IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

# Document Clustering for College Student Database

**Amol Patil[1], Shahezad Arif[2], Ajit Patil[3], Prof. Deveshree Wankhede[4]**

Student of Computer Engineering, Shivajirao S. Jondhale C.O.E Dombivli, Mumbai University[1, 2, 3]

Assistant Professor of Computer Department, Shivajirao S. Jondhale C.O. E Dombivli, Mumbai University[4]

**Abstract:** Large number of data is available on internet today, these textual data constitutes resources. It is a difficult and important challenge to discover knowledge from textual database or for short text mining. The reason behind this is its richness and its ambiguity of natural language, which also affects analyzing of the data. Thus the question arises who is responsible to read and analyse the data? In this context, manual analysis and effective extraction of useful information may be possible. We think the solution is that it is relevant to provide automatic tools for analyzing large textual collections by automatically finding relevant Info. It depends on keyword features for discovering association rules amongst keywords and labelling the documents. In this work, system ignores the order in which the words occur, but instead it focuses on the words and their statistical distributions in documents. The main contributions of the technique are that to convert the document from unstructured to structured form with Information Retrieval scheme i.e. TF-IDF (for keyword/feature selection that automatically selects the most frequently occurred keywords to generate association rules) and use Data Mining technique for association rules discovery. The system requires Pre-processing phase such as transformation, stemming and indexing of the documents. Stemming is common requirement of natural processing function. The main purpose of stemming is to reduce different grammatical word forms of a word (noun, adjective, verb, adverb etc.) to its root form. Association Rule Mining (ARM) phase and Visualization phase i.e. visualization of results. The input is selected as static web pages related to road accidents on district level and its preventive measures.

**Keywords:** HARMT algorithm; Porter Stemming Algorithm; Text Preprocessing phase; Association Rule Mining Phase; network lifetime.

## I. INTRODUCTION

Now a days large number of data is available on Web, digital libraries. These textual data constitute a resource that is worth exploiting. In this way knowledge searching from textual database, or for short, text mining (TM) is an important and difficult challenges, because of the richness and ambiguity of natural language Therefore, there is problem of analyzing those data. So the question is who is able to read and analyze it? In this context, manual analysis and effective extraction of useful info are not possible.

We think the solution is that It is relevant to provide automatic tools for analyse big textual collections by auto find relevant Info. Text mining is an Increasingly Important research field because of the necessity of knowledge from the enormous number of text documents available, We describe use of Association Rules in TM Association rules highlight correlation between features In the texts, e.g. keywords a Word is selected as a keyword It' It does not appear In a predefined stop-words 1ist Moreover, association rules are easy to understand and Interpret for an analyst or may be for normal users. We have described system for automatically extracting association rules from Web page documents

A stemming algorithm is a process of linguistic normalization in which the variant forms of word are reduce the common forms, for example, **(Connection, Connections, Connective, Connected)** has root word as **connect.** It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related word map to the same stem, even if this stem is not in itself a valid root. For this reason, a number of so-called stemming Algorithms or stemmers. which attempt to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems .This not only means that different variants of a term can be merged to a single representative form – it also reduces the dictionary.

For accurate stemmer the author Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol 2 (6), (year 2011) has concluded there is no perfect stemmers has designed so far to match all the requirements. But in the porter stemmer it produces the best output as compared to other stemmers. And also has less error rate [2].

For generating association rule on stemmed tokens "Qiankun Zhao Nanyang Technological University, Singapore and Sourav S. Bhowmick Nanyang Technological University, Singapore (2003)" stated that Apriori is a great improvement in the history of association rule mining, Apriori algorithm was first proposed by Agrawal in [Agrawaland Srikant 1994]. The AIS is just a straightforward approach that requires many passes over the database, generating many object itemsets

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 2, February 2016*

and storing counters of each candidate while most of them turn out to be not frequent. Apriori is the more efficient during the candidate generation process for two reasons, Apriori employees a different candidates generation method and a new pruning technique [5].

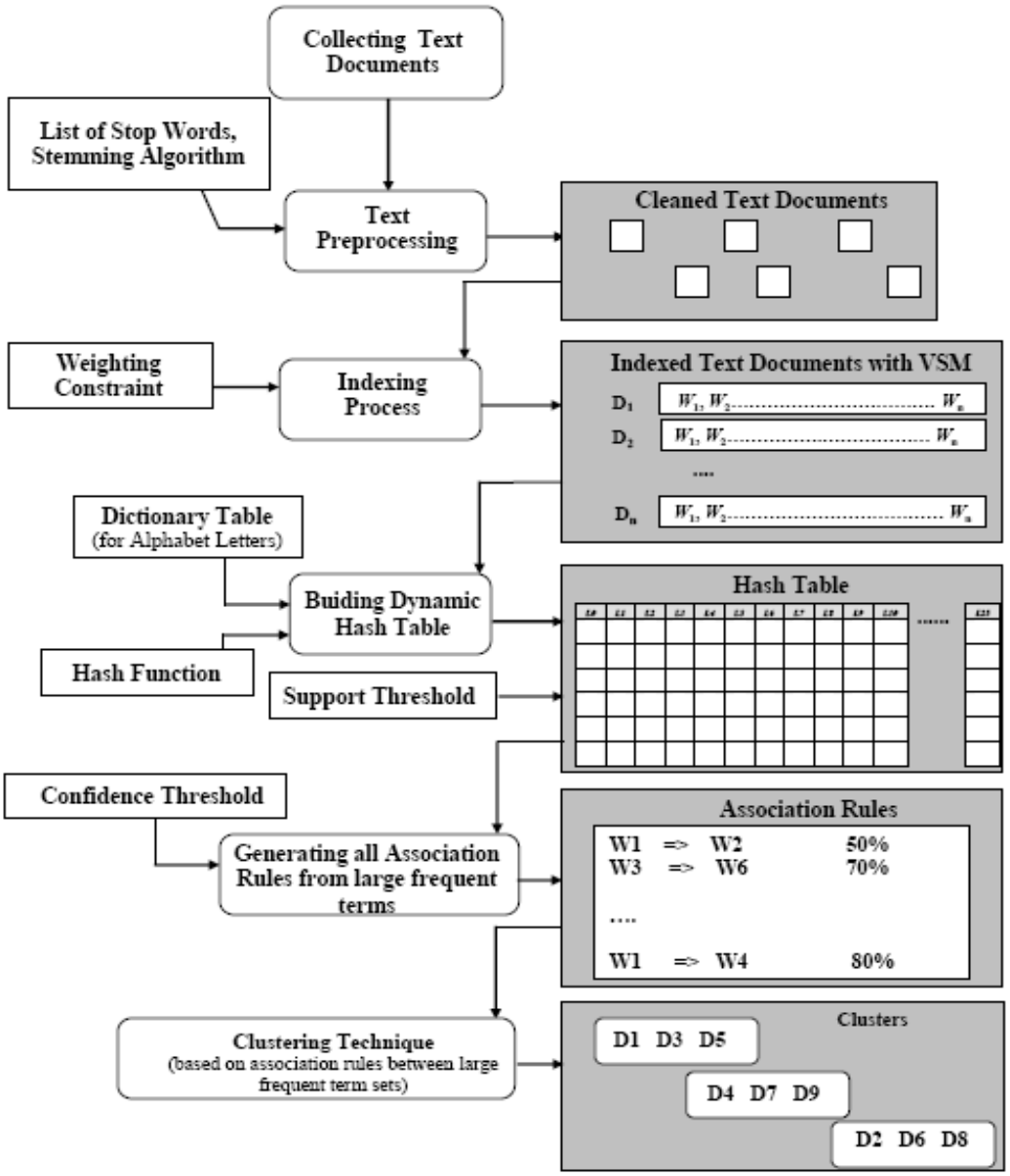
## II. RELATED WORK

As we have seen from the current systems, we cannot extract accurate information from large database. To retrieve such information from that large database user need to read all the database and extract the current information. There is very low possibility of getting the desired information .similarly it is not feasible and very time consuming. Earlier systems were also complicated for decision support as reading had to be done manually.
Thus system requires automatic analyzing of large amount of data which should be feasible and less time consuming. It should also possess a property of extracting the desired information from the database. So we are creating a Association Rule Extraction System.

## III.PROPOSED ALGORITHM

In our application web pages are taken as input. Then the textual data present in webpage are converted into tokens. The tokens are compared with the stop wards list, then unwanted tokens are filtered the afterward remaining tokens are proceed through porter stemmer through which root words tokens are taken as next output and then association rule is applied on remaining tokens. Here, when we provide the Input as any webpage to our System, we get the desired output accordingly.

## IV. PSEUDO CODE





DOI 10.17148/IJARCCE.2016.52105

## V. SIMULATION RESULTS

We presented unique approach for clustering in this report. We used HARMT algorithm for generating association rule mining. This algorithm scans documents one time and stores all term sets in dynamic hash table .Our approaches out to be performs in terms of efficiency, accuracy and scalability. We are currently in the stage of implementation and optimizing the performance.

## VI. CONCLUSION AND FUTURE WORK

There is problem of analyzing huge amount of data. So the question is who is able to read and analyze it? In this context, manual analysis and effective extraction of useful information are not possible. We think the solution is that it is relevant to provide automatic tools for analyzing large textual collections by automatically find relevant Information.

Text mining is an Increasingly Important research field because of the necessity of knowledge from the enormous number of text documents available, We describe use of Association Rules in TM Association rules highlight correlation between features In the texts, e.g. keywords a Word is selected as a keyword It' It does not appear In a predefined stop-words 1ist Moreover, association rules are easy to understand and to Interpret for an analyst or may be for a normal user. We've described system for automatically extracting association rules from Web page documents.

## REFERENCES

[1] Vaishali bhujade, N. J. Janwe, 2011 "Knowledge discovery in text mining technique using association rules extraction", International conference on computational intelligence and communication system.
[2] Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol 2 (6), 2011 "A Comparative Study of Stemming Algorithms", The Maharaja Sayajirao University of Baroda Vadodara, Gujarat, India
[3] Atika Mustafa, Ali Akbar, and Ahmer Sultan, 2009, "Knowledge discovery in text mining technique using association rules extraction, National University of Computer and Emerging Sciences-FAST, Karachi Pakistan
[4] Hany Mahgoub, 2008, "Mining Association Rules from Unstructured Documents", "World Academy of Science, Engineering and Technology"
[5] Qiankun Zhao Nanyang and Sourav S.Bhowmick, Singapore,2003"Association Rule Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore

## BIOGRAPHIES

**Prof. Deveshree Wankhede** is Assistant Professor of Computer Department, Shivajirao S. Jondhale C.O.E Dombivli, Mumbai University.

**Amol Bhaiyasaheb Patil** is a Student of Computer Engineering, Shivajirao S. Jondhale C.O.E Dombivli, Mumbai University.

**Ajit Shrawan Patil** is a Student of Computer Engineering, Shivajirao S. Jondhale C.O.E Dombivli, Mumbai University.

**Shahezad Arif Akhtar** is a Student of Computer Engineering, Shivajirao S. Jondhale C.O.E Dombivli, Mumbai University.