IMPROVING TEXT MINING USING DISCOVERY OF RELEVANT FEATURES BY NLP

Miss. Roshani S. Khule
CS&IT, HVPM, COET
Amravati & SGBAU , Maharashtra, India
khule.roshani@gmail.com

Prof. Ranjit R. Keole
HVPM, COET,
Amravati & SGBAU Maharashtra, India
ranjitkeole@gmail.com

---***---

Abstract - *Text mining is a technique for analysing text documents to extract useful knowledge and information. Most text mining methods such as classification, clustering, and summarisation require features such as terms (words), patterns (frequent term sets), or phrases (n-grams) to represent text documents. To enhance the performance of text mining methods, text feature selection is a process to select a subset of text features relevant to the mining task, and use these features to represent the document of interest. However, guaranteeing the high quality of selected features from text is a challenge because of the large amount of irrelevant (noisy) information in text document is a big challenge to guarantee the quality of discovered relevance features in text documents for describing user preferences because of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. Over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences; yet, how to effectively use large scale patterns remains a hard problem in text mining. To make a breakthrough in this challenging issue, this paper presents an innovative model for relevance feature discovery.*

1.INTRODUCTION

Text Mining is equivalent to Information Extraction. The first approach assumes that text mining essentially corresponds to information extraction the extraction of facts from texts. in another word Text Mining is equal to Text Data Mining. Text mining can be also define similar to data mining as the application of algorithms and methods from the fields machine learning and statistics to texts with the goal of finding useful patterns. [1]For this purpose it is necessary to pre-process the texts accordingly. Many authors use information extraction methods, natural language processing or some simple preprocessing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied.

1.1 Related Research Areas

Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modelling of hidden patterns. [8] In this context influence of domain knowledge and domain specific procedures plays an important role. Therefore, an adaptation of the known data mining algorithms to text data is usually necessary. In order to achieve this, one frequently relies on the experience and results of research in information retrieval, natural language processing and information extraction. In all of these areas we also apply data mining methods and statistics to handle their specific tasks:

Information Retrieval (IR):- Information retrieval is the finding of documents which contain answers to questions and not the finding of answers itself.[8] In order to achieve this goal statistical measures and methods are used for the automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval for an overview).

Natural Language Processing (NLP):- The general goal of NLP is to achieve a better understanding of natural language by use of computers . Others include also the employment of simple and durable techniques for the fast processing of text, as they are presented e.g. in. [8]The range of the assigned techniques reaches from the simple manipulation of strings to the automatic processing of natural language inquiries. In addition, linguistic analysis techniques are used among other things for the processing of text.

Information Extraction (IE):- The goal of information extraction methods is the extraction of specific information from text documents. These are stored in data base-like patterns and are then available for further use.[5] In the following, we will frequently refer to the above mentioned related areas of research. We will especially provide examples for the use of machine learning methods in information extraction and information retrieval

1.2 Relevant text feature

Using data mining methods, a large number of terms can be extracted from text files. Over the years, data mining and information retrieval have developed many methods to extract and reduce these features from text documents to fulfil user information needs [3]. The feature selection technique is one of the methods that rely on selection according to the weight of the extracted feature. Usually, terms appear to be general if the term has a large weight, because it frequently appears in both relevant and irrelevant documents [12].

Therefore, a good feature selection method should be able to find the relevant features and disregard the noisy and redundant features. Two feature qualities must be considered by feature selection methods: relevancy and redundancy. The feature is considered as relevant to user needs if it can be predicted; otherwise, it is an irrelevant feature. A redundant feature is one that is highly correlated with other features in the document [5]. The relevance of the extracted text can be classified into three different degrees: strongly relevant, weakly relevant, and irrelevant features, as shown in Figure. The relevance of features has been defined in different studies [2,7]. The strongly relevant feature is one that cannot be removed without some loss of accuracy, while the weakly relevant feature is not strongly relevant but holds some relation to other features in the document. On the other hand, irrelevant features are those that can be removed from the document and do not have any relation with other features, and could be noisy or redundant features.

2.PROPOSED METHODOLOGY

In the current approach, researchers are using N-Gram based technique for detection of relevant features in text mining. This approach is well known for accuracy if the related words in the vicinity of the current word are proper action words, and are relevant to the meaning of the sentence. In the research done, the researchers have used various techniques of N Gram like Bi-gram and tri-gram to incorporate the detection of relevance features in the given text, but these techniques if used alone give in-accurate results due to the lack of pre-processing done on the text to find the input keywords

Our approach uses an improved and efficient Natural Language Processor based on the Word Net API, which performs pre-processing to give better and improved results for relevant feature detection and it's application to text mining. Our approach works in the following manner,

Input Text -> Application of NLP to get action Words -> Application of N-Gram approach for Text Mining -> Mined text results

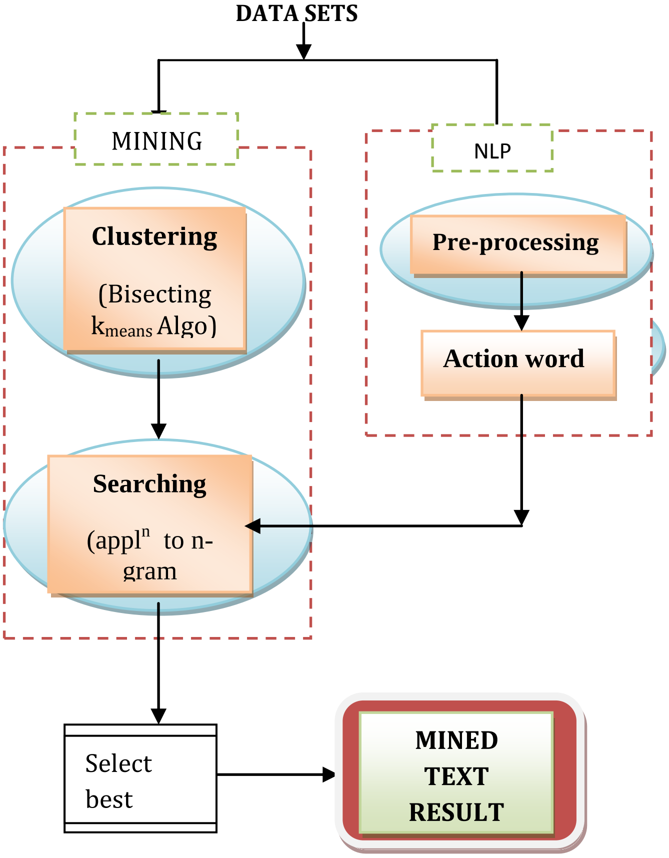


Fig -1: Architecture of Proposed System

2.1 Dataset

Datasets are used to evaluate the proposed technique. Twitter is having one of the most popular database. Evaluation point of view collects the twitter webpage, which contain number of twit on multiple subjects. It is used to test the approach in effective manner.

2.2 Natural Language Processing

NLP provides means of analyzing the text. The goal of NLP is to make computers analyze and understand the languages that humans use naturally. As the reader has probably already deduced, the complexity associated with natural language is especially key when retrieving textual information to satisfy a user's information needs. This is why in Textual Information Retrieval, NLP techniques are often used both for facilitating descriptions of document content and for presenting the user's query, all with the aim of comparing both descriptions and presenting the user the documents that best satisfy their information needs

The pre-processing consisted of deferent steps, as shown in Figure

1. text fields in the XML document to represent the documents' content.
2. All the stopwords were removed to reduce the noise in documents. Stopwords can be denned as common words that frequently occur in the document such as articles, prepositions, and conjunctions. Common stop words in English include [6]:
a; about; an; are; as; at; be; by; for; from; how; in; is; of; on; or that; the; these; this; to;was ;what;when;where;who;willwith:
3. Word stemming is a further step in preprocessing to remove some noise in the document. Thus, word stemming tries to solve the problem in a variety of forms of the word by in flecting words into their stem or root form. In this thesis the Porter algorithm is used for suffix stripping
4. Deferent k-words were selected to reduce further noise keywords.
5. The final preprocessing step was to transform the (title) and (text) fields in the document into paragraphs which were tagged with (p) in an XML file, where the (text) section contained one or more paragraphs and each paragraph consisted of low-level features (terms).

2.3 Clustering

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

2.3.1 Bisecting k-Means

Bisecting k-Means is like a combination of k-Means and hierarchical clustering.
It starts with all objects in a single cluster.
The pseudocode of the algorithm is displayed below:
Basic Bisecting K-means Algorithm for finding K clusters:-

```
1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic k-Means algorithm (Bisecting step)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.
```

2.4 Searching

Searching can be considered the most important part of proposed system. In proposed system searching is done with the help of N-Gram technique which is describe in section 2.2 . and the steps are given below.

2.4.1 N-gram Extraction Using Relevant Feature Selection (GERS)

Extracting the best features and weighting them is an important task in data mining and information retrieval. Many features such as high-level n-grams can be extracted from datasets using various methods; however, many of these features are irrelevant to user needs. In addition, calculation of the high-level n-gram's weight could be inaccurate if the distribution of the n-gram's contents (low-level terms) is not considered.

Procedure for Evaluation of N-gram Extraction .The steps required for the GERS evaluation procedure are listed as follows:

1. The system starts with one of the dataset topics and retrieves the related information with regard to the training set, such as the file name list and the number of documents.

2. Each document is preprocessed with word stemming and stopword removal and transformed into a set of transactions based on its document structure (paragraphs).

3. The best top-k low-level terms extracted from some models such as Rocchio and BM25 are selected to be used in high-level n-gram extraction.

4. The system then extracts the high-level n-grams a using window size based on the selected low-level features.

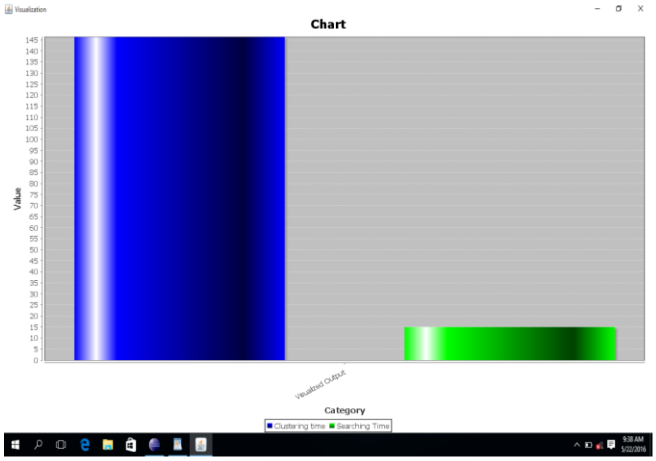
5. The system calculates the probability for each extracted term.

6. The random set is extended to calculate the n-gram probability using the terms' probability, as calculated in the previous step.

7. Finally, the results of n-gram extraction using the ERS model are evaluated.

3.RESULTS AND DISCUSSION

ScreenshotNo.3.1 :- graph between clustering time and searching time on x-axis with praposed algorithem and on y-axis without algorithem



In this section describe the result of proposed approach "improving text mining using discovery of relevant features by natural language processing". Experiment shows the proposed system is better than existing system because it takes less time and more accuracy. Following graph shows the clustering time and searching time of one of the example, this is taken in to the screenshot section. Graph shows the time taken by the clustering phase searching phase on x-axis with praposed algorithem and on y-axis without algorithem . Hence, it proves that the propose system is better in time than any another system

4.APPLICATIONS

In this section we briefly discuss successful applications of text mining methods in quite diverse areas as patent analysis, text classification in news agencies, bioinformatics and spam filtering. Each of the applications has specific characteristics that had to be considered while selecting appropriate text mining methods.

4.1 Patent Analysis:-In recent years the analysis of patents developed to a large application area. The reasons for this are on the one hand the increased number of patent applications and on the other hand the progress that had been made in text classification, which allows to use these techniques in this due to the commercial impact quite sensitive area. Meanwhile, supervised and unsupervised techniques are applied to analyze patent documents and to support companies and also the European patent office in their work.

4.2 Text Classification for News Agencies :-In publishing houses a large number of news stories arrive each day. The users like to have these stories tagged with categories and the names of important persons, organizations and places. To automate this process the Deutsche Presse-Agentur (dpa) and a group of leading German broadcasters (PAN) wanted to select a commercial text classification system to support the annotation of news articles.
4.3 Bioinformatics Bio-entity :- recognition aims to identify and classify technical terms in the domain of molecular biology that correspond to instances of concepts that are of interest to biologists. Examples of such entities include the names of proteins, genes and their locations of activity such as cells or organism names

4.4 Anti-Spam Filtering of Emails :-The explosive growth of unsolicited e-mail, more commonly known as spam, over the last years has been undermining constantly the usability of e-mail. One solution is offered by anti-spam filters. Most commercially available filters use black-lists and hand-crafted rules.

5.CONCLUSION

The paper proves that the use of N-gram approach is significant for improving the performance of relevance feature discovery models. It provides a promising methodology for developing effective text mining models for relevance feature discovery. Automatic text mining techniques have a long way to go before they rival the ability of people, even without any special domain knowledge, to glean information from large document collections.

6.FUTURE SCOPE

It has been proven that a text document includes a lot of information regarding to different kind of topics. However, not all information in the document is useful for learning relevant features to a given topic. To improve the quality of extracted features, the proposed model and patterns to select term features. this dissertation shows that selected pattern are good for representing documents but not good enough to represent queries for answering what users want.

REFERENCES

[1] Yuefng Li, Abdulmohsen Algarni,Mubarak Albathan, Yan shen, and moch Arif Bijaksana"Relevance feature discovery for text mining" IEEE transaction on knowledge and data engineering,vol.27,no.6,june2015

[2] Ning Zhong, Member of the IEEE, Yuefeng Li, Member, IEEE, and Sheng Tang Wu, Member, IEEE " Effective Pattern Discovery for Text Mining"IEEE, transaction on knowledge and data engineering,vol.24,no. 1, January 2012

[3] D.M.Kulkarni and S.K.Shirgave"using data mining methods knowledge discovery for text mining" QUT E-Discovery lab,

[4] V.Sharmila,I.Vasudevan, Dr.g.Tholkappia Arasu "Pattern based classification for text mining using fuzzy similarity algorithm" Journal of theoretical and applied information technology,vol 63,May 2014.

[5] Mallareddy kiran, R.Ravikanth ME "Classification of documents using effective pattern taxonomy"International journal of computer applications,vol 86, no. 6, January 2014.

[6] Miss. Dipti charjan and prof. Mukesh Pund "Pattern discovery for text mining using pattern taxonomy"International journal of engineering trends and technology,vol 4,issue 10,October 2013

[7] Yuefeng Li,Xiaohui Tao, Abdulmohsen Algarni,Sheng Tang Wu " Mining specific and general features in both positive and negative relevance feedback" Australian research council, July 2010

[8] Vishal Gupta and Gurpreet S. Lehal" A survy of text mining techniques and application" Journal of emerging technologies in web intelligence, vol 1, no. 1, August 2009

[9] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, “Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,” Proc. IEEE Int’l Forum on Research and Technology Advances in Digital Libraries (ADL ’98), pp. 2-11, 1998.

[10] Andreas Hotho “A Brief Survey of Text Mining” KDE Group University of KasselMay 13, 2005

[11] 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.