International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-021 PAGE NO: 111-115

A Recent Approach on Identifying the ARM Based Clustering Algorithm on Text and Image Context

Dr. T. Karthikeyan [1], J. Mohana Sundaram [2], R. Karthik Raj [3]

[1]*Associate Professor, Department of Computer Science*, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India
[2]*Assistant Professor, Department of Computer Science*, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India
[3]*PG Student, Department of Computer Science*, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India

[1]t.karthikeyan.gasc@gmail.com,[2]majornavy@gmail.com,[3] karthik.grm@gmail.com.

Abstract— **Each and every association rule has an aid and a confidence. Association Rules (AR) can also considered as itemset. Association Rule Mining (ARM) gets into the new technological fields. The fields consist of Classification Association Rule Mining (CARM) and distributed ARM. Now a day's work in progress for developing faster execution of ARM algorithm which will reduce the execution time. Text mining is dealt with the application of data mining techniques to document sets. Image Association Rule Mining is involved with the application of Association Rule Mining techniques to image collections. Here large number of textual data contains many itemsets. If we mined the itemsets many times in short interval of time it reduce the dimensionality of the given text document extremely. To get the meaningful information for the given text document many data mining techniques used. This technique contains clustering. Text clustering means a group of documents into many groups.**

Keywords: **Association Rule Mining (ARM), Classification Association Rule Mining (CARM), Text mining, Image ARM, Text Clustering.**

INTRODUCTION

The growth of database form the point of human gives many tools for giving data to effective knowledge. The important research areas called Data Mining (DM) and Knowledge Discovery in Databases (KDD). Data Mining is done on data in textual manner. Today there are large quantities of information's in the form of text, documents, etc. This textual data can be used in more effective manner. A recent approach for text mining with the help of item sets. Knowledge Discovery can done form textual database is also called Text Mining. Text Mining is used to get knowledge from the huge number of text files such as documents, especially on web. Information Mining is a combination of text mining and data mining. For text mining data mining methods also used which will be more useful during mining of text. Text mining is an application of DM. ARM provides a statistical relation between more than ones variables such that systematic changes in the value of single variable are accompanied by systematic changes with huge amount data items.

I. CLUSTER ANALYSIS

The process of partitioning the data objects into meaningful cluster or group of objects within the cluster. The analogous aspect of something to the object in other cluster is called as Cluster analysis. With the help of cluster analysis we can make better accuracy and retrieve the information, from the obtained information it is easy to get the neighbour document.

II. A RECENT APPROACH FOR TEXT CLUSTERING

There are huge quantity of document available in electronic form which consist of more hidden content in the given text document. So with the help of recent approach of ARM we can find the purpose of the text document that is we can get the hidden message. This can be done with help of clustering technique. Text clustering contains the collection of text document which gives the same contents. There are several ways to make better performance of text document with the respective algorithm for clustering. The recent approaches for the text clustering are:

1) Text pre-processing
2) Frequent mining of itemsets
3) Dividing the text content with respect to the itemsets
4) Clustering within the given partition.

A. Text pre-processing: The text pre-processing consists of changing of raw text files form one format to another. Normally text files in the form of digital form that is bits. We can done text pre-processing with different language. Here conversion of the given text document form one form to another is done easily. That is the encryption text document is decrypted for the text processing, extraction of the text document is called as tokenization. There are two techniques for text processing they are:

1. Stop word removal
2. Stemming algorithm

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-021 PAGE NO: 111-115

Stop word removal: Eliminate the linking words such as have, then, it, can, need, but, they, form, was, the, to, also for the corresponding document.
Stemming algorithm: Eliminate the prefix and suffix from the given document.

B. Frequent mining of itemset: Mining of several itemset plays essential role in pattern mining. To develop the skill of the mining process there is a framework called tree based framework for the identifying the itemsets. There is an algorithm proposed for framework is called as Frequent Pattern Tree to detect the database transaction.

C. Dividing the text content with respect to the itemsets: Frequent item set contain the collection of words in a group of documents.
For developing the initial partition, we should use the itemset which mined previously so it can reduce the dimension of the text document, it will be more efficient. Here due the group of itemset there will be overlapping of itemset arise it can be rectified during the result.

D.Clustering within the given partition: Here we should show the cluster technique obtained from the above step. With the help of the cluster from the previous step again make a partition to obtain the sub cluster from this cluster derive the outline of the document in an effective manner after this we not require any specific number of cluster.

III. TEXT MINING

Text Mining is abstract of separation of a whole into its constituent parts in order to study the parts and their relations. It is a process of formulating the input text and derive the output pattern with meaningful data in the form of pattern. High Quality in the text explains the concept of relation of something, originality of the text content, and power of attracting something. With the help of text mining we can derive the high quality information form the text document. It is based on the application of data mining with association of document collection. The main goal of ARM is to deduct the patterns and relations which links the document inside the collection. The Vector Space Model is the technology used to represents ARM in text. Here document is shown as Single, High Dimensional, and Numeric Vector. The whole document is based on the group of word with particular document which is equivalent subset collection In this context the word of document is arrange the group of words, in this original document is lost. There is a disadvantage in bag of words approach in this we want to pick out limited words form the whole document. There is a technology for reducing the number of words they are

1. Stop Lists
2. Stemming or Word Normalisation Technique

There as approach called "Bag of term" approach which is alternative for "Bag of words" approach. Here Phrase based approach is also used which is an alternative way, which will work better than keyword approach. The phrase based approach contains of phrase which would contain more information. The advantage of phrases are

a. Phrases have lower level statistical properties
b. Phrases have inferior frequency of repetitiveness.
c. Bag of phrases include many require phrases

In this also many disadvantage which can be rectifies by CARM algorithm. There is a set of rules followed for phrase based approach they are

i. Remove more identical words.
ii. Terminate the rarely applicable words
iii. Select the important word from the remaining words available.
iv. Make a new significant word and connecting words.

Here identical words and rarely applicable words are called as noise. The words which has the ability of Upper Noise Threshold and these words are called as Upper Noise Words. The word which has the ability of Lower Noise Threshold called as Lower Noise Words.

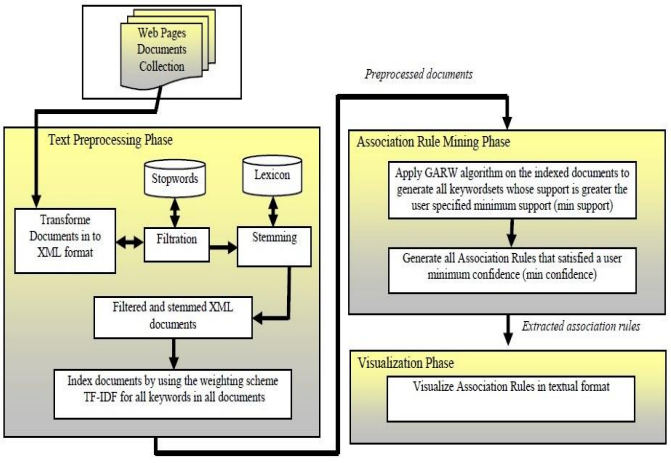

Fig 1: Text mining Architecture

Text mining is an Extraction Association Rule from Text (EART)
A.The aim of the text pre-processing is to optimize the performance of the next phase

a. Transformation

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-021 PAGE NO: 111-115

The system accepts a different number of documents formats (doc, txt, rtf, etc.) and structures to convert them into the XML format amenable for the further processing.

b. Filtration

In this process the documents are filtered by removing the unimportant words form documents content.

C.*Indexing*

The filtered and stemmed XML document are then index by using the weighting scheme

B.Visualization Phase: The extracted association rule can be reviewed in textual format or tables, or in graphical format. In this phase, the system is designed to visualize the extracted association rule in textual format or tables.

IMAGE MINING

Image Mining is also associated with Association Rule Mining technique to the group of image. The several reasons for analysing the set of images. Common image analysis application is done mostly in medical fields. There are many application using image mining which will again bring back the image and converted into a meaningful form and use data mining process. With the help of ARM we can transform the given image into another form. The image is represented in the form of numeric term and represented in vector elements. Here another approach for the analysis of image is Tesseral Approach which has two main disadvantages. Another method form image mining is quad tree representation in which image is tessellated to many stages of resolution which is vary in degree of different area of image. The quad tree is sequenced into vector form. Image mining is deals with two approaches first it is retrieved from database and group of image sets, second is to mining the image with αnumeric and image sets.

IV. IMAGE MINING TECHNOLOGY

Image mining is used to mine data to get knowledge form the corresponding image. There are several technique they are:

Object Recognition:

- It is capable to identify an object form the image with the help of object recognition.

Image Indexing:

- It gives an effective support for recovering information.

Image retrieval:

- Image are get back based on level of being intricate the requirement.

V. IMAGE CLASSIFICATION AND CLUSTERING

It classify the images into sets based on the observation and non-observation manner.

ARM:

- With the help of Association Rule Mining we can obtain several pattern.

Neural Network:

- With the help of neural network we can retrieve and store more knowledge presents.

VI. MINING USING APRIORI ALGORITHM

Apriori is an algorithm for mining the frequent itemset and AR learning within the database.

Example: Consider the below database, in which it consist of row and column or cell here row is called as transaction, column or cell is called as item of the corresponding transaction:

A	β	γ
A	β	θ
A	β	ε
A	β	θ

The AR can be easily understand from the above table how the content are associated with each other:

- 100% of table with α also contain β
- 25% of table with α, β also have γ
- 50% of table with α, β also have

Concepts: The concepts contains some rule, conditions and interest. The common condition is there should be minimum thresholds on the support.

Support: The Support Supp(X) of the item set X is the proportion of transactions in the given data set in which item set presents.

Supp(X)= no. of transactions which contain the itemset X / total no. of transactions

Confidence: This is one of the rules

$$\text{Conf}(X \rightarrow Y) = \text{Supp}(X \cup Y) / \text{Supp}(X)$$

Lift: The lift rule is defined as

$$\text{lift}(X \rightarrow Y)$$

Conviction: The conviction rule is said by

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-021 PAGE NO: 111-115

Item	Support
1	6
2	7
3	9
4	8
5	6

$$\text{conv}(X \rightarrow Y) = \frac{1 - \text{Supp}(Y)}{1 - \text{conf}(X \rightarrow Y)}$$

VII. APRIORIALGORITHM

The usage of apriori algorithm is

- Find the item sets in the given database
- With the help of the item set and the confidence condition frame the rules

Obtaining the item sets in the given database is very hard because it should find all the corresponding item sets. The set of all itemset is power to the set I and the size is $2_n - 1$. Even though the size of power set is increment exponentially in the total number of items that is n in I, here search is possible using the closure property.

Apriori Algorithm Pseudocode

procedure **Apriori** (T, *minSupport*) { //T is the database and *minSupport* is the minimum support
L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++) **{**
Ck= candidates generated from Lk-1
for each transaction **t** in database
do
{
#increment the count of all candidates in Ck that are contained in t
Lk = candidates in Ck with *minSupport*
}//end for each
}//end for
return $\mathbf{U_k}$ $\mathbf{L_k}$;
}

It is the common ARM, in which the set of item set of the algorithm search for the subset which is occurred in common that is minimum number C of the item sets. The Apriori uses Bottom Up Approach. The algorithm ends when there will be no extension are occurred. This algorithm uses breadth first search and the structure of tree is used to count item set easily. It contains the item set of K from the item set with length K – 1.

VIII. SAMPLE USAGE OF APRIORI ALGORITHM

In a supermarket sales data is tracked by SKU that is Stock-Keeping-Unit it contains the item purchased together. With the help of apriori algorithm we can construct list of item purchased. The Database contains following transaction they are, {1,2,3,4}, {1,2,3,4,5}, {2,3,4}, {2,3,5}, {1,2,4}, {1,3,4}, {2,3,4,5}, {1,3,4,5}, {3,4,5}, {1,2,3,5}. Each number is the product points the butter or water.

The first step of this algorithm is to count the frequencies which is also called as support.

Now generate the list of 2-pairs of items

Item	Support
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

Now generate the list of 3-pairs of items

Item	Support
{1,3,4}	4
{2,3,4}	4
{2,3,5}	4
{3,4,5}	4

Now the algorithm going to end because the pair {2, 3, 4, 5} does not have corresponding support.

Result:

Step 1:

Item	Support
1	6
2	7
3	9
4	8
5	6

Step 2:

Item	Support
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-021 PAGE NO: 111-115

Applications of Apriori

1. Apriori algorithm used in the field of adverse drug reaction detection

The apriori algorithm application is used in Adverse Drug Reaction (ADR). The Apriori algorithm is used to surveillance the patients, what are the drugs the patient using, identify the nature. This ADR is based on AR which indicates what combinations of medications and patient characteristics should lead to ADRs.

2. Apriori Algorithm also used in Oracle Bone Inscription Explication

Oracle Bone Inscription (OBI) is one of the oldest writing in the world. There are six thousand words are found till date. But only Thousand five hundred words are found explicitly. This explication in OBI is the key problem. First and foremost thing is the OBI data should be extracted from the OBI corpus that should be pre-processed; these processed data should be given input to apriori algorithm so we get item sets. And now the OBI words are found. That this method is feasible and effective in the finding correlation.

IX. CONCLUSION

In this short paper two application areas have been presented which may prove fruitful in the application of current ARM technology. In particular the application domains of text and image mining have been considered in some detail. In all cases the fundamental challenge is how best to translate the input data into a tabular format that will permit the application of ARM technology.

REFERENCES

[1] Agrawal, R, Imielinski, T. and Swami, A. (1993). *Mining association rules between sets of items in large databases*. Proc. ACM SIGMOD Conference on Management of Data, pp 207-216.

[2] Agrawal, R. and Srikant, A. (1994). *Fast algorithms for mining association rules*. Proc. VLDB'94, pp 487-499.

[3] Borglet, C. and Berhold, M.R. (2002). *Mining Molecular Fragments: Finding RelevantSubstructures of Molecules*. Proc ICDM'02, IEEE, pp 51-58.

[4] Ding, Q., Ding, Q. and Perrizo, W. (2002). *Association Rule Mining on remotely sensed images using P-trees*. Proc PAKDD. pp 66-79.

[5] Han, J., Pei, J. and Yiwen, Y. (2000). *Mining Frequent Patterns Without Candidate Generation.* Proceedings ACM-SIGMOD International Conference on Management of Data, ACM Press, pp 1-12.

[6] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. VLDB'98, 594-605, New York, NY.

[7] Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey, "A Text Mining Technique Using Association Rules Extraction", *International Journal of Computational Intelligence*, Vol. 4; No. 1, 2008

[8] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", *in proceedings of the KDD-2000 Workshop on Text Mining*, Boston, MA, pp. 109-111, 2000.

[9] Association Rule Mining in The Wider Context of Text, Images and Graphs by F Coenen Department of Computer Science, The University of Liverpool.

[10] A Text Mining Technique Using Association Rules Extraction by Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey.

Published by SK Research Group of Companies (SKRGC).