Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2014; 2(1):80-84 **ISSN 2347-9523 (Print)**
©Scholars Academic and Scientific Publisher
(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Improving the Effective Pattern Discovery for Text Mining

[1]Vijayakumar T, [2] Priya R

[1]Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode DT, Tamilnadu, India

[2] Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode Tamilnadu, India

***Corresponding author**
Priya R
Email: priyar.se12@bitsathy.ac.in

Abstract: Huge data mining techniques have been used for mining useful pattern in text document. Text mining can be used to extract the data in document. It is effectively use and update the discovered pattern; still the research is not yet completed. The existing approach is term based approach; they suffer the problem of polysemy and synonymy. In the past years, people have used pattern based approaches for hypothesis which perform better than the term based ones, but many of the experiments do not support this hypothesis. In this paper present a new idea about the effective pattern discovery technique which involved the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and useful information.

Keywords: Text mining, Text classification, pattern mining, pattern evolving, information filtering

INTRODUCTION

Due to high range of data available in past few years, the data mining techniques are used for extracting useful information and knowledge [1]. For example, market analysis and business management it is used to extracting the large amount of data. The data mining techniques available to perform effective pattern mining and text mining are: The association rule mining. The frequent item set mining, The sequential pattern mining, The maximum pattern mining and The closed pattern mining. The data mining techniques used to find large number of pattern to find effectively use and update.

Text mining is used to extracting information from the text document [2, 5]. It is used to find help the user want they want. First, information retrieval provided term based method to solve probabilistic models and support vector machine (it is used to find filtering the data model). The advantages of term based method to find efficient computational problem and term weight. But the term based method suffer the problem of polysemy (the word has multiple meanings e.g. machine) and synonymy (multiple words having the same meaning e.g. object).

In the past years, people have used pattern based approaches for hypothesis which perform better than the term based ones, but many of the experiments do not support this hypothesis. They have less ambiguous and more discriminative problem. They include the performance like 1) inferior statistical properties to terms, 2) low frequency of occurrence, and 3) numbers of redundant and noisy phrases among them.

In the sequence, mining used to discover set of attributes from the large of database. For example bookstore it is used to find frequent number of item occurs.

The pattern based approach includes two fundamental issues: low frequency and misinterpretation problem. It overcome problem of association rule mining (e.g., support and confidence).

Related Work

These work is based on the term frequency (tf) and inverse document frequency (df) and the weighting scheme is used for text representation and they improve the performance of the data.

Information filtering is used for relevant and irrelevant document in the text classification. The techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining.

The data mining techniques used to find large number of pattern to find effectively use and update.

MODULES IN PATTERN DISCOVERY

The modules describe how the document is evaluated in the effective pattern discovery.

Modules used are

A. Dataset Collection and Tokenizing

B. Text Preprocessing
C. Term Selection model
D. Frequent Pattern Analysis
E. Pattern Clustering
F. Sequence Analysis
G. Pattern Evaluation and Reports

Dataset collection and Tokenizing

The first step is extraction of data from the given dataset. The sentence which contains set of text will be extracted for the analysis. Identifying data's and splitting into terms is the major process [5].

The document with the all kind of words and symbols are given as the input to the system. This document is then sent for the text mining process. This is here done for the segregation of the important words from the document. This process has been carried out by selecting the each important word and removing the unwanted word and by replacing the tabs, the other non text characters by a single space.

The common existing words from all the documents are collected. These words are here collected by merging the documents together. This collection word from the all documents by merging them is called as dictionary of document collection.

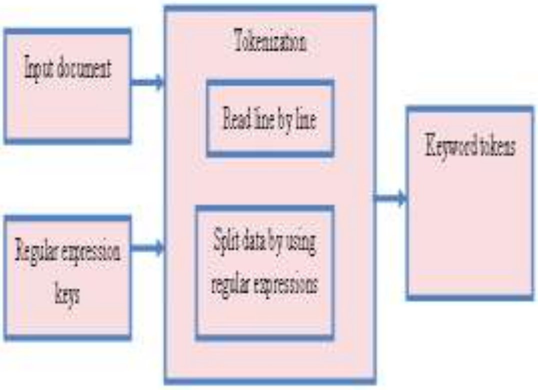


Fig.1. Data set collection and tokenization

Text Preprocessing

Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. The Stop-words are used to extracting the words in the document words. The frequent word does not carry any information. The unwanted words are removed and to create a new dataset [4,6].

In text preprocessing includes Word Disambiguation. Each words or phrases will have their ambiguity in their meanings. This problem of ambiguity could be resolved using this text preprocessing [3,5]. Stemming process is to reduce the normalization of document. For example the stemming algorithm take the acceptable data of walk are walk, walking, walked.

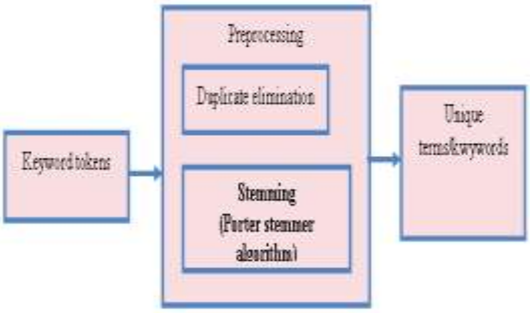


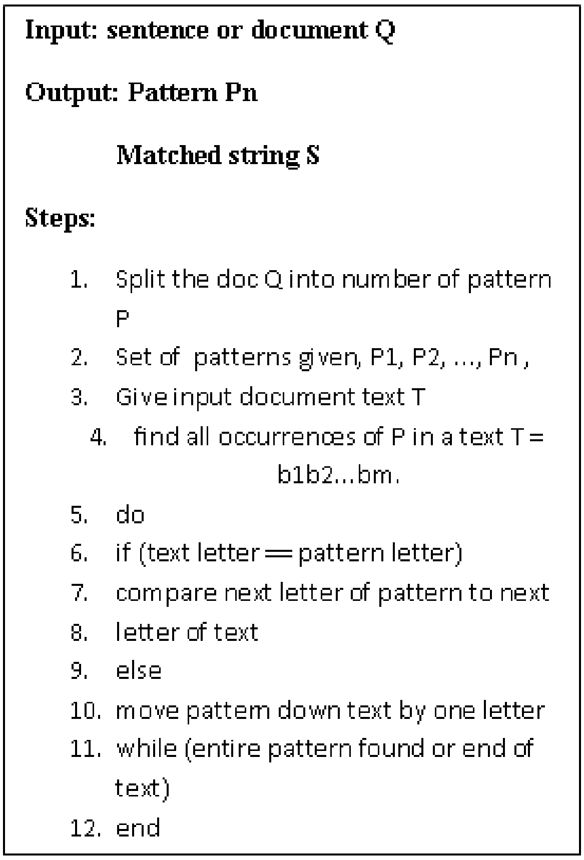
Fig.2 Text preprocessing

Term Selection Model

To decrease the number of words that should be used in the document is done by indexing or keyword selection algorithms. In this paper, the documents are described; these documents are described by using the selected keyword.

Based on the calculated entropy value, the keywords are extracted from the document. This type extracting the keywords by using the entropy value calculation is the simple method. Here the module represents different type of steps to extract term selection.

The unique terms should be identified for the term selection; the identification is done by extracting all texts and eliminating the unwanted words. This type process leads to unique identification of the terms. Pattern matching concepts has been applied in this module. For unique term selection patter matching steps has been implemented.



Input: sentence or document Q

Output: Pattern Pn

Matched string S

Steps:

```
1.  Split the doc Q into number of pattern
    P
2.  Set of  patterns given, P1, P2, ..., Pn ,
3.  Give input document text T
4.  find all occurrences of P in a text T =
                      b1b2...bm.
5.  do
6.  if (text letter == pattern letter)
7.  compare next letter of pattern to next
8.  letter of text
9.  else
10. move pattern down text by one letter
11. while (entire pattern found or end of
    text)
12. end
```

The main objective of document indexing is to increase the efficiency by extracting from the resulting document a selected set of terms to be used for indexing the document. Document indexing is the three step process:

Step 1: The set of keywords are selected from the corpus database. These keywords are here based on the documents.
Step 2: The each input document here is selected and scanned, from that the needed keywords are selected.
Step 3: The weight for each keyword is here assigned to each of the keyword.
Step 4: Then the whole document is transformed into the vector of keywords.

The frequency of occurrence of the term in the particular document and the number of documents is used.

Frequent Pattern analysis

This module helps to make the frequent pattern count and text snippets value with combine to making the final values. The matching is totally based on the correlation between the words. This process is an invisible this automatically occurs while the extraction happened.

Pattern clustering

Clustering module makes the output from the pattern extraction in to similar items in a set of groups. This process helps to avoid the unrelated items from the resultant value.

Clustering algorithm groups the items from page count as well from the text snippets. The pattern extraction omits some unrelated items and the clustering will removes all unwanted item from the result.

Sequence analysis

A sequential pattern analysis module is used to frequent pattern if its relative support (or absolute support) _ min sup, a minimum support.

Here minimum support and confidence will be calculated. The property of closed patterns can be used to define closed sequential patterns.

A frequent sequential pattern will be analyzed by the identification of closed pattern and support and confidence of every data.

Pattern evaluation and reports

After the clustering and sequence identification of data set, the system will apply priority measure for accurate pattern. It also consists of the filtering process called ranking based on the sequence as well as frequent. This module analysis the existing clustered group and also extracts data from semantic library.

Porter Stemmer Algorithm

The preprocessing process includes the stemming process, which eliminates unnecessary keys. All stemming algorithms can be roughly classified as affix removing, statistical and mixed. Affix removal stemmers apply set of transformation rules to each word, trying to cut off known prefixes or suffixes.

Porter stemmer utilizes suffix stripping techniques rather than prefix methods. The porter stemmer Algorithm dates from 1980.

Step 1: Gets rid of plurals and -ed or -ing suffixes
Step 2: When the vowel is identified, then turns the terminal “y” to “i”.
Step 3: Maps double suffixes to single ones:- ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off the words such as -ant, -ence, etc.
Step 6: Removes a final –e

The above steps represent the process and elimination of porter stemmer algorithm.. The importance of the stemmer algorithm is, it reduces the difficulties of data classification when the training data’s are insufficient. This effectively eliminates the suffix words such as ‘ed’, ‘ing’ etc.,

The pseudo code for the above algorithm is represented below

```
1.  String s
2.  Split string s and stored into s[].
3.  For each word in s[]
4.  If S[i].text end with “ed”
5.  Remove the two keys from the word.
6.  Store s1[i].
7.  Else If S[i].text end with “ing”
8.  Remove the three keys from the word.
9.  Store s1[i].
10. else if ends("s") || ends("ss")
11. do step 9
```

RESULTS

The support vector machines used to filter the positive and negative document. The documents are evaluated by the threshold value. Threshold (DP) = $min_{p\in DP}(\sum_{(t.w)\in\beta(p)} support(t))$ in these formula to remove the noisy negative document in the result.

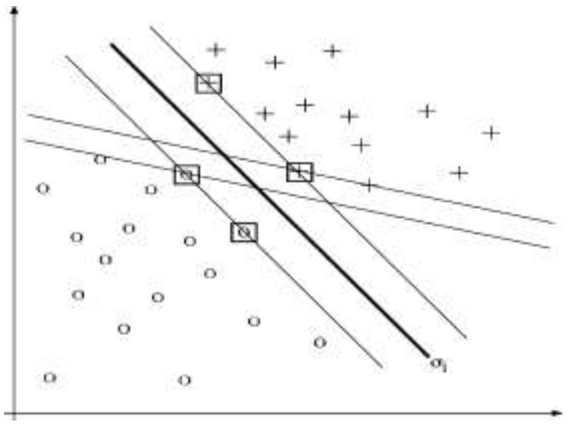


Fig 3: Positive and Negative Classifiers

In the support vector used to classify the document. The positive and negative training examples are here represented by the small circles and crosses.

The result is based on precision and recall to evaluating the document. First the document loaded in the database and each document are split into paragraph to remove the unwanted word and noise in the document.

The following figure describes about the average value taken to evaluate the graph.

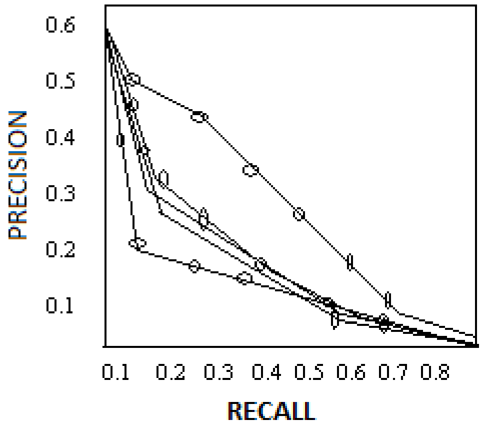


Fig. 4: The graph shows the precision and recall the relevant and irrelevant document

CONCLUSION

The data mining techniques includes association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. In these work, the problem of the low-frequency and misinterpretation problem of text mining have been overcome.

REFERENCES

1. Pal JK; Usefulness and applications of data mining in extracting information from

different perspectives. Annals of Library and Information Sciences, 2011; 58:7-16.

2. Li Y, Zhou X, Bruza P, Xu Y,. Lau RY; A Two-Stage Text Mining Model for Information Filtering, Proc. ACM 17th Conf. Information and Knowledge Management (CIKM '08), 2008;1023-1032.
3. Aas K, Eikvil L; Text Categorisation: A Survey. Technical Report Raport NR 941, Norwegian Computing Center, 1999.
4. Sebastiani F; Machine Learning in Automated Text Categorization, ACM Computing Surveys, 2002; 34(1):1-47.
5. Ning Zhong N, Li Y, Wu ST; Effective Pattern Discovery for Text Mining, 2012; 24.
6. Salton G, Buckley C; Term-Weighting Approaches in Automatic Text Retrieval, Information Processing and Management: An Int'l J., 1988; 24(5):513-523.