methods help to do this since the number of classes and clusters are usually way less than the number of items in the dataset.

Classifying or clustering the data items online, and then predicting the missing classes, provides valuable association information on the Web. On shopping websites, a large number of products are available which are listed in various categories and sub-categories. Instead of establishing direct association between individual data items, it is sufficient to establish association between the various classes available on the website.

5.4 Prediction Algorithm

The prediction algorithm employed is the associative classification algorithm proposed in [13] which makes use of ComboMatrix. The algorithm first constructs the graph from a given set of transactions that have already occurred. Based on the occurrences of various items in the transaction history, the weights of the graph are updated. The prediction algorithm is as shown below.

Algorithm predict(G[ ][ ], trans[ ])
1. for each item ∈ trans
2. do adjacentVertices ← G[item][item]
3. for each v ∈ adjacentVertices
4. do weight ← G[item][v]
5. if(weight >= threshold)
6. add vertex to the list
7. sort the list and display to the user

Algorithm 3: Prediction Algorithm

For predicting missing items, given an input transaction, the adjacent vertices are extracted from the diagonal entries in the ComboMatrix for each item present in the input transaction. The weights of the edges between the given item and its adjacent vertices are then checked to be above a particular threshold value. If the weights are above the threshold value, the vertices are listed to the user in decreasing order of their weights.

This algorithm is simple yet efficient for predicting missing items. The added advantages for this algorithm are reduced rule generation complexity and single pass over the database to predict the missing items.

6. RESULTS

The system is implemented using Java and is compared to the existing ComboMatrix algorithm [13]. The system is analyzed for varying lengths of transaction. The factors considered for evaluation are time and memory requirements.

The training set consists of 100 records, containing 10 classes. The data set is extracted from Yahoo! Shopping websites. The classes present in the training data are Mobile Phones, Camcorders, Cameras, Projectors, Audio Systems, Video Systems, GPS, Car Audio System, MP3 Players and Television. The number of items present in the transaction history is 40.

For text classification the feature used is all the words from the title. To further enhance the classification process only nouns from the title can be used thus reducing the vocabulary size and thereby reducing the training complexity.

The method first classifies the items present in a transaction and constructs the classified transaction. In this experiment, the transactions of varying lengths are reduced to transaction having a maximum length of 10units since the number of classes is 10. The graph is constructed based on the classified transactions. The table below compares the maximum memory required for the modified and existing algorithm.

Table 1. Memory Requirements for modified and existing algorithm

Number of Items in Database	Maximum Memory required for modified algorithm	Maximum Memory required for modified algorithm
10	100	100
20	100	400
30	100	900
40	100	1600
50	100	2500
60	100	3600
70	100	4900

The memory requirements for Modified ComboMatrix algorithm and the existing ComboMatrix algorithm are as shown below. The memory requirement for the modified algorithm is constant at 100 because the number of classes is 10 so the items in the database are mapped to one of these 10 classes and therefore the matrix needs a maximum of 100 entries.

The overhead cost of employing a classifier or clustering mechanism is negligible since training is a performed only once and the conditional probabilities are computed. The test operation does not result in performance degradation.

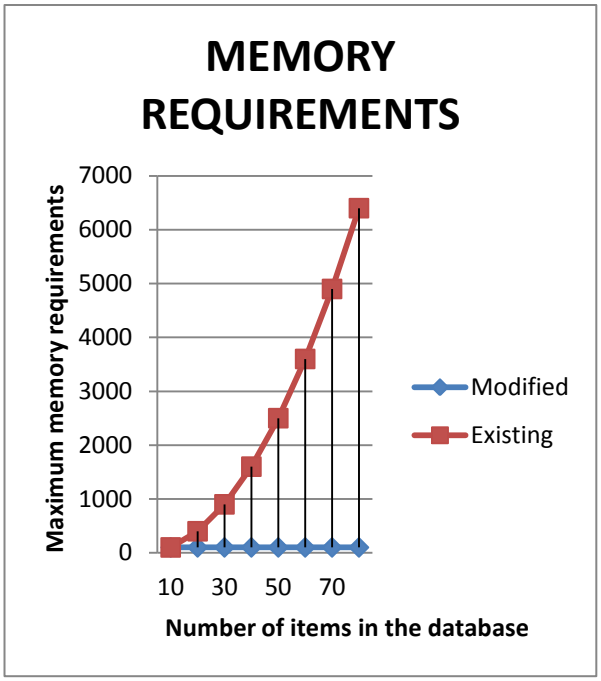

Fig 4: Graph comparing the Maximum Memory Requirements for the existing algorithm and the algorithm modified by using classification

The test results also show that the performance of the modified algorithm is better than that of the existing algorithm. For 40 items in the database, the original algorithm needs to compute a matrix of size 40*40, for every transaction in the history. However the modified algorithm needs to compute matrices of size 10*10 for every classified transaction in history. The performance can be further enhanced by dividing the processes into forground and background processes. The classification of the transaction and graph construction can be done as a backgorund process while the prediction can be done as a foreground process.

7. CONCLUSION

Web mining is a vast area of research. This paper aims to design a prognostic recommender system using supervised and unsupervised technique. This system predicts missing items based on the past information and suggests the same to the users. To fulfill this task the system uses classification techniques prior to the prediction process. The advantage of using classification/clustering is that the prediction is done at a higher level of abstraction and the cost of rule generation in association rule mining is minimized. Out of the various options available, Naïve Bayes classifier is chosen for classification since this classifier has worked well for large data sets and in relatively simple to implement. Hierarchical clustering mechanism is chosen for clustering. The data extraction is done using an automated data extraction tool web Harvest. The data structure chosen for graphs is a ComboMatrix which is a variation of adjacency matrix. The prediction algorithm used in associative classification mining using ComboMatrix. The proposed approach is discussed and the design of the system is presented. The advantages of this method are as follows:

- Prediction at a higher level of abstraction
- Better Performance
- Scalable
- Reduces Rule generation complexity

The results show that the proposed method performs better and has lower memory requirements.

8. REFERENCES

[1] Kasun Wickramaratna, Miroslav Kubat and Kamal Premaratne, “Predicting Missing Items in Shopping Carts”, IEEE Trans. Knowledge and Data Eng., vol. 21, no. 7, July 2009.

[2] Srivatsan. M, Sunil Kumar. M, Vijayshankar. V, Leela Rani P, “Predicting Missing Items in Shopping Carts using Fast Algorithm”, International Journal of Computer Applications (0975 – 8887) Volume 21– No.5, May 2011

[3] Faustina Johnson and Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.

[4] MENG Xiaofeng, LU Hongjun, WANG Haiyan and GU Mingzhe, “Data Extraction from the Web Based on Pre-Defined Schema”, J. Comput. Sci. & Technol., Vol.17 No.4, July 2002.

[5] David F. Barrero and David Camacho and Maria D. R-Moreno, “Automatic Web Data Extraction based on Genetic Algorithms and Regular Expressions”.

[6] Yanhong Zhai and Bing Liu, “Web Data Extraction Based on Partial Tree Alignment”, WWW 2005, May 10-14, 2005.

[7] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara and Robert Baumgartner, “Web Data Extraction, Applications and Techniques: A Survey”, ACM Computing Surveys, Vol. V, No. N, July 2012.

[8] Alberto H. F. Laender, Berthier A. Ribeiro Neto, Altigran S. da Silva and Juliana S. Teixeira, “A Brief Survey of Web Data Extraction Tools”.

[10] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen, “Collaborative Filtering Recommender Systems”.

[11] Hemalatha Chandrashekhar and Bharat Bhasker, “Personalized Recommender System Using Entropy Based Collaborative Filtering Technique”.

[12] Benjamin C. M. Fung, Ke Wang, and Martin Ester, “Hierarchical Document Clustering”

[13] Ila Padhi , Jibitesh Mishra, Sanjit Kumar Dash, “Predicting Missing Items in Shopping Cart using Associative Classification Mining”, International Journal of Computer Applications (0975 – 8887) Volume 50 – No.14, July 2012.

[14] http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.