Improving Text Search Process using Text Document Clustering Approach

Ruchika Mavis Daniel[1], **Arun Kumar Shukla**[2]

[1]Department of CS & IT, SSET, SHIATS (Deemed-to-be-University), Allahabad, India

[2]Department of CS & IT, SSET, SHIATS (Deemed-to-be-University), Allahabad, India

Abstract: ***Knowledge discovery and data mining is a process of retrieving the meaningful knowledge from the raw data, using different techniques. Therefore, text mining is a sub domain of knowledge discovery from the text data. This paper provides a different way of understanding the text mining and their applications in different real time applications. This paper also includes the design of a hybrid text document clustering approach by which the document organization becomes easier for text search process. That includes the concept of fuzzy calculation, text summarization and traditional k-means clustering for cluster analysis of data. The implementation of the proposed methodology is summarized using a new kind of document clustering algorithm, and implemented using visual studio environment. After implementation of the proposed methodology the performance of the designed system is evaluated using the N-cross validation process. According to the obtained results the proposed system outperformed for text categorization with small amount of resources consumption. So, the proposed methodology is efficient and adoptable for different real time application.***

Keywords: text search, text clustering, text classification and categorization

1. Introduction

The main aim in this paper is to improve the technique of automatic text categorization process. Basically, text mining is a domain of unsupervised learning techniques due to the absence of a single domain found with the class labels. Additionally, the data is always found in unstructured way. The text categorization includes the various semantically and statistical issues. In this proposed work, a hybrid approach to optimize the process of text categorization using the text summarization techniques is proposed. This technique is employed for finding the essential features form the large text documents for preserving the computational resources during text analysis.

In this era of computational intelligence most of the data resources are available in text format. The amount of data is too big to analyse and finding the important knowledge is too complex. Therefore, that is an interesting domain of research and development. The text categorization includes a wide range of applications that is part a of knowledge process and natural language processing in artificial intelligence. The content mining, artificial intelligence and content mining techniques that include various applications such as semantics analysis, compiler design, document and big data analysis are the sub domain of text and semantic analysis.

Text mining deals with the computational analysis of text for knowledge discovery and data pattern analysis. These techniques provide ease in information extraction, natural language processing, and information retrieval. Additionally, these domains are included with algorithms and KDD methodologies. [12] A similar procedure cannot be followed with this domain of KDD process, where data is not in general format and similar to each other. Therefore, text documents are required to be seriously analysed. From this a new question for data mining techniques is the data modelling perception for unstructured data sets [1].

If making effort for finding the definition of text mining, then that can be nearer to related research domain's or application's specific. At this point each of them can provide a different meaning of text mining, which is inspired by the specific viewpoint of the application area [1]:

- Text Mining = Information Extraction.
  In this context the text mining is essentially parallel to information extraction, which means the information extraction from texts.
- Text Mining = Text Data Mining.
  Text mining can be also defined by approximating to data mining as the application of algorithms and technique from the domain of machine learning and statistics analysis over text documents with the aim of discovering fruitful patterns. For that motive it is required to pre-process the text documents more appropriately. Different researchers use natural language processing, information extraction techniques or other simple pre-processing algorithms to find meaningful information from texts.
- Text Mining = KDD Process.
  In the knowledge discovery process model, it is commonly found that the text mining is a process with a series of fractional steps, among the use of data mining or statistical analysis. In general way, the extraction of information that is not yet discovered in collection of texts documents also text mining as procedure orientated methodology on texts.

2. Background

In recent years, due to technological growth various information and data are converted into digital formats for ease of storage and maintenance. These formats of data are available in huge amount. Manual evaluation of individual data files and classification is a complex and time consuming process. Therefore, a number of automatic text categorization and classification techniques are recently proposed and implemented in order to find most appropriate

Licensed Under Creative Commons Attribution CC BY

document organization or classification for efficient data identification and retrieval process. In this context term based, summarization based, machine learning based techniques are available. Most of them results poor outcomes due to unlabelled text.

On the other hand during search process, the text mining algorithms are employed over documents in order to find most relevant text document. This may arise the issue of high resource consumption, therefore it is required to develop and design an efficient solution that categorise the document and adopt new documents in a domain specific document cluster, with optimal resource consumption. Therefore, the proposed work is intended to improve the content based text search process model of text retrieval, using the proposed text document clustering approach.

Text mining is a technique to recover meaningful and effective data from available text data bases. In this context the following is the main area of investigation under which the solution is required to be found.

1. For data categorization, each time the evaluation of large text files is required. Additionally, a significant amount of time is also consumed during the evaluation process.
2. Evaluation of huge text files require to pre-load the data in the main memory. Thus a significant amount of memory resources are consumed.
3. Text summarization technique for data categorization is much popular for its low amount of text evaluation processes. But the summarization process can affect the semantics of document by which proper categorization of document is difficult in a specific domain.

Text categorization faces various complex issues related to semantics and resources in order to find an efficient technique for text categorization by which the resource consumption (i.e. memory and time) can be optimized. Additionally, a technique is required to enhance the classification accuracy during text categorization. Therefore, a hybrid text reduction and document clustering approach is proposed which reduces the amount of text for finding a small set of data that represent the whole text document. Additionally, the information is preserved for future categorization by which the additional document evaluation overhead can be reduced. Using this technique the categorization of text documents becomes efficient and effective in terms of memory, accuracy and time.

3. Proposed System

The proposed concept is a complex data model that involves various sequence of process that consumes the data and produces the outcomes. The figure 1 demonstrates the proposed methodology for implementing the efficient technique for text categorization.

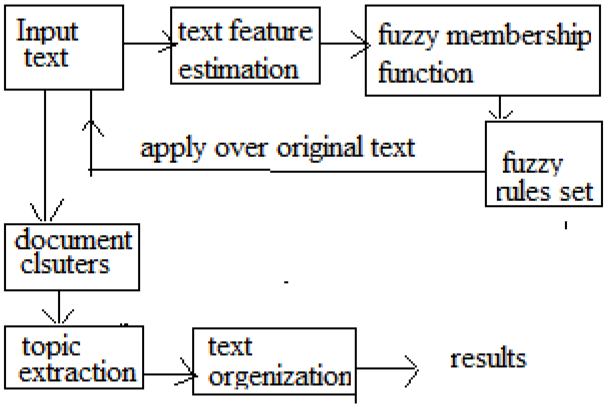

Figure 3.1: system design

According to the given system design the process is initialized with the input text. The system, consumes the input text and pre-process technique is applied for obtaining cleaned data. The text data base may include different data formats such as HTML documents therefore recovery of original text from document is necessary in this phase. After cleaning process the features are calculated first such as word frequency, sentence formation probability, these feature are used to reduce the significant amount of text from the input document text. Than after required to find the distribution of the estimated features for a specified domain, for that purposes the fuzzy logic based membership function is called to discover the word probability distribution over the available text domains. Using the evaluated text features and estimated probability distribution the size of original data is reduced.

In next step, a K-mean clustering inspired process is applied for topic wise text categorization for the available text document. A noteworthy point is that the performance of K-mean is less for topic wise text categorization due to random cluster centre election. Therefore the K-mean technique is not applied directly for data cluster generation that is optimizing for efficient use.

A. System Architecture

This section of the document describes the system architecture for implementation of the proposed working model. The given figure 2 demonstrates the system architecture to find the text summarization based text categorization technique for unlabelled data clustering process. The sub components of the data model are given as:

The proposed text summarization and categorization process is divided into three essential modules namely Feature extraction, summarization and cluster organization the brief description of each module is provided here.

1. Text document feature extraction: The first and important module of the system is recovery of text feature vector like word frequency and statement formation probability. These features reflect the important attributes in the given text document. Using these features, the document is identified with their domain.

2. Rule mapping and text reduction: In this phase of process, the input text is processed again to reduce the size of the text document. Here the features calculated in the previous step are utilized for reducing the text size. The less amount of text is easy to handle and processing with the clustering algorithms.

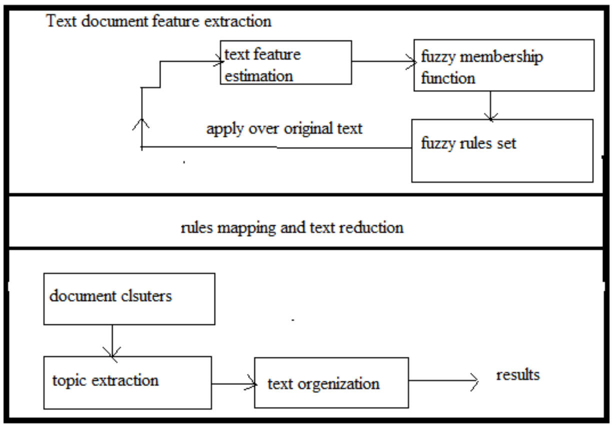

Figure 2: system architecture

3. Document generation and results: In this phase the reduced text is used with the clustering algorithm, by which the document is clustered topic-wise and the list of clusters are produced as final outcome.

B. Proposed Algorithm

This section includes the proposed algorithm for optimum cluster organization using hybrid approach of fuzzy logic and the traditional K-mean clustering algorithm. Basically the proposed technique first selects the features using their probability distribution and then the amount of data is reduced for efficient data processing. And finally using the iterative distance measurement techniques the data is categorized for predefined class labels.

```
Input: domain name, data samples
Output: categorized data list
Process:
1. Read the whole document
2. Pre-process in order to remove unwanted tag and text
3. Find features
    a. Term frequency
       $Tf = \sum_{i=0}^{n} \frac{word_i}{total\ words\ in\ document}$
    b. Sentence formation frequency
       $Sf = \sum_{i=0}^{n} \frac{word_i}{total\ sentences}$
    c. Create name value pair using array list
4. Sort array list
5. Find top 50 words as feature set
6. Eliminate remaining text from the input text
7. Find distance between domain keywords and input text using
       $d(x,y) = \sum_{i=0}^{n} \sqrt{(x_i - y_i)^2}$
8. Closest distance indicate the document's class
```

4. Results Analysis

After implementing the desired algorithm for document clustering and categorization, the performance of the system is estimated in order to find effectiveness in terms of resource consumption (i.e. memory and time consumption). Additionally, the accuracy and error rate to provide efficiency of the system.

A. Accuracy

The accuracy of the system demonstrates how efficiently a document recognizes their domain in order to categorize under a domain. The accuracy of the system is evaluated during different experiments and different documents, and best obtained results are listed in this section. The following formula is used for calculating the accuracy of the desired system.

$$accuracy = \frac{total\ correctly\ classified\ data}{total\ samples\ avilable\ to\ classify}$$

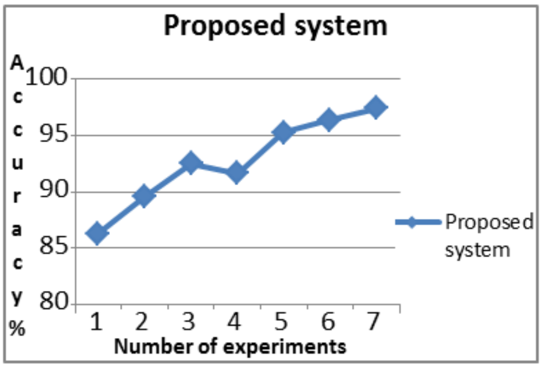

Figure 3: Accuracy of system

B. Error Rate

Error rate of the system reflects the outcome is how far from the existing solution, therefore the error rate of the system can be given using the below formula

$$error\ rate = 100 - accuracy$$

C. Memory Usage

The amount of memory consumed during the process execution is known as the memory consumed. The figure 5 shows the memory usage by the system that is calculated on the basis of peak working set.

D. Clustering Time

The total amount of time is consumed during the categorization of documents are known as clustering time. The time is given using figure 6.

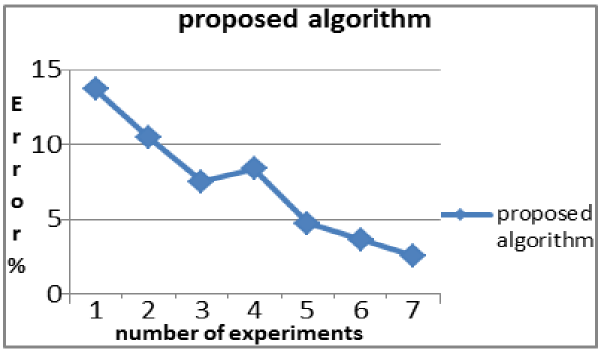

Figure 4: Error Rate

Licensed Under Creative Commons Attribution CC BY

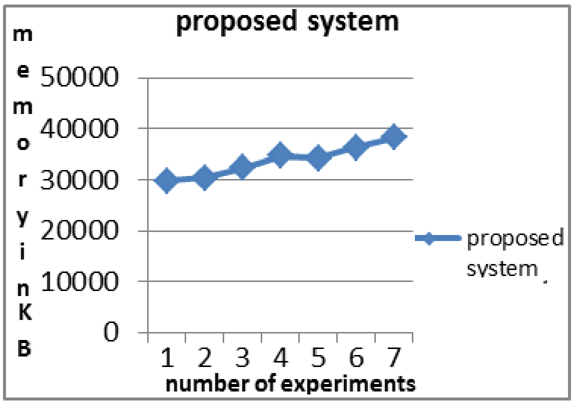

Figure 5: Memory Uses

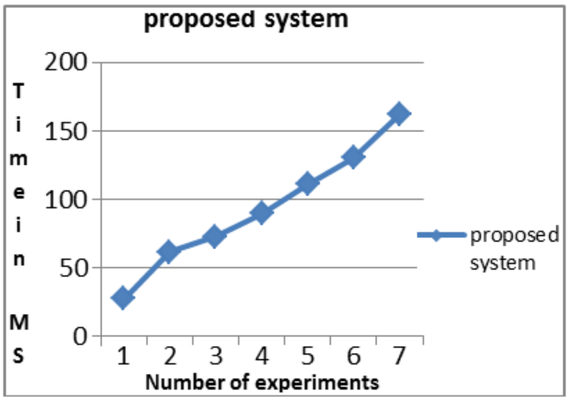

Figure 6: Time Consumption

5. Conclusions

The proposed work is intended to provide efficient text categorization technique using the text summarization technique and cluster analysis technique. Therefore, in this study a hybrid text clustering technique is developed for categorizing the text in a given domain. In text mining and text categorization, the resource consumption is the major issue, therefore feature extraction technique is utilized to reduce the amount of text. This reduced text represents the whole text document. Additionally, the k-mean clustering algorithm's Euclidian distance based approach is utilized for finding similarity between domain knowledge and available document.

The proposed technique is implemented and demonstrated using the visual studio environment and for performance analysis the well-known data mining validation method namely N-cross validation process is consumed. The performance of designed system is demonstrated in previous section additionally the performance summary is given using the table 1

Table 1: Performance summary

Parameter	*Remark*
Accuracy	The accuracy of the system is found approximately 80-95% in complex cases. Therefore the system is adoptable for text categorization process.
Error rate	The evaluated error rate is low, therefore that is adoptable in performance issue.
Memory	Less memory is consumed during huge data process, the amount of memory is below than 40000KB, during different document clustering conditions
Time	Less time consuming for finding the similar domain of a given document

The overall performance of the designed system is adoptable and useful for text mining processes.

References

[1] Andreas H., Andreas N., Gerhard P, Fraunhofer A, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005

[2] Anne S, Johanna B, "Applying text-mining to personalization and customization research literature – Who, what and where", 2012 Elsevier Ltd. All rights reserved

[3] Daniel R, Christopher D. M, Susan D, "Partially Labeled Topic Models for Interpretable Text Mining", KDD'11, August 21–24, 2011, Copyright 2011 ACM 978-1-4503-0813-7/11/08.

[4] Deng C and Xiaofei H, "Manifold Adaptive Experimental Design for Text Categorization", accepted 17 Sep. 2010

[5] Fuzhen Z, Ping L, Hui X, Qing H, Yuhong X and Zhongzhi S, "Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization", 27 October 2010, DOI:10.1002/sam.10099, Wiley Online Library.

[6] G. Koteswara R and Shubhamoy D, "DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH", International Journal of Managing Information Technology, Vol.3, No.3, August 2011

[7] Koby C, Mark D, Fernando P, "Confidence-Weighted Linear Classification for Text Categorization", Journal of Machine Learning Research 13 (2012) 1891-1926

[8] Luís Moreira-M, João Mendes-M, João G, and Pavel B, "Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix", MLDM 2012, pp. 525–539, 2012. © Springer.

[9] Miloš R, Mirjana I, "Text Mining: Approaches and Applications", Abstract Methods and Applications in Computer Science, Vol. 38, No. 3, 2008, 227-234

[10] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, "Knowledge Extraction Using Rule Based Decision Tree Approach", International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[11] Roberto N, Stefano F, Aitor S, Oier de L, Eneko A, "Two Birds with One Stone: Learning Semantic Models for Text Categorization and Word Sense Disambiguation", CIKM'11, Copyright 2011 ACM 978-1-4503-0717-8/11/10

[12] Umajancy. S, Dr. Antony S T, "An Analysis on Text Mining –Text Retrieval and Text Extraction",

International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[13] Vishal G., Gurpreet S. L, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in web Intelligence, VOL. 1, NO. 1, AUGUST 2009

[14] Zhong, N, Li, Y, & Wu, Sheng-T, “Effective pattern discovery for text mining”. IEEE Transactions on Knowledge and Data Engineering, (2010)

Licensed Under Creative Commons Attribution CC BY