Journal of Theoretical and Applied Information Technology
30th November 2011. Vol. 33 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING EFFICIENCY OF TEXTUAL STATIC WEB CONTENT MINING USING CLUSTERING TECHNIQUES

R.MANIKANDAN
Senior Assistant Professor, School of Computing , SASTRA University,Thanjavur-613401,India
E-mail: manikandan75@core.sastra.edu

ABSTRACT

There are several efficient methods for the discovery or mining of various types of data , methods devised for mining textual static web content have always been proved less efficient due to the data's ambiguous , unclassified, unstructured or unclustered nature. Various association rule mining algorithms like Generalized pattern algorithm are being implemented to mine the web content but again due to the above setbacks the efficiency expected from the algorithm is not obtained. Since the dip in the efficiency of these algorithms is amounted to the nature of the textual web content, an algorithm which may deal with , if not all the anomalies at least the unclustered nature of the content may increase the efficiency drastically. This paper emphasizes the same point, making the assumptions that the web content is static and there is at least one common pattern found in the given datasets.

Keywords: *Textual Web Content, Unclustered Nature, Generalized Pattern Algorithm, Datasets, Clustering Algorithm*

1. INTRODUCTION

This paper describes the issues of "Improving Efficiency of Textual static Web Content Mining Using Clustering Techniques".

Our intent is to draw a comparison between the text mining techniques using clustering algorithms and without clustering algorithms. This paper describes the need for clustering of data and present conclusions about the improvement in efficiency when it is done.

This paper provides strong mathematical, graphical and hypothetical explanations as of how the clustering techniques will optimize the mining algorithm efficiency.

Rest of the document is divided into three parts viz. Existing system, proposed system and Representations.

Existing system discusses about the prevailing system of text mining, algorithms being used and its drawbacks.

Proposed system will give an insight of the intended idea by explaining the design of the clustering algorithm and an existing associate rule text mining algorithm of which the efficiency can be improved.

Representations give mathematical, hypothetical and graphical representation as of how the efficiency is being improved because of using the clustering techniques.

2. EXISTING SYSTEM

Text mining is an emerging technology for extracting meaning from the "unclustered and unstructured" text that constitutes a majority of enterprise information assets. Applied to routine customer and business interactions, regulatory and financial fillings, operational reports and documentation and new articles, Text mining extracts concepts to create and apply taxonomies that categorize the wealth of the enterprise information. Just as data mining discerns patterns in numeric data for predictive modeling, text mining identifies conceptually-based interconnections to classify documents and automate process ranging from customer service to the prevention any sort of attacks.[4],[5]

Existing techniques mainly transform text documents into simplistic intermediate forms, such as term vectors and bags of keywords. As terms are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

treated as individual items at such simplistic representations the original text patterns lose its actual meaning and the intended results may not show up. The draw back of such systems as GENARALIZED ASSOCIATION PATTERN ALGORITHM is their inherent way of treating each intermediate form as an individual item rather than a clustered pattern.[1][2],[3]

To Overcome this limitation, we need a new technique to discover the text based on the clustered text patterns which can achieved by using clustering algorithms such as K-Means Algorithm or PSO Clustering algorithm.

3. PROPOSED SYSTEM

The existing i.e. the normal implementation of generalized pattern mining algorithm has a few disadvantages. It has to check for patterns every time it is executed which actually consumes more time, i.e. the algorithm scans every time the whole collection of text documents and the pattern search space is large. This project aims to overcome this limitation by integrating generalized pattern mining algorithm and clustering techniques. Clustering is applied to group similar text documents based on the concepts chosen. Then the generalized pattern algorithm is applied on each cluster to mine significant patterns. Here the pattern search space depends on the cluster size.

While Generating the relation association patterns, the time spent in mining cluster is less than the time spent in mining whole collection. Time efficiency is achieved and scalability is improved due to decrease in number of scans through the collection.

Though both algorithms are well known the integration of these two is the striking feature of the project. Rest of the section will give an overview of the algorithms to be used and a brief explanation.

A. Clustering algorithm

The clustering algorithm to be used in this project is a simple K-Means clustering algorithm. During clustering, K-Means algorithm is applied to discover clusters on the documents with the textual web content. Number of clusters depends on the number of concepts specified in the framework.

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. These centroids have to be selected carefully since their placement will always affect the end result. So, the better choice is to place them as far as possible. The algorithm is composed of following steps:

(1) Place k points into the space represented by the objects that are being clustered. These points represent initial group of centroids.
(2) Assign each object the group that has the closest centroid.
(3) When all objects have been assigned, recalculate the positions of the k centroids.
(4) Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups of the objects into groups from which the metric to be minimized can be calculated.

The sensitiveness of the algorithm to the initial set of randomly chosen centroids can be reduced by running the algorithm multiple times.

K-Means algorithm is a simple algorithm which can be used evidently to improve the performance of the mining process together with the generalized pattern algorithm.

B. Generalized pattern mining

A number of data mining algorithms have been recently developed that greatly facilitate the processing and interpreting of large stores of data. One example is the generalized pattern algorithm, which discovers correlations between items in transactional databases.

For discovering associations between items across different levels of taxonomy, generalized pattern algorithm is proposed. In particular, Users may require the generated rules to have a large support to avoid trivial knowledge being discovered. In this case, a large portion of the rules that include only the leaves of the item taxonomy may be filtered away.
Generalized pattern algorithm allows users to extract small set of useful rules instead of generating a large set of trivial ones. This Property is especially important for mining associations

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

from large text data sets, in which the support of most items (words) is very low.
The generalized pattern algorithm can be given in two modules. It is composed of following steps:

Module 1: GPC

Input : A Simple collection of files with multiple patterns.
Output: The set of all closed frequent generalization closures: C

(1) root=0;
root.supp=1 //initialize closure enumeration tree root
(2) Constructing closures of frequent relation sets as child closures of roots.
(3) sort (root.children) //sort child-closures in a length decreasing or support increasing manner.
(4) Closure-Enum(root,C=0)
(5)return C

Module 2:
Input : A node in the closure enumeration tree :n
A set of discovered frequent closed generalization closures
Output : The expanded set of frequent closed generalization Closures:C

- //Subtree pruning.
- Children-prune(n.children**)** // pruning child-closure
- c=closed-closure(n) //generate a locally closed generalization closure c and c must also be globally closed according to our subtree pruning strategy.
- //insert c into the frequent closed generalization closure set.
- For each child-closure ichild of n where n belongs to n.children do
- Generate the child-closures of ichild by merging ichild with one of its subsequent siblings.
- Closure-enum(ichild,c) //Recursively visit the child-closure of the current tree node n
- Return C.

C.EXPLANATIONS

This section as already stated will give an overview of the various arguments viz. Hypothetical, mathematical and graphical representations to support the integration of clustering techniques with the existing mining techniques.

Following are the initial hypothetical arguments of this proposal since the implementation was done only in the later stage. The arguments are:

A. This paper has proposed an approach for discovering knowledge quickly from textual web content.
B. Integration of clustering and generalized pattern algorithm can substantially reduce the pattern redundancy and perform much better than the ordinary implementation of the generalized pattern algorithm without clustering in the terms of efficiency.

C. Since clustering is applied to group of similar text documents, relative pattern search space is minimized.
Scalability is improved due to small number of database scans
Possible mathematical explanation of this change or improvement can be given as follows:
For a Tree Based algorithm the average case scenario of efficiency can be given in the order n*logn i.e. O(n *logn)
This algorithm shows a direct impact by reducing n i.e. the pattern search space.
For Example, when applied in business transactions, Searching for profit transactions of the year in all the records is rather meaning less. Instead just clustering all records into various clusters like profit cluster, loss cluster etc. will reduce the pattern search space

Now if we have totally 50,000 files in my operational database or my data warehouse. The efficiency of the algorithm as we said will be in the order of 50000 * log (50000) and if the number of profit related files in those 50000 are only some 2000 the efficiency or the time taken can be given in the order of 2000 * log 2000. Similarly for various number of files in a data warehouse, assuming there are only 4% profit related files, the hypothetical tabulation can be given as follows:

S.N	Total	Order for n files	Order of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

o	Files(n)		searching Profit based files
1	50000	234948.500	6602.0599
2	40000	184082.3996	5126.5919
3	30000	134313.63	3695.0174
4	20000	86020.56	2322.471

Hence by observing the above tabulation we can find that there will be an improvement in the efficiency of the algorithm when a clustered input is applied.

4. CONCLUSION

Given calculations in this paper are almost hypothetical though there is a glint of practicality in it. Practical work on the same is ensuing nevertheless the hypothetical explanations give base of what the idea is basically behind clustering the inputs.

Though the practical efficiency may not be in the order of what is shown in the hypothetical tabulation given. There won't be a radical shift in practical calculations. Hence it is safe to conclude that there will be definitely an improvement in the efficiency in the algorithms such as generalized pattern search rule algorithm.

With the support of the hypothetical arguments we have, the idea of the paper is proved and further work can also be done in this direction by taking the concepts of unclassified and ambiguous nature of the data.

REFERENCES

[1] R.Srikant and R. Agarwal, "Mining Generalized Association Rules", Proc.Conference on very large databases,1995.

[2] N.Pasquier,Y.Bastide,R.Taouil, and L.Lakhal," Discovering Frequent Closed Item sets for Association Rules" , Proc.Internatonal Conference on Database Theory,1999.

[3] J. Wang, J. Han and J. Pei, "Closet:Searching for the best strategies for mining frequent closed item sets",Proc.Ninth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining,2003.

[4] J.D.Holt and S.M.Chung ," Multi pass Algorithms foe Mining Association Rules in Text Databases",Knowledge Information System,2001.

[5] M.M.Y.Gomez,A.F.Gelbukh, and A.Lopez-Lopez, "Text Mining Detail Level", McGraw hill Publications,2002.