International Journal of Applied Engineering and Technology ISSN: 2277-212X (Online)
An Open Access, Online International Journal Available at http://www.cibtech.org/jet.htm
2014 Vol. 4 (4) October-December, pp.6-9/Harpreet and Rupinder
Research Article

KEYWORDS EXTRACTION AND TEXT SUMMARY GENERATION BASED ON WORDS HISTOGRAM AND SYNONYMS

***Harpreet Kaur and Rupinder Kaur**
DIET, Kharar, Punjab, INDIA
**Author for Correspondence*

ABSTRACT
In today's digital era of information technology, it is very challenging to extract the useful information from a chunk of large data base of textual document. To counter this challenge, accurate keywords extraction from the digital documents is presented that not only generate the summary of the documents containing the keywords sentences and also helps in clustering the similar documents based on either keywords vector set of text summary. The key words extraction is based on the frequency of a word in documents and its synonyms occurrence in the entire document. Further, for a word to be a potential keyword, its font attributes are also considered. The text summary also includes the factual information that is quoted in double inverted comas. The performance of the keywords and then text summary is computed by using the precision and recall.

Keywords: *Relevance Feedback, Text Mining*

INTRODUCTION
As the amount of information in the modern world grows rapidly, it is becoming more and more difficult to maintain and process document archives for Knowledge Management Systems, Information Retrieval Systems, and Digital Libraries. This is true especially for very large archives with millions or more articles. Processing all the words in the documents, as if they are of equal importance, as basis for finding relevant articles would be slow and not practical. That is why it is important to have a set of good keywords that represent the actual contents of the document. However, it is not possible to have all documents labelled by experts. It is therefore useful to be able to automatically identify keywords in the documents that are just as good as assigned keywords.
Keywords are often the basis of document processing methods such as clustering and retrieval since processing all the words in the document can be slow. Common models for automating the process of keyword extraction are usually done by using several statistics-based methods such as Bayesian, K-Nearest Neighbour, and Expectation-Maximization. These models are limited by word-related features that can be used since adding more features will make the models more complex and difficult to comprehend.

Related Works
This research proposes a new neural network for text categorization which uses alternative representations of documents to numerical vectors. Since the proposed neural network is intended originally only for text categorization, it is called NTC (Neural Text Categorizer) in this research (Raymond *et al.,* 2003).
They concluded that many data mining and data analysis techniques operate on dense matrices or complete tables of data. Real world data sets, however, often contain unknown values. Even many classification algorithms that are designed to operate with missing values still exhibit deteriorated accuracy (Govindarajan and Chandrasekaran, 2007).
This paper proposes a new email classification model using a teaching process of multi-layer neural network to implement back propagation technique (Aurangzeb *et al.,* 2010).
In this paper they concluded Text Mining is around applying knowledge discovery techniques to unstructured text is termed knowledge discovery in text (KDT), or Text data mining or Text Mining. In Neural Network that address classification problems, training set, testing set, learning rate are considered as key tasks (Taeho, 2010).

However, how to effectively use and update discovered 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 problems of polysemy and synonymy (Taiwo *et al.*, 2010).

In this paper they are tried to give the introduction of text classification, process of text classification as well as the overview of the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principal and performance (Ning *et al.*, 2012).

The aim of this paper is to highlight the important techniques and methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. This paper provides a review of the theory and methods of document classification and text mining, focusing on the existing literature (Vandana and Namrata, 2012).

Algorithm

Since documents are unstructured data by themselves, they cannot be processed directly by computers. They need to be encoded into structured data for processing them for text categorization. This section will describe the two strategies of encoding documents with the two subsections: the traditional strategy and the proposed strategy. The first subsection describes the former and points out its demerits, and the second subsection describes the latter and mentions its merits.

The proposed algorithm is implemented in following steps:

- Document Scanning
- Low Level Features Extraction using term based approach
- Relevant Features (RF) Extraction
- Portioning of RF into GT and ST
- Key Text optimization as o/p of BPN
- Key words validation
- Large font sized words extraction
- Underlined words extraction
- Back Propagation Neural Approach using RF, ST and GT
- Keywords Extraction

Results

The proposed algorithm has been tested on no. of text documents in notepad format. Following one case study is given below:

Original Text

For too many students, a college education that is supposed to create opportunities can also mean years of struggle to pay off tens of thousands of dollars in debt. Schools must be required to do more to educate students about the real cost of their education and about a complex borrowing process that even the most sophisticated people have trouble understanding.

An article in The Times last week described the experience of 23-year-old Kelsey Griffith. She currently earns a meager wage as a restaurant worker and owes $120,000 in student loans for an undergraduate degree from Ohio Northern University, a college whose recent graduates are among the most indebted in the country.

Nationally, about two-thirds of bachelor's degree recipients now borrow from either public or private lenders, up significantly from the early '90s, when about 45 percent of graduates borrowed from all sources, including family. According to an analysis by the Federal Reserve Bank of New York, the average debt for student borrowers last year was about $23,300, while 10 percent owed more than $54,000 and 3 percent owed more than $100,000.

Ms. Griffith's debt was worsened by the fact that she changed majors and took five years to graduate. And because federal loans did not cover her total costs, she had to take out more expensive private loans that offer fewer protections than federal loans — like deferments and income-based repayment plans. Ms. Griffith voiced an increasingly common sentiment when she told The Times: "I knew a private school would cost a lot of money. But, when I graduate, I'm going to owe like $900 a month. No one told me that."

Text Summary when using the proposed Algorithm

For too many students, a college education that is supposed to create opportunities can also mean years of struggle to pay off tens of thousands of dollars in debt. An article in The Times last week described the experience of 23-year-old Kelsey Griffith. Ms. Griffith's debt was worsened by the fact that she changed majors and took five years to graduate. Ms. Griffith voiced an increasingly common sentiment when she told The Times: "I knew a private school would cost a lot of money.

Ideal Text Summary

For too many students, a college education that is supposed to create opportunities can also mean years of struggle to pay off tens of thousands of dollars in debt. An article in The Times last week described the experience of 23-year-old Kelsey Griffith. Ms. Griffith's debt was worsened by the fact that she changed majors and took five years to graduate. Ms. Griffith voiced an increasingly common sentiment when she told The Times: "I knew a private school would cost a lot of money.

The algorithm performance is measured by computing the precision and recall as follows:

$$\text{Precision} = \frac{N}{N_{IS}}$$

$$\text{Recall} = \frac{N}{N_{CS}}$$

Where,

N_{IS} = No. of Sentences in ideal summary

N_{CS} = No. of sentences in computed summary

N = No. of common sentences in Ideal and Computed Summary

In the presented case study,

No. of sentences in Ideal Summary = 5;

No. of Sentences in computed summary = 5;

Therefore, $\text{Precision} = \frac{5}{5} = 1$

And $\text{Recall} = \frac{5}{5} = 1$

The results are computed for other different domain documents and precision and recall are computed and tabulated as follows:

Documents	**N_{IS}**	**N_{CS}**	**N**	**P**	**R**
D-1	24	20	19	0.80	0.95
D-2	51	49	45	0.88	0.96
D-3	38	35	34	0.90	0.92

Conclusion

It is observed that the presented algorithm works fine in extracting the text summary to a fair extent. However, the time consuming may vary when the documents size increases. The algorithm is tested on notepad format document and the same may extended to pdf, word and other text document format. Also, the extension of the work is also sought in tabular data format text summarization.

REFERENCES

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee and Khairullah Khan (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology* **1**(1).

Govindarajan M and Chandrasekaran RM (2007). Classifier Based Text Mining for Neural Network" World Academy of Science, Engineering and Technology. *International Journal of Computer, Information Science and Engineering* **1**(3).

Luepol Pipanmaekaporn (2013). Feature Discovery in Relevance Feedback Using Pattern Mining 978-1-4799-0174-6/13/$31.00 ©2013 IEEE.

Michael S Gashler, Michael R Smith, Richard Morris and Tony Martinez (2013). Missing Value Imputation with Unsupervised Backpropagationar **Xiv** 1312.5394v1 [cs.NE].

Ning Zhong, Yuefeng Li and Sheng-Tang Wu (2012). Effective Pattern Discovery for Text Mining. *IEEE Transactions on Knowledge and Data Engineering* **24**(1).

Nitu Mathuriya and Ashish Bansal (No Date). Comparison of K-means and Backpropagation Data Mining Algorithms, *International Journal of Computer Technology and Electronics Engineering (IJCTEE)* **2**(2).

Raymond Chan, Qiang Yang and Yi Dong Shen (2003). Mining High Utility Itemsets. *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)* 0-7695-1978-4/03 $ 17.00 © 2003 IEEE.

Taeho Jo (2010). School of Information Technology & Engineering Ottawa University Ontario, Canada "NTC (Neural Text Categorizer): Neural Network for Text Categorization. *International Journal of Information Studies* **2**(2).

Taiwo Ayodele, Shikun Zhou and Rinat Khusainov (2010). Mail Classification Using Back Propagation Technique. *International Journal of Intelligent Computing Research (IJICR)* **1**(1/2).

Vandana Korde and Namrata Mahender C (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications (IJAIA)* **3**(2).