INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# A SURVEY ON TEXT MINING TECHNIQUES

[1]**K.THILAGAVATHI**, MCA,[2] **V.SHANMUGA PRIYA**, MCA, M.PHIL
E-MAIL- k.thilagamca@gmail.com
E-MAIL- vspriyarjpm@gmail.com

Author correspondence: mobile no:962969893,e-mail:kthilagamca@gmail.com

## Abstract

Text mining is a technique to find meaningful patterns from the available text documents. The pattern discovery from the text and document organization of document is a well-known problem in data mining. Analysis of text content and categorization of the documents is a complex task of data mining. In order to find an efficient and effective technique for text categorization, various techniques of text categorization and classification is recently developed. Some of them are supervised and some of them unsupervised manner of document arrangement. This presented paper discusses different method of text categorization and cluster analysis or text documents. In addition of that a new text mining technique is proposed for future implementation.

**Keywords:** Text mining, classification, cluster analysis, survey, Information Retrieval and Indexing Techniques**.**

## 1. INTRODUCTION

Now in these days, due to computational automation various different text document sources are available. Extraction of patterns and arranging the text document is a key goal of text mining technique development. Text mining is related to data mining, except that data mining tools are considered to handle structured data, but text mining can work with formless or semi- structured data sets. The application of text mining is very popular in emails analysis, digital libraries and others. The text mining techniques starts with collection of text documents (text repository), than a text mining tool for pre-processing is applied. The preprocessing technique clean and format the data, additionally that is responsible for extracting the meaningful features from these documents. In next step the text mining techniques such as clustering or classification algorithm is taken place to arrange the documents.

## 2. RECENT STUDIES

This section of the paper explores recent efforts and contributions on text mining techniques. Therefore a number of research article and research papers and their contributions are placed in this section. Many data mining techniques have been planned for mining valuable patterns in text documents. However, how to successfully use and update exposed patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the troubles of polysemy and synonymy. This paper presents an inventive and valuable pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to advance the effectiveness of using and updating discovered patterns for finding

appropriate and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

The "helpfulness" characteristic of online user reviews helps consumers deal with information overloads and facilitates decision-making. However, many online user reviews require sufficient helpfulness votes for other users to assess their true helpfulness level. Text mining techniques are employed to remove semantic characteristics from review texts. Our findings also advise that reviews with strong opinions receive more kindness votes than those with mixed or neutral. This paper sheds light on the considerate of online users' helpfulness voting activities and the design of a enhanced helpfulness voting mechanism for online user review systems.

## 3. TEXT MINING TECHNIQUES

There are different kinds of techniques available by which the text pattern analysis and mining is performed. Some of the essential techniques are discussed in this section.

### 3.1. *Information Extraction*

A starting point for computers to examine unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships with in text. The software infers the relations between all the identified people, places, and time to deliver the user with significant information. This technology can be very helpful when dealing with large volumes of text. Traditional data mining assumes that the information to be "mined" is previously in the form of a relational database. Unfortunately, for many applications, electronic information is only obtainable in the form of free natural language documents rather than structured databases. Since IE addresses the difficulty of transforming a corpus of textual documents into a extra structured database, the database constructed by an IE module can be provided to the KDD module for advance mining of knowledge.

### 3.2. *Topic Tracking*

A topic tracking system mechanism by custody of user profiles and, based on the documents the user views, guess other documents of interest to the user. Yahoo offers (www.alerts.yahoo.com) free topic tracking tool that permits users to choose keywords and informs them when news relating to those topics becomes existing. Topic tracking methodology has its own limitations, however. For example, if a user sets up an alert for "text mining", we will receive numerous news stories on mining for minerals, and very few that are really on text mining. Some of the improved text mining tools let users select specific categories of interest or the software routinely can even infer the user's concern based on his/her reading history and click-through information.

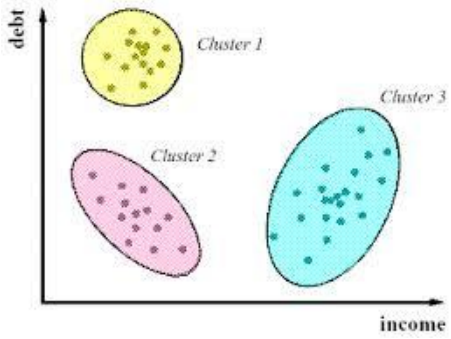
.

### 3.3. *Categorization*

Categorizations engage identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often delight the document as a "bag of words. "Rather, categorization only calculates words that emerge and, from the counts, identifies the main topics that the document covers. Categorization often relies on a vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, synonyms, and related terms. Categorization utensils normally have a technique for grade the documents in order of which documents have the most content on a specific topic.

### 3.4. *Clustering*

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results. A basic clustering algorithm generates a vector of topics for each document and determines the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of Management information systems, which may contain thousands of documents.




### 3.5. *Concept Linkage*

Concept linkage tools attach related documents by identifying their commonly-shared idea and help users find information that they perhaps wouldn't have establish using conventional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations.
But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

### 3.6. *Information Visualization*

Visual text mining or information visualization puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. DocMiner Is a tool that shows mappings of large amounts of text, allowing the user to visually analyze the Content. The user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them, with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own.

### 3.7. *Association Rule Mining*

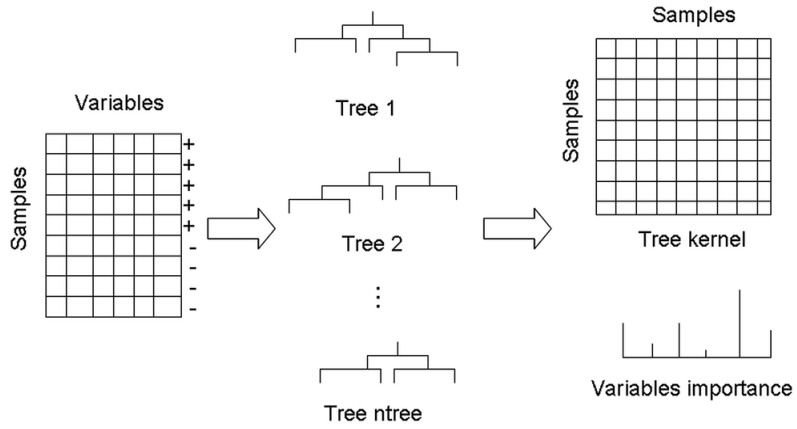
Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. ARM refers to the discovery of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently occur. Similar to the idea of correlation analysis (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may Contain two or more variables. This section provides the overview of text mining techniques and methodologies by which suitably text data becomes classifiable in next we discuss the data mining algorithms that are frequently consumed in the text mining and classification tasks.

## 4. TEXT MINING ALGORITHMS

There are various algorithms of data mining is available for efficient classification and Categorization. The discussion about whole methods and technique are not much feasible here therefore a little overview is proving in this section.

### 4.1. *k nearest neighbor*

In the text mining domain the k nearest neighbor algorithm is a classical and frequently used technique. In order to find a query text k nearest neighbour classifier is outperforms. This method estimates the distance between two strings for comparison and classify the text on the basis of distance. Where x and y represents the data instances and d is distance between x and y. The main advantage of this algorithm is high accurate classification. On the other hand the major disadvantage is resources consumption such as memory and time.




### 4.2. *Support vector machine*

This approach is a one of most efficient and accurate classification algorithm. In this approach concept using hyper-plans and dimension estimation based technique are used to discover or classify the data. The main advantage of this algorithm is to achieve high accurate classification results. But that is quite complex to implement.

### 4.3. *Bayesian classifier*

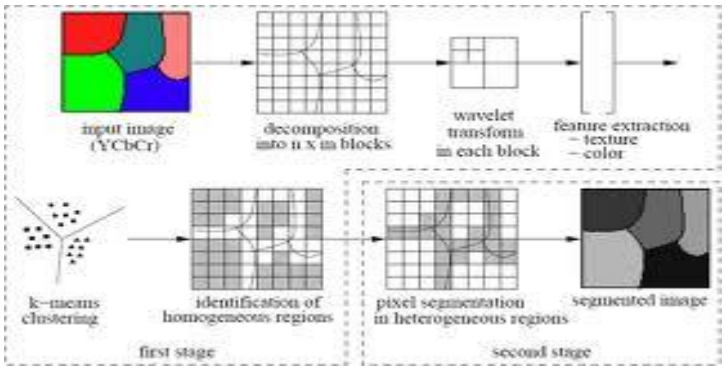
That a probability based classification technique that is uses the word probability to classify the text data. In this classification scheme based on previous text and patterns data is evaluated and the class possibility is measured. That is some time slow learning classifier additionally that do not produces the more accurate results.

| Outlook | P | N |
|---|---|---|
| sunny | 2/9 | 3/5 |
| overcast | 4/9 | 0 |
| rain | 3/9 | 2/5 |
| Tempreature | | |
| hot | 2/9 | 2/5 |
| mild | 4/9 | 2/5 |
| cool | 3/9 | 1/5 |

| Humidity | P | N |
|---|---|---|
| high | 3/9 | 4/5 |
| normal | 6/9 | 1/5 |
| | | |
| Windy | | |
| true | 3/9 | 3/5 |
| false | 6/9 | 2/5 |

### 4.4. K-mean clustering

This technique is also a classical approach of text categorization. That uses the distance function as k nearest neighbour classifier to cluster data. That is an efficient method of text mining in order to preserve the resources, but accuracy of this cluster approach is susceptible due to initial cluster center selection process. In addition of that hierarchical schemes of text categorization is available which are not much efficient for cluster formation or categorization but comparative accuracy is much reliable than k-mean clustering.




## 5. INFORMATION RETRIEVAL

Information retrieval is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance. Due to the abundance of text information, information retrieval has found many applications.

There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines. A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some adhoc information need, such as finding information to buy a used car. When a user has a long-term information need , a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process is called information filtering, and the corresponding systems are often called filtering systems or recommender systems. From a technical viewpoint, however, search and filtering share many common techniques. Below we briefly discuss the major techniques in information retrieval with a focus on search techniques.

## 6. MEASURES FOR TEXT RETRIEVAL

The set of documents relevant to a query be denoted as *{*Relevant*}*, and the set of documents retrieved be denoted as {Retrieved}. The set of documents that are both relevant and retrieved is denoted as {Relevant} ∩ {Retrieved}, as shown in the Venn diagram of Figure 1. There are two basic measures for assessing the quality of text retrieval.

Precision: This is the percentage of retrieved documents that are in fact relevant to the query . It is formally defined as

Precision=ı{Relevant}∩{Retrieved}ı/ı{Retrieved}ıRecall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

Recall=ı{Relevant}∩{Retrieved}ı/ı{Relevant}ı An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used tradeoff is the F-score, which is defined as

The harmonic mean of recall and precision F-score =recall×precision/(recall+ precision)/2 The harmonic mean discourages a system that sacrifices one measure for another too drastically

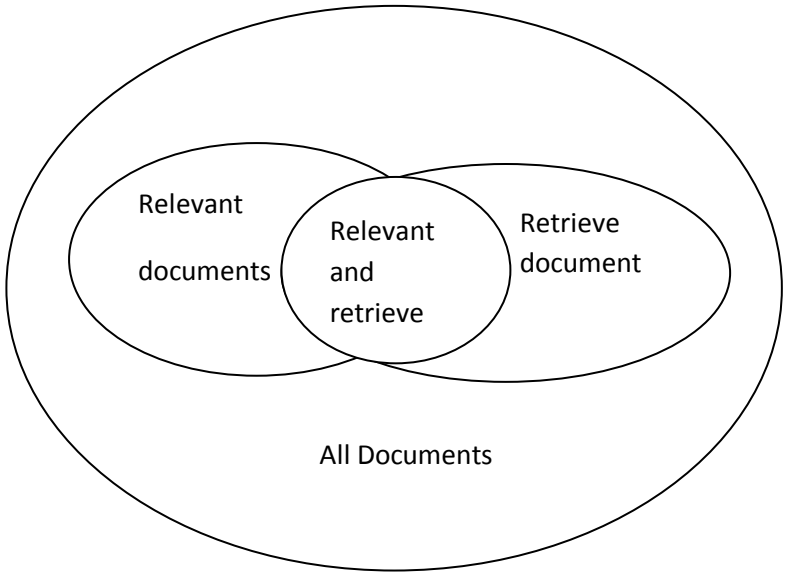



**Figure 1.** Relationship between the set of relevant documents and the set of retrieved documents.

Precision, recall, and F-score are the basic measures of a retrieved set of documents. These three measures are not directly useful for comparing two ranked lists of documents because they are not sensitive to the internal ranking of the documents in a retrieved set. In order to measure the quality of a ranked list of documents, it is common to compute an average of precisions at all the ranks where a new relevant document is returned. It is also common to plot a graph of precisions at many different levels of recall; a higher curve represents a better-quality information retrieval system. For more details about these measures, readers may consult an information retrieval textbook.

## 7. TEXT INDEXING TECHNIQUES

There are several popular text retrieval indexing techniques, including inverted indices and signature files. An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: document table and term table, where document table consists of a set of document records, each containing two fields: doc id and posting list, where posting list is a list of terms that occur in the document, sorted according to some relevance measure. term table consists of a set of term records, each containing two fields: term id and posting list, where posting list specifies a list of document identifiers in which the term appears. With such organization, it is easy to answer queries like “Find all of the documents associated with a given set of terms,” or “Find all of the terms associated with a given set of documents.” For example, to find all of the documents associated with a set of terms, we can first find a list of document identifiers in term table for each term, and then intersect them to obtain the set of relevant documents. Inverted indices are widely used in industry. They are easy to implement. The posting lists could be rather long, making the storage requirement quite large. They are easy to implement, but are not satisfactory at handling synonymy (where two very different words can have the same meaning) and polysemy

(where an individual word may have many meanings). A signature file is a file that stores a signature record for each document in the database. Each signature has a fixed size of *b* bits representing terms. A simple encoding scheme goes as follows. Each bit of a document signature is initialized to 0.A bit is set to 1 if the term it represents appears in the document. A signature *S*1 matches another signature *S*2 if each bit that is set in signature *S*2 is also set in *S*1. Since there are usually more terms than available bits, multiple terms may be mapped into the same bit. Such multiple-to one mapping make the search expensive because a document that matches the signature of a query does not necessarily contain the set of keywords of the query. The document has to be retrieved, parsed, stemmed, and checked. Improvements can be made by first performing frequency analysis, stemming, and by filtering stop words, and then using a hashing technique and superimposed coding technique to encode the list of terms into bit representation. Nevertheless, the problem of multiple-to-one mappings still exists, which is the major disadvantage of this approach. Readers can refer to for more detailed discussion of indexing techniques, including how to compress an index.

## 8. QUERY PROCESSING TECHNIQUES

Once an inverted index is created for a document collection, a retrieval system can answer a keyword query quickly by looking up which documents contain the query keywords. Specifically, we will maintain a score accumulator for each document and update these accumulators as we go through each query term. For each query term, we will fetch all of the documents that match the term and increase their scores. More sophisticated query processing techniques are discussed in .When examples of relevant documents are available, the system can learn from such examples to improve retrieval performance. This is called relevance feedback and has proven to be effective in improving retrieval performance. When we do not have such relevant examples, a system can assume the top few retrieved documents in some initial retrieval results to be relevant and extract more related keywords to expand a query. Such feedback is called pseudo-feedback or blind feedback and is essentially a process of mining useful keywords from the top retrieved documents. Pseudo-feedback also often leads to improved retrieval performance. One major limitation of many existing retrieval methods is that they are based on exact keyword matching. However, due to the complexity of natural languages, keyword based retrieval can encounter two major difficulties. The first is the synonymy problem: two words with identical or similar meanings may have very different surface forms. For example, a user's query may use the word "automobile," but a relevant document may use "vehicle" instead of "automobile." The second is the polysemy problem: the same keyword, such as mining, or Java, may mean different things in different contexts.

## 9. INFORMATION EXTRACTION

The general purpose of Knowledge Discovery is to "extract implicit, previously unknown, and potentially useful information from data". Information Extraction IE mainly deals with identifying words or feature terms from within a textual file. Feature terms can be defined as those which are directly related to the domain

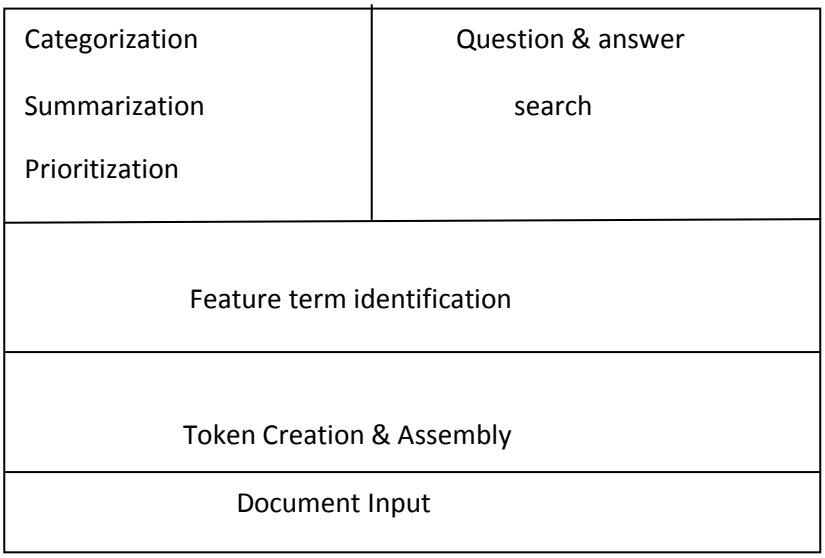



Figure 2. A layered model of the Text Mining Application

These are the terms which can be recognized by the tool. In order to perform this function optimally, we had to look into few more aspects which are as follows:

## 9. STEMMING

Stemming refers to identifying the root of a certain word. There is basically two types of stemming techniques, one is inflectional and other is derivational. Derivational stemming can create a new word from an existing word, sometimes by simply changing grammatical category (for example, changing a noun to a verb). The type of stemming we were able to implement is called Inflectional Stemming. A commonly used algorithms is the 'Porter's Algorithm' for stemming. When the normalization is confined to regularizing grammatical variants such as singular/plural or past/present, it is referred to inflectional stemming .To minimize the effects of inflection and morphological variations of Words (stemming), our approach has pre-processed each word using a provided version of the Porter stemming algorithm with a few changes towards the end in which we have omitted some cases.e.g. apply – applied – applies

print – printing – prints – printed

In both the cases, all words of the first example will be treated as 'apply' and all words of the second example will be treated as 'print'

.

## 10. DOMAIN DICTIONARY

In order to develop tools of this sort, it is essential to provide them with a knowledge base. A collective set of all the 'feature terms' is the Domain dictionary (our source was www.webopedia.com). The structure of the Domain dictionary which we implemented consisted of three levels in the hierarchy. Namely, Parent Category, Sub-category and word. Parent categories define the main category under which any sub-category or word falls. A parent category will be unique on its level in the hierarchy. Sub-categories will belong to a certain parent category and each subcategory will consist of all the words associated with it. As an example, consider the following

Table 1.Structure of the Domain Dictionary

| Parent Category | Sub Category | Words |
|---|---|---|
| Hard ware | Data storage | Grabber |
| | Input devices<br>Modems | Light pen<br>Joy stick |
| | Mother boards | Contact image sensor |
| | Networking | Digital camera |

Table 1 is an example that shows how we identify words Which belong to the Parent Category 'Hardware' and Subcategory 'Input Devices'.

## Exclusion List

A lot of words in a text file can be treated as unwanted noise. To eliminate these, we devised a separate file which includes all such words. These include words such as the, a, an, if, off, on etc.

## Research Directions

With abundant literature published in research into frequent pattern mining, one may wonder whether we have solved most of the critical problems related to frequent pattern mining so that the solutions provided are good enough for most of the data mining tasks. However, based on our view, there are still several critical research problems that need to be solved before frequent pattern mining can become a cornerstone approach in data mining applications. First, the most focused and extensively studied topic in frequent pattern mining is perhaps scalable mining methods.

The set of frequent patterns derived by most of the current pattern mining methods is too huge for effective usage. There are proposals on reduction of such a huge set, including closed patterns, maximal patterns, approximate patterns, condensed pattern bases, representative patterns, clustered patterns, and discriminative frequent patterns, as introduced in the previous sections. However, it is still not clear what kind of patterns will give us satisfactory pattern sets in both compactness and representative quality for a particular application, and whether we can mine such patterns directly and efficiently. Much research is still needed to substantially reduce the size of derived pattern sets and enhance the quality of retained patterns. Frequent pattern mining: current status and future directions. Second, although we have efficient methods for mining precise and complete set of frequent patterns, approximate frequent patterns could be the best choice in many applications. For example, in the analysis of DNA or protein sequences, one would like to find long sequence patterns that approximately match the sequences in biological entities, similar to BLAST. Much research is still needed to make such mining more effective than the currently available tools in bioinformatics. Third, to make frequent pattern mining an essential task in data mining, much research is needed to further develop pattern based mining methods. For example, classification is an essential task in data mining. Fourth, we need mechanisms for deep understanding and interpretation of patterns, e.g. semantic annotation for frequent patterns, and contextual analysis of frequent patterns. The main research work on pattern analysis has been focused on pattern composition (e.g., the set of items in item-set patterns) and frequency.

The semantic of a frequent pattern includes deeper information: what is the meaning of the pattern; what are the synonym patterns; and what are the typical transactions that this pattern resides? In many cases, frequent patterns are mined from certain data sets which also contain structural information. Finally, applications often raise new research issues and bring deep insight on the strength and weakness of an existing solution. This is also true for frequent pattern mining. On one side, it is important to go to the core part of pattern mining algorithms, and analyze the theoretical properties of different solutions. On the other side, although we only cover a small subset of applications in this article, frequent pattern mining has claimed a broad spectrum of applications and demonstrated its strength at solving some problems. Much work is needed to explore new applications of frequent pattern mining. For example, bioinformatics has raised a lot of challenging problems, and we believe frequent pattern mining may contribute a good deal to it with further research efforts.

## CONCLUSIONS

In this paper various techniques and methods are discussed for efficient and accurate text mining. In addition of that the efficient algorithms are also learned. Due to observation a promising approach is obtained given in . According to the analyzed methods an improvement over this is suggested. In near future the proposed technique is implemented using JAVA technology and the comparative results are provided.

## ACKNOWLEDGMENT

I wish to thanks who directly and indirectly contribute in paper, First and foremost, I would like to thank Prof. V.Shanmuga priya for his most support and encouragement. She kindly read my paper and offered valuable details and provides guidelines. Second, I would like to thanks all the authors whose paper i refer for their direct and indirect support to complete my work.

## REFERENCES

[1] Vishal Gupt a, Gurpreet S. Lehal, " A Survey of Text Mining Techniques and Applicat ions", JOURNAL OF EMERGING T ECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009
[2] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effect ive P att ern Discovery for Text Mining", IEEE Transact ions on Knowledge and DataEngineering. C Copyright 2010 IEEE
[3] Qing Cao, Wenjing Duan, Qiwei Gan, "Exploring det erminant s of vot ing for t he "helpfulness" of online user reviews: A t ext mining approach', 0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009
[4] Hamid Mousavi, Shi Gao, Carlo Zaniolo, " IBminer: A Text Mining Tool for Const ruct ing and Populat ing InfoB[ox Databases and Knowledge Bases", Proceedings of the VLDB Endowment , Vol. 6, No. 12, Copyright 2013 VLDB Endowment 21508097/13/10...$
[5] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky, "HierarchicalTopics: Visually Exploring Large Text Collect ions Using Topic Hierarchies", IEEE T RANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 19, NO. 12,DECEMBER 2013
[6] Liwei Wei, Bo Wei, Bin Wang, "Text Classificat ion Using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58,