IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

A Combined Approach for Effective Text Mining using Node Clustering

Yogendra Singh Rajput[1], Priya Saxena[2]

Research Scholar, CSE, Sanghvi Innovative Academy, Indore, M.P, India[1]

Lecturer, CSE, Sanghvi Innovative Academy, Indore, M.P, India[2]

Abstract: Text Mining is an important step of knowledge discovery process. Text mining extracts hidden information from not-structured to semi-structured data. It is the discovery by automatically extracting information from different written resources. There is a two basic methods used in text mining which is term based and phrased based approach; both methods have their merits and demerits. We proposed a combined approached to mine effective text form text documents using advantageous of methods, term based and phrases based simultaneously. The experiment results show the efficiency of the work.

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

1. INTRODUCTION

Data mining technology helps to extract useful information from various databases. Data warehouses are good for only numerical solution but unsuccessful when it came to textual information. As text mining is extraction of useful information from text data it is also known as text data mining or knowledge discovery from textual databases. It is challenging issue to find accurate knowledge in text documents to help users to find what they want.

Text mining process starts with a document collection from various resources. Text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

Text mining process is as shown in following fig.1

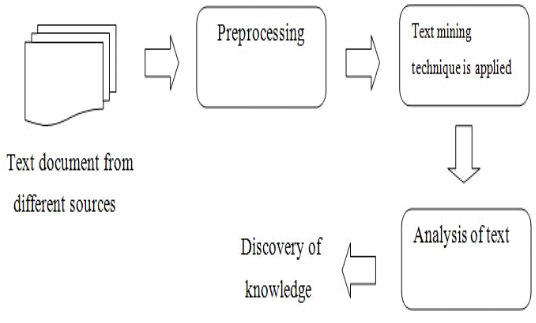

Text document from different sources
Preprocessing
Text mining technique is applied
Analysis of text
Discovery of knowledge

Fig. 1 Text mining process

In this paper we focus on two methods which is term based and phrased based.

Term Based Method

Term in document is word having semantic meaning. In term based method document is analyzed on the basis of term and has advantages of efficient computational performance as well as mature theories for term weighting. These techniques are emerged over the last couple of decades from the information retrieval and machine learning communities. Term based methods suffer from the problems of polysemy and synonymy[1]. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods to solve this challenge.

Phrase Based Method

Phrase carries more semantics like information and is less ambiguous. In phrase based method document is analyzed on phrase basis as phrases are less ambiguous and more discriminative than individual terms[2]. The likely reasons for the daunting performance include:

Phrases have inferior statistical properties to terms,
They have low frequency of occurrence, and
Large numbers of redundant and noisy phrases are present among them.
Rest of this paper presents related work, proposed work, implementation & result, and conclusion.

2. TECHNIQUES USED IN TEXT MINING

To teach computers how to analyze, understand and generate text, technologies are produced by natural language processing. The technologies like information extraction, summarization, categorization, clustering and

DOI 10.17148/IJARCCE.2016.5483

information visualization, are used in the text mining process. In the following sections we will discuss each of these technologies and the role that they play in text mining. The types of situations where each technology may be useful in order to help users are also discussed.

2.1 Information Extraction

Information extraction is initial step for computer to analyze unstructured text by identifying key phrases and relationships within text. To do this task process of pattern matching is used to look for predefined sequences in text. Information extraction task includes tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment. Firstly phrases and sentences are parsed and semantically interpreted then required pieces of information entered into the database. General information extraction process is as shown in fig.2. The most accurate information extraction systems involve handcrafted language processing modules, substantial progress has been made in applying data mining techniques to a number of these steps. This technology can be very useful when dealing with large volumes of text. For many applications challenging is electronic information is in the form of free natural language documents rather than structured databases like relational databases. Information extraction solves this problem of transforming a corpus of textual documents into a more structured database. For further mining of knowledge database constructed by an information extraction module can be provided to the KDD module.

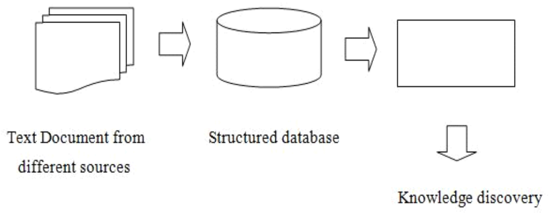


Fig.2 Information Extraction

2.2 Categorization

Categorization automatically assigns one or more category to free text document. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction, and classification [3][4]. The goal of categorization is to train classifier on the basis of known examples and then unknown examples are categorized automatically. Statistical classification techniques like Naïve Bayesian classifier, Nearest Neighbour classifier, Decision Tree, and Support Vector Machines can be used to categorize text.

2.3 Clustering

Clustering method can be used in order to find groups of documents with similar content. The outcome of clustering is typically a partition called clusters P and each cluster consists of a number of documents d. The contents of the documents within one cluster are more similar and between the clusters more dissimilar then the quality of clustering is considered better. Even though clustering technique used to group similar documents it differs from categorization because in clustering documents are clustered on the fly instead of use of predefined topics. As documents can appear in multiple subtopics clustering ensures that a useful document will not be omitted from search results [7]. In data mining K-means is frequently used clustering algorithm; in text mining field also it obtains good results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. The organization of management information systems makes use of clustering technology as organizational database contain thousands of documents.

2.4 Visualization

In text mining visualization methods can improve and simplify the discovery of relevant information. To represent individual documents or groups of documents text flags are used to show document category and to show density colours are used. Visual text mining puts large textual sources in a visual hierarchy. The user can interact with the document by zooming and scaling. Information visualization is applicable to government to identify terrorist networks or to find information about crimes. Following fig.3 shows steps involved in visualization process.

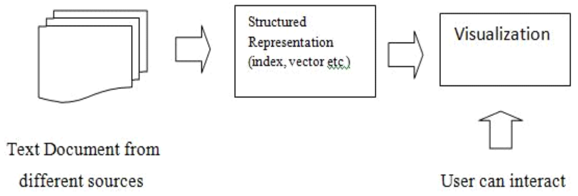

Text Document from different sources
Structured Representation (index, vector etc.)
Visualization
User can interact

Fig.3 Visualization

The goal of information visualization divided into three steps:

- Data preparation step includes deciding and obtaining original data of visualization and form original data space.
- The process of analyzing and extracting visualization data needed from original data and to form visualization data space is known as Data analysis and extraction.
- Visualization mapping step employ certain mapping algorithm to map visualization data space to visualization target.

2.5 Summarization

Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning. Text summarization is helpful for to figure out whether or not a lengthy document meets the user’s needs and is worth reading for further information hence summary can replace the set of documents. In the time taken by the user to read the first paragraph text summarization software processes and summarizes the large text document. It is difficult to teach software to analyze semantics and to interpret meaning of text document even though computers are able to identify people, places, and time. Humans first reads entire text section to summarize then try to develop a full understanding, and then finally write a summary, highlighting its main points.

Summarization process include following steps:
(1)Pre-processing obtain a structured representation of the original text.

- To transform summary structure from text structure algorithm is applied in next processing step.
- In the invention step the final summary is obtained from the summary structure.

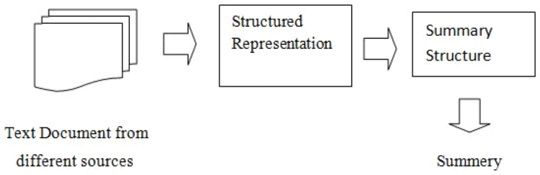


Fig.4 Summarization

3. PROPOSED WORK

There is no such method have been proposed in previous methods to combined the term and phrased based approach. We proposed a method to overcome the problems of polysemy and synonymy to combine the advantageous of term and phrased based approach.

3.1 PROPOSED ALGORITHM

The working procedure of an algorithm is as follows

```
D ←New Document File.
**for** each sentence s in D **do**
    Create node for every new document
    Each node has two field L1 &L2.
    L1 take the all main heading and sub heading from
    text document.
    L2take subheading of all related to contains of L1.
    M ←Empty List {M is a list of matching phrases}
    **for** each file fi ∈ {f2, f3. . . fk} in s **do**
        **if** (fi−1, fi) is an edge in Node **then**
            Extend phrase matches in M for sentences that
              Continue along (fi−1, fi)
            Add new phrase matches to M
        **else**
            Add into node
            Update sentence path in nodes fi−1 and fi
        **end if**
    **end for**
**end for**
Output matching term and phrases in M.
```

3.2 SYSTEM ARCHITECTURE

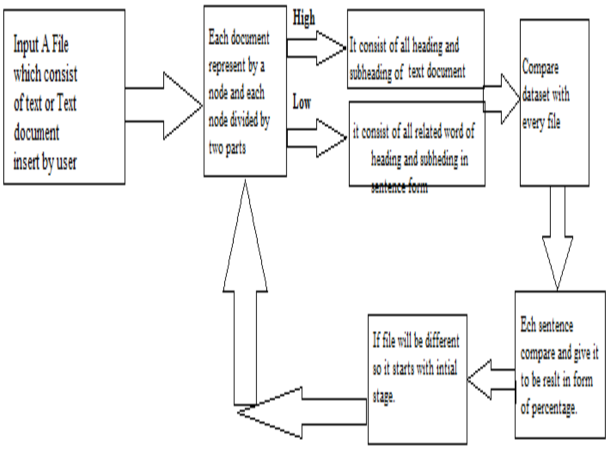

Input A File which consist of text or Text document insert by user
Each document represent by a node and each node divided by two parts
High
Low
It consist of all heading and subheading of text document
it consist of all related word of heading and subheding in sentence form
Compare dataset with every file
Ech sentence compare and give it to be reslt in form of percentage.
If file will be different so it starts with intial stage.

Fig.5 System Architecture

Fig.5 shows the system architecture of proposed architecture. The working process of system architecture is as follows:

- It takes any text document as a input.
- Every time a new file mark as a node.
- Each node consists of two pointers High and Low.
- High pointer consists of main heading and subheading.
- Low pointer consists of all related contents of sub heading.
- If document match with this dataset result show in the form of percentage otherwise it will create another node.

4. IMPLEMENTATION AND RESULT

In this section we discuss about the result of text mining result. In fig. 6, 7, 8 show the whole implementation result in form of filtering and comparison.
In fig. 6&7 shows the input area, how take input, count the word, make sentences and eliminate unnecessary word from it.
In fig.8 give the comparison result of any file or text whatever user want.

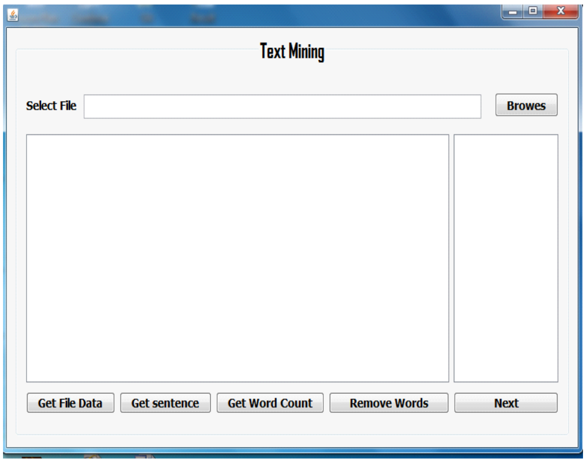
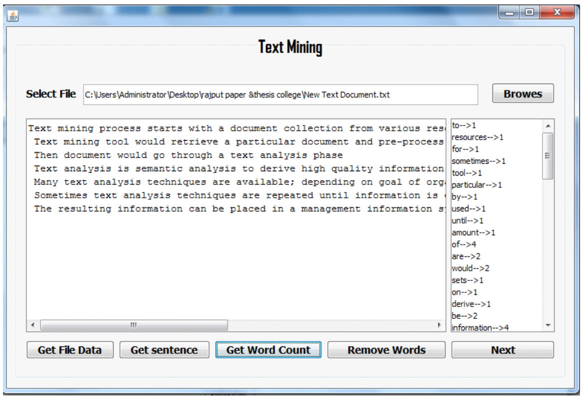


Fig.6 Front Page

DOI 10.17148/IJARCCE.2016.5483

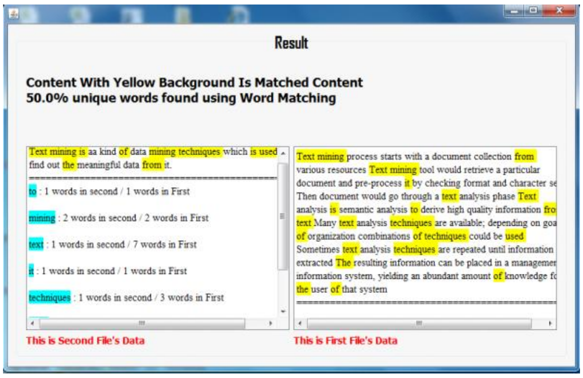


Select File C:\Users\Administrator\Desktop\rajput paper &thesis college\New Text Document.txt

```
Text mining process starts with a document collection from various res
 Text mining tool would retrieve a particular document and pre-process
 Then document would go through a text analysis phase
 Text analysis is semantic analysis to derive high quality information
 Many text analysis techniques are available; depending on goal of org
 Sometimes text analysis techniques are repeated until information is
 The resulting information can be placed in a management information s
```

```
to-->1
resources-->1
for-->1
sometimes-->1
tool-->1
particular-->1
by-->1
used-->1
until-->1
amount-->1
of-->4
are-->2
would-->2
sets-->1
on-->1
derive-->1
be-->2
information-->4
```

Fig.7 Filtering Process



Content With Yellow Background Is Matched Content
50.0% unique words found using Word Matching

Text mining is aa kind of data mining techniques which is used
find out the meaningful data from it.

to : 1 words in second / 1 words in First

mining : 2 words in second / 2 words in First

text : 1 words in second / 7 words in First

it : 1 words in second / 1 words in First

techniques : 1 words in second / 3 words in First

Text mining process starts with a document collection from
various resources Text mining tool would retrieve a particular
document and pre-process it by checking format and character se
Then document would go through a text analysis phase Text
analysis is semantic analysis to derive high quality information fro
text Many text analysis techniques are available; depending on goa
of organization combinations of techniques could be used
Sometimes text analysis techniques are repeated until information
extracted The resulting information can be placed in a managemen
information system, yielding an abundant amount of knowledge fo
the user of that system

Fig.8 Comparison

So the whole process of combine the result of term and phrased basedis shown.it result make it sense and it can be further used in artificial intelligence or any other information retrievalsystem.

5. CONCLUSION

This paper has presented the combined approach of text mining using term based and phrased based approach simultaneously. Different methods have their advantageous and disadvantageous like term based approach suffer from polysemy and synonymy while phrase based approach performs better as phrase carries more semantics like information and is less ambiguous. Two terms can have same frequency from statistical analysis this problem can be solved by combined two methods in a single framework. This approach helps to mine efficient pattern and avoid unnecessary time wastage.

REFERENCES

[1] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” Information Processing and Management:An Int’l J., vol. 24, no. 5, pp. 513-523, 1988.

[2] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I.

[3] Verkamo, “Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document

[4] Collections,” Proc. IEEE Int’l Forum on Research and Technology Advances in Digital Libraries (ADL ’98),

[5] pp. 2-11, 1998.

[6] W. Lam, M.E. Ruiz, and P. Srinivasan, “Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[7] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, “Text Classification Using String Kernels,” J. Machine Learning Research, vol. 2,p. 419-444, 2002.

[8] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, “Automatic Pattern- Taxonomy Extraction for Web Mining,” Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence (WI ’04), pp. 242-248, 2004.

[9] S.-T. Wu, Y. Li, and Y. Xu, “Deploying Approaches for Pattern Refinement in Text Mining,” Proc. IEEE Sixth Int’l Conf. Data Mining (ICDM ’06), pp. 1157- 1161, 2006.

[10] S. Shehata, F. Karray, and M. Kamel, “Enhancing Text Clustering Using Concept-Based Mining Model,” Proc. IEEE Sixth Int’l Conf. Data Mining (ICDM ’06), pp. 1043-1048, 2006.

[11] S. Shehata, F. Karray, and M. Kamel, “A Concept- Based Model for Enhancing Text Categorization,” Proc. 13th Int’l Conf. Knowledge Discovery and Data Mining (KDD ’07), pp. 629-637, 2007.

DOI 10.17148/IJARCCE.2016.5483