A Novel approach for Information Retrieval of Text Documents

Miss.K.S.Hantodkar Dr.S.S.Sherekar Dr.V.M.Thakare

Abstract— Text mining is gaining attention because of its automatically discovering technique of knowledge assets buried in unstructured text. In text mining techniques, the frequency of a word or phrase is observed to gain the importance of the term in the document. The similarity between documents is also measured by one of several similarity measures by text mining that are based on such a feature vector. This paper, discusses briefly five methods i.e Concept-based mining model, Text-driven D-matrix method, Ontology based Text Mining Method, PPSGEN method and VarifocalReader method.

This paper proposes a method for information retrieval from text documents which is the combination of text driven D matrix, concept based similarity model and SVR model. This paper proposed a method which is a combination of some model that will help user to retrieve information from similar topic text documents among various given text documents.

Index Terms— Concept-based mining, Text-driven D-matrix, Ontology based Text Mining, PPSGEN, VarifocalReader

1 INTRODUCTION

TEXT mining is a method that discovers new and previously unknown information by applying natural language processing and data mining technique. Text mining is gaining attention because of its automatically discovering technique of knowledge assets buried in unstructured text [1]. In text mining techniques, the frequency of a word or phrase is observed to gain the importance of the term in the document. However, two terms can have the same frequency in the document, but one term contributes more meaning to its sentences than other. The similarity between documents is also measured by one of several similarity measures by text mining that are based on such a feature vector. Text mining is also done for selection of research projects and analyzing recurring activity in many organizations such as government research funding agencies [2]. Academic papers generally have a similar structure. They contain several sections like abstract, introduction, related work, proposed method, experiments and conclusions. When dealing with textual data, natural language processing techniques are unsurprisingly method of choice for automatic analysis method [3].

This paper, discusses various methods used for text retrieval such as Concept-based mining model [5], Text-driven D-matrix method [6], Ontology based Text Mining Method [8], PPSGEN method [11] and VarifocalReader method [12]. To improve the performance of the text mining for information retrieval from text documents, integrate SVR model, concept based similarity measure model and text driven method.

2 RELATED WORK

The study on mining of text documents discusses the most relevant mining techniques developed in recent years. Concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure [4]. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only. A new concept based mining model composed of four components, is proposed to improve the text clustering quality. The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. This method has drawback that it cannot work on web documents [5].

Ontology-based text mining method (OTMM) to cluster research proposals was a hybrid method for grouping Chinese research proposals for project selection. It uses text-mining, multilingual ontology, optimization, and statistical analysis techniques to cluster research proposals based on their similarities. The proposed OTMM was used together with statistical method and optimization models. The proposed method promotes the efficiency in the proposal grouping process. By manual grouping, users need to spend at least one week, while the grouping can be finished within hours using the proposed methods. Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically [6].

Ontology-based text mining methodology constructs the D-matrices [7] by automatically mining the unstructured repair verbatim data collected during fault diagnosis. The method is implemented as a prototype tool and validated by using real-life data collected from the automobile domain. The performance of the text-driven D-matrix when compared with other method demonstrated improved fault detection and fault isolation rate while exhibiting lower error rate. The text-driven D-matrix of huge size helped the engineering users to perform the root-cause analysis accurately [8].

Paper presen

tation slide generation (PPSGen) generates presentation slides from academic papers which trains a sentence scoring model based on support vector regression (SVR)[9] and use the integer linear programming (ILP)[10] method to align and extract key phrases and sentences for generating the slides. The advantage of this method is that it generates the slides auto-

matically and it is well structured slides that can be easily understandable by humans. People can quickly catch the key points when looking at the phrases and get more information after reading the sentences below the phrases. Additional information such as other relevant papers and the citation information can be used to improve the generated slides. This issue can be considered in the future [11].

VarifocalReader is a technique that helps to solve some of the problems by combining characteristics. VarifocalReader supports intra-document exploration through advanced navigation concepts and facilitates visual analysis tasks. VarifocalReader was designed to let scholars perform a variety of analysis tasks of diverse abstraction, varying granularity, and different requirements for automatic analysis quality in one interactive visual tool. VarifocalReader can also display scanned images of the original pages next to the detail layer. Missing feature of this method is a language guesser [12].

This paper analyses five methods i.e Concept-based mining model, Text-driven D-matrix method, Ontology based Text Mining Method, PPSGEN method and VarifocalReader method and these are organizes as follows. Section I Introduction. Section II discusses related work done. Section III discusses existing methodologies. Section IV discusses attributes and parameters. Section V proposed method and outcome result possible. Finally section VI concludes this review paper.

3 EXISTING METHODOLOGIES

Many text mining methods has been implemented over the last decades. There are different methodologies that are implemented for mining text document i.e Concept-based mining model, Text-driven D-matrix method, Ontology based Text Mining Method, PPSGEN method and VarifocalReader method.

Concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure, as depicted in Fig. 1.

OTMM use statistical and optimization models and have four phases. First, a research ontology containing the projects of latest five years is organized according to keywords. Then, new research proposal is classified to its area using a sorting algorithm. Next, the new proposals in each discipline are clustered using a self-organized mapping algorithm. Finally, if the number of proposals in each cluster is still very large, they will be further decomposed into subgroups.

Ontology-based text mining method D-matrix automatically constructs and updates mining thousand of repair verbatim collected during the diagnosis episodes. In this, it first constructs the fault diagnosis ontology consisting of concepts and relationships commonly observed in the fault diagnosis domain. Next, employ the text mining algorithms that make use of this ontology to identify the necessary artifacts.

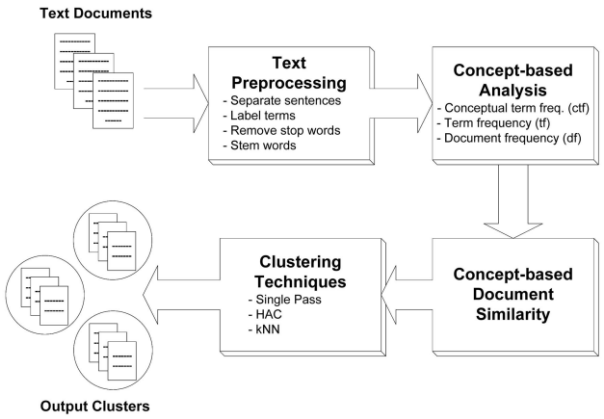

Fig. 1 Concept-based mining model system

The VarifocalReader approach supports users in exploring and understanding complex text documents by visualizing them at various aggregation levels. It provides users with means for navigating the visualization, tracing the current position across all aggregation levels, and offers automatic support from text processing and machine learning algorithms.

PPSGen is a system that automatically generates slides that have good structure and contents from academic papers. The SVR-based sentence scoring model is used to assign an importance score for each sentence. Then, it generates slides from the given paper by using ILP. This approach selects a number of important sentences. After the selection of sentences and phrases can be constructed well-structured slides.

4 ANALYSIS AND DISCUSSION

The text is directly analyzed without using metadata of the text documents. This demonstrates the effect of using concepts on the text mining process. The similarity of text is calculated on sentence-based, document-based, and corpus-based used to compute similarity matrices of documents. The concept-based weighting is one of the factors having importance of a concept. When concept based similarity was introduced using the combined weighting scheme among the tf, ctf, and df, the quality of clusters produced was pushed close to that produced by HAC and k-NN [5].

The typical criterion for text clustering F measurement is used to measure the quality of clustering research projects. For generated cluster c and predefined research topic t, the corresponding Recall and Precision can be calculated as follows: Precision(c, t) =n(c, t)/nc

Recall(c, t) =n(c, t)/nt where n(c, t) is the project number of the intersection between cluster c and topic t. nc is the number of projects in cluster c, and nt is the number of projects in topic t. Comparison of clustering quality of OTMM and TMM, both methods are same. The relations between F measurement and the number of research projects n in these two disciplines can be found. The performance of OTMM is better than that of TMM. Therefore, the OTMM is an alternative for clustering research proposals [6].

The performance of text-driven D-matrix was compared with

the other technology i.e. historical data-driven D-matrix. The metrics of comparison were computed by when a single failure occurred at a given instance. The fault detection is defined as the percent of faults detected by the symptoms at failure modes. It used to determine the fault coverage to determine acceptability. Less than 100% fault detection indicated that there are faults, which cannot be detected [8].
All layers of VarifocalReader are always synchronized. This makes approach less powerful since it is not possible to move to another point in a layer without losing text details. For the class of text analysis tasks, has no disadvantage since they always involved the text source itself, which switch the text details to the new location. VarifocalReader has strength in intradocument analysis. Comparing two text documents works well by this method [11].

TABLE 1: Comparisons between Concept-based mining, Text-driven D-matrix, Ontology based Text Mining, VarifocalReader and PPSGEN

Text Mining Techniques	Advantages	Disadvantages
Concept-Based Mining Model	It efficiently finds significant matching concepts between documents.	This method does not work on web documents.
Ontology Based Text Mining To Cluster	It improves the efficiency and effectiveness of the research project selection process.	This method does not work properly when the number of proposals in one cluster is very large.
Text-driven D-matrix	The text-driven D-matrix improves fault detection and fault isolation rate while exhibiting lower error rate.	The main drawback of the text-driven D-matrix to enhance its performance extension of LDA model is incorporated.
VarifocalReader	The presented approach can be flexibly extended with additional text mining and NLP methods as well as corresponding visual, interactive feedback loops.	Missing feature is a language guesser.
PPSGen System	It generates the slides automatically and slides are well structured that can easily understand by humans.	The main drawback of the method is that generates the slide with any slide design which cannot be changed.

5 PROPOSED METHODOLOGY

Many mining strategies have been used, such as the Concept-based mining, Text-driven D-matrix, Ontology based Text Mining, PPSGEN and VarifocalReader, each of which has its own special characteristics. In text clustering, selecting important feature is important, which has critical effect on output of clustering algorithm. Hence by using concept based mining text clustering can be enhanced. Also, it is used for matching the concept between documents with respective to their semantics of sentences. In this concept based similarity measure is used on the documents for clustering the set of similarity. By using Text driven D-matrix method, important term from the text document can be extracted from unstructured text. Hence text driven D matrix can be used for information retrieval from document. In PPSGen, sentence importance is a key step for analyzing importance of each sentence in given paper. Support vector regression model is used for learning importance of each sentence, because its regression score is easier to use for sentence selection. Hence the proposed method for information retrieval from text documents which is the combination of text driven D matrix, concept based similarity model and SVR model. When a collection of text document is given, then important sentences are selected those documents with SVR model and then the similarity of those sentences is checked by concept based similarity measure. Finally the important information is retrieved from those similar text documents by using text driven model.

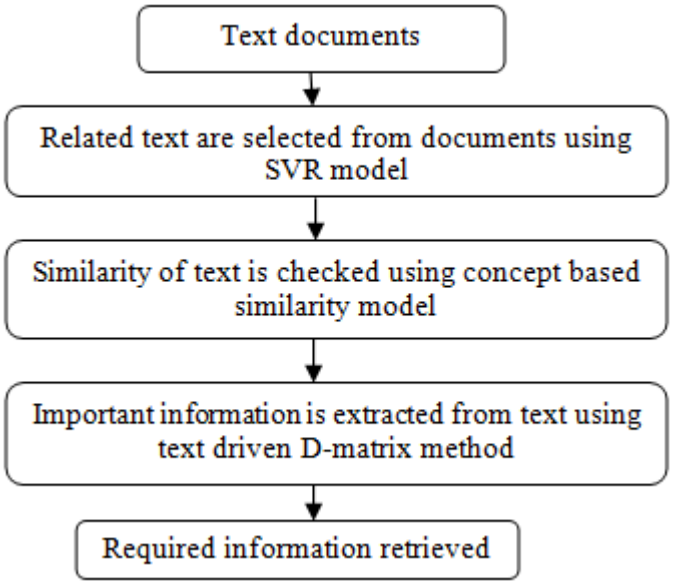

Fig. 2. Proposed framework

6 EXPECTED RESULTS

The expected result for this proposed method will be that, it will give efficient information from no. of text documents which contains similar information. This method will first find the similar text documents from given documents and then will find important information for which user has demanded. This proposed method is more efficient than other information retrieval methods applied earlier. This method consumes less time than any other method.

7 CONCLUSION

This paper focused on efficient method for text retrieval from given document. This paper proposed a novel approach, which consist of the combination of Concept-based mining, Text-driven D-matrix, PPSGEN and SVR model that will help user to retrieve information from similar topic text documents among various given text documents.

The novel approach proposed for text retrieval will derived better and efficient result in terms of retrieval time and the contextual text which will be more appropriate as compared to the existing methodologies.

REFERENCES

[1] T. Hearst, "Untangling text data mining," in Proc. 37th Annu. Meeting Assoc. Comput. Linguist., 1999, pp. 3–10.

[2] Q. Tian, J. Ma, and O. Liu, "A hybrid knowledge and model system for R&D project selection," Expert Syst. Appl., vol. 23, no. 3, pp. 265–271, Oct. 2002.

[3] D. Oelke, D. Kokkinakis, and M. Malm. Advanced visual analytics methods for literature analysis in natural language processing. Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '12, pages 35–44, Stroudsburg, PA, USA, 2012.

[4] S. Shehata, F. Karray, and M. Kamel,"Enhancing Text Clustering Using Concept-Based Mining Model,"Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2006.

[5] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," IEEE Transaction on Knowledge And Data Engineering, VOL. 22, NO. 10, OCTOBER 2010.

[6] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transaction on System, Man, And Cybernetics—Part A: System And Human, VOL. 42, NO. 3, MAY 2012

[7] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. Kavuri, "A review of process fault detection and diagnosis Part I: D matrix methods," Comput. Chem. Eng., vol. 27, no. 3, pp. 293–311, 2003.

[8] Dnyanesh G. Rajpathak and Satnam Singh, "An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text," IEEE Transaction on System, Man, And Cybernetics: System, VOL. 44, NO. 7, JULY 2014.

[9] V. Vapnik, Statistical Learning Theory. Hoboken, NJ, USA: Wiley, 1998.

[10] T. Shibata and S. Kurohashi, "Automatic slide generation based on discourse structure analysis," in Proc. Int. Joint Conf. Natural Lang. Process., 2005, pp. 754–766.

[11] Yue Hu and Xiaojun Wan, "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers," IEEE Transaction on Knowledge And Data Engineering, VOL. 27, NO. 4, APRIL 2015.

[12] Steffen Koch, Markus John, Michael Worner, Andreas Muller and Thomas Ertl, "VarifocalReader – In-Depth Visual Analysis of Large Text Documents," IEEE Transactions on Visualization and Computer Graphics, VOL. 20, NO. 12, DECEMBER 2014.