Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Novel Analysis Technique Using Text Mining for Scientific Papers

Dheeraj Chandra Murari* Vinod Kuamr Verma
Computer Science & Engineering, Bipin Tripathi Kumaon Institute Of Technology
Dwarahat , Almora (UTTARAKHAND)
263653, India

Abstract: Nowadays many students, researchers are publishing their ideas in text format online through papers publication. In doing their researches, they need to read and analyze multiple research papers, e-books and other documents and then determine what they contain and discover knowledge from them. Many available resources are in the form of unstructured text format of long text pages which require a lot of time to read. In this paper, we propose an efficient scientific papers analysis using text mining that facilitates quick analysis of many research papers. Our approach applies clustering algorithm to group similar papers based on their topics for fast access and analysis. In clustering, the proposed approach uses compound word feature in selection of initial centroids. Conducted experiments show that our approach is effective and can help readers or analyzers know the content of many scientific research papers in a short time.

Keywords: text mining, information extraction, sentence extraction, text clustering, similarity measure, paper analysis.

I. INTRODUCTION

Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents [11]. Text mining methods have been developed and applied in all fields, where large amounts of data were available: analysis of large archives of news documents, internet documents. Researchers have put a lot of interest on mining data that are in the form of structured format where they assume that the information to be mined is already in the form of a relational database [22]. Unfortunately some research papers, e-books and new articles are in the form of unstructured format which is not easy to apply data mining or knowledge discovery directly. A large amount of data available on internet is in the form of unstructured text. This information is being analyzed by many different people for different purposes like knowledge discovery, for decision-making and knowledge management through text mining.

In this paper, we combine both information extraction and text clustering and propose an efficient scientific papers analysis that can help researchers or a leaders to know the content of many scientific research papers in a very short time and can facilitate do a quick analysis of them. Document clustering (also known as text clustering) is one of the most important text mining methods that are developed to help users effectively navigate, summarize, and organize text documents[4]. Document clustering has many purposes including expanding a search space, generating a summary, automatic topic extraction, browsing document collections, organizing information in digital libraries and detecting topics[17]. In our approach it has the purpose of browsing paper collections in which related or similar scientific papers are grouped together in a cluster. In this paper, the content of many scientific papers are processed and stored in summary format, which is database that contains main information(information mainly needed by the reader) of papers and then a clustering algorithm is applied to group related papers together for other processing. The summarized, structured and similar grouped papers will help the reader to easily get many similar papers' information in a very short time without go through the whole papers and can speed up the analysis.

We propose a new cluster clustering algorithm that is based on the use of compound words feature in selecting initial centroids. With this new method, all research papers are stored in databases and their descriptions (Title, authors, main idea/key sentences and results) can be queried. For example the reader or analyzer can see graphically how many papers have been published in a certain field or she/he can know that such author has published such number of papers, etc. Having the needed information in summarized, viewed, structured format with similar (related) papers grouped together in clusters will help the reader to access quickly many papers' information, know their contents in a very short time and then can decide whether to read the paper deeply or not. Text summarization is needed because it present information in shorter way, it saves reader's time and it reduces storage space [20].

II. RELATED WORK

A starting point for computers to analyze unstructured text is to use information extraction [23]. Facing with so large data set, it is very difficult to find the desired information quickly and accurately. It is helpful for users to retrieve the summary of original article instead of original article, with the purpose of finding the desired information efficiently and

accurately [20]. Many text mining methods have been developed for all kind of text data. Those methods can only extract simple words from an unstructured text. Mostly useful information such as names of people, places or organization mentioned in the text is extracted without a proper understanding of the text.

Text clustering has been used in many applications such as text summarization [8, 14], navigation of large document collections and organization of Web search results. Eloize Rossi Marques Seno et al[6] in their experiments on identifying and clustering similar sentences from one or multiple documents of new articles, proposed an evaluation framework based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences. Their approach detects and clusters similar sentences of texts written in Brazilian Portuguese. Their approach is limited on working only for text written in Brazilian Portuguese and also works for only news articles. Few works have been conducted to analyze scientific papers, one of them is the analysis of scientific papers in the field of radiology and medical imaging included in Science Citation Index Expanded and published by Turkish authors done by Erhan Akpınar et al [7]. This study is concerned only with counting the number of published paper per year in certain period by Turkey authors. It doesn't group similar paper and do more analysis.

III. THE PROPOSED APPROACH

A considerable amount of research is being conducted by many people (researchers, graduate students, professors etc) everyday. People doing research have to read many papers so that they can see what others have done and they can contribute to the existing approaches or can propose new ones. They spend a lot time on searching and reading papers related to their research topics. The reader of a scientific paper can be only interested in the main idea of the paper and the results from experiments and he/she doesn't need to read the whole paper. Information extraction can help to automatically extract only the needed information to the reader and text clustering can help to group similar papers together for the benefit of the reader.

We extend our previous proposed approach [9] where title, authors, main content and experiments results) were extracted from a research papers and stored in database [9]. In this paper, a new clustering method is proposed and then applied to the resulted database so that papers are grouped according to their topic, which means that similar or related papers are grouped together. Our clustering algorithm uses compound words feature in selecting initial centroids. The output of clustering methods shows papers in different topics and presents also graphically the number of published papers from different fields, thus speed up the analysis. Figure 1 shows the proposed approach.

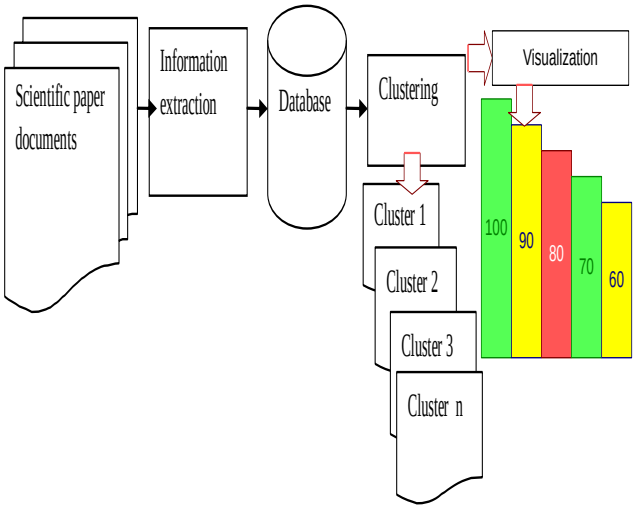

Fig1. The proposed approach framework.

IV. PAPER CLUSTERING

The clustering of text documents is a central technique in text mining which can be defined as grouping documents into clusters according to their topics or main contents [17]. A *cluster* is defined as a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Congnan Luo et al[4] defined the problem of document clustering as follows: given a set of documents, they can be automatically grouped into a predetermined number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics.

We apply data mining technique (clustering) on the constructed database to group similar (or related) paper together. This technique is applied to a multi word attributes database as it is made of sentences extracted from text documents (research papers). In our case we are concerned with scientific papers and a cluster refers to a group of related papers(the papers are similar since they talk on similar field or topic). In data mining area(super topic), for example all papers related to text mining, database mining, multimedia mining, biological data mining and World wide web can be grouped in different clusters, since they are different topics.

V. SIMILARITY MEASUREMENT

The most important factor in a clustering algorithm is the similarity measure. All clustering algorithms are based on similarity measures and each clustering method use a similarity function. Before clustering, a similarity/distance measure must be determined [1]. Many methods have been proposed to measure the similarity between words, sentences paragraphs and documents. Document similarity is often represented by the Vector Space Model (VSM). Documents are represented by the bag of words, and the meanings of documents are presented by vectors. A well-known similarity measure is the cosine function, which is widely used in document clustering algorithms and is reported performing very well [21]. Sentences are represented by a vector of weights while computing cosine similarity. The cosine function can be used in the family of k-means algorithms to assign each document to a cluster with the most similar cluster centroid in an effort to maximize the intra-cluster similarity. In order to achieve high efficiency of our method, we have also chosen cosine similarity as it has found performing well.

VI. CLUSTERING ALGORITHM

In many previous methods, K-means has been used for many methods and has reported to perform well [1, 4]. The k-means algorithm is based on the idea that a centroid can represent a cluster. The k-means algorithm starts with initial cluster centroids, and sentences are assigned to the clusters iteratively in order to minimize or maximize the value of the global criterion function [4]. We adopt this algorithm in grouping similar papers of our constructed database. The concept of clustering used in this work is similar to the one used in [1, 6, 13]. The difference relies on estimating the initial K centroids i.e. sentences centroids.

a) *Selection of initial cluster k centroids*

It is not easy to determine the initial centroids of clusters in a text database. Many methods have been proposed in estimating number of clusters centroids such as random selection and buckshot. The random algorithm randomly chooses *k* documents from the data set as the initial centroid.The buckshot [5] algorithm picks $\sqrt{nk}$ documents randomly from the data set of *n* documents, and clusters them using a clustering algorithm. The *k* centroids resulting from this clustering become the initial centroids. It is known that the clustering algorithms based on this kind of iterative process are computationally efficient but often converge to local minima or maxima of the global criterion function. There is no guarantee that those algorithms will reach a global optimization [4]. We wish to have a good set of initial cluster centroids in order to overcome this problem. We propose a new centroids selection method based on weight of compound words that composed all paper titles. The words are assigned weight according to their frequencies. The frequent words are given higher weight. If the word is more frequent, it means that it appears in many paper titles. The computer related topics and other engineering subjects are made of compound words for example, artificial intelligence, civil engineering, data mining, wireless network, information systems, text mining, information extraction, computer network, information technology, etc, so dealing with those topics or subjects it is better to take them as compound words instead of single words, because they are composed of many words and their meanings are based on those compound words. In a text document, a compound word meaning is more meaningful than a simple word meaning.

We use the compound word feature in order to achieve the best results, meaning that if compound word is frequent in many titles, it is likely that it is represents one of the topics that have been covered by many papers. The method starts at first record i.e. first paper title and counts the frequency of each word of the title's attribute. After counting the frequency of each word, all words having more frequency are checked if there are used as compound words in titles, if yes, their frequencies are counted as compound word and those having many frequencies are selected to represent the topics in clusters, that are centroids. In the case that more frequent
word is not found in compound words, the title containing this word will also selected as a centroid. In short the proposed method work like this:

After removing the stop words, the algorithm follows the
following steps

Step 1. Count the frequency of every word in
title attribute.
Step 2. Select words having more frequency.
Step 3. Check if those words are used as
compound words
Step 4. Count the frequency of those compound Words
have many frequencies
Step 5. Select titles whose compound words.
Step 6. Selected titles are used as initial centroids.

b) Titles clustering

After selecting k initial centroids(records from database), each title is assigned to a cluster based on a distance measure (between the title and each of the k centroids), then k centroids are recalculated.

This step is repeated until all titles are assigned to clusters. Cosine similarity [3] measure is used to calculate similarity between titles. The following is the clustering algorithm:

Input: Database of n records and K centroids clusters

Output: Database of n records grouped in different K clusters according to their topics.

Steps

Step1. Select *k* records (centroids), records having titles that have been selected in selecting initial centroids.

Step 2. Select one record *r* from remaining records

Step 3. Compute cosine similarities between r and *k* centroids

Step 4. Put *r* in the closest cluster and recompute the centroid.

Step 5. Repeat Steps 2 to 4 until the centroids don't change

VII. CLUSTERING RESULTS

Based on our proposed method that estimates initial k centroids based on frequency of compound words of paper titles, we conducted experiments to cluster similar papers using K means method and cosine as similarity measure. We conducted our experiment on 200 paper titles that was extracted and stored in database of our previous method[9]. The obtained results are compared with the previous works [1, 4, 6]. Table 1 shows the results of our proposed method and Figure 2 and 3 show details of the comparison results of both methods. As it is seen in both figures, our method outperforms previous works in term of F-measure, entropy and purity. Figure 2 depicts the comparison results of our method with [4] in term of F-measure and Figure 3 depicts the comparison results of our clustering method with [1, 6] in terms of entropy and purity. The best performance of our approach is based on best selection of initial centroids while clustering the papers.

In the below table a comparison is depicted in different domains of a relation.

Table 1.1

papers range	F-measure	Entropy	Purity
1-40	**0.874**	**0.255595**	**0.880**
41-80	**0.800**	**0.397331**	**0.900**
81-120	**0.896**	**0.21366**	**0.900**
121-160	**0.902**	**0.202153**	**1.000**
161-200	**0.916**	**0.194209**	**1.000**
average	**0.878**	**0.253**	**0.936**

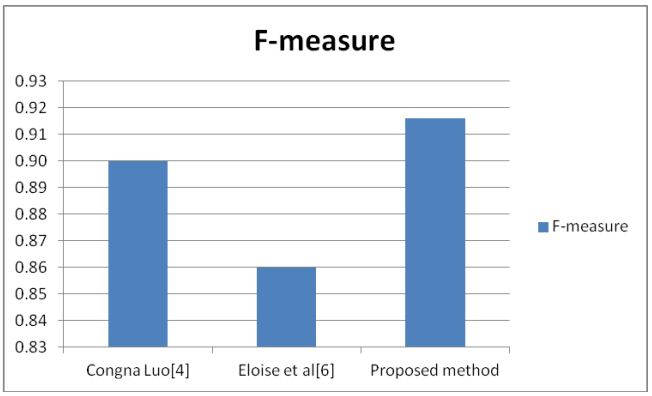

Fig 2: F-Measure results comparison with different methods

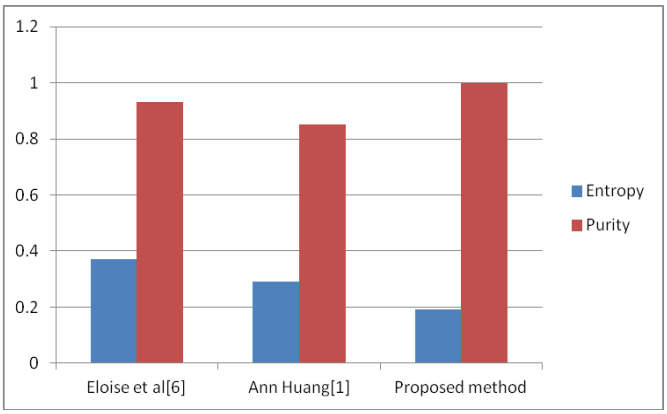

Fig 3. Entropy and Purity results comparison with different methods

In general, the smaller the Entropy value, the better the clustering result, or the larger Purity and F-measure values the better the clustering result.

VIII. CONCLUSIONS

In this paper we proposed an efficient scientific papers analysis approach using text mining that can help do quick analysis of research papers. It extracts title, author names, the main content and the experiments results from scientific research paper. The extracted information is then stored in database where all related papers are grouped in same cluster for fast access and analysis. The main contribution of this paper is that the reader can get the paper's information in summarized, viewed, structured format for easy, and fast access, so that he or she can read the content of the paper in a very short time reducing time and space. At the same time the reader can get all related papers grouped together quickly and this can help do quick analysis. Our clustering experiment results show an average of 87.8%, 0.252 and 0.936 of F-measure, entropy and purity respectively as shown in table 1. The clustering results of our proposed approach give the values of 91.6%, 0.19 and 1 of F-measure, entropy and purity respectively for the best case. It outperforms also previous proposed methods [1, 4, 6] as shown in Figure 2 and 3. The best performance of our clustering approach is based on best selection of initial centroids while clustering. We are planning to extend our approach so that it can extract sentences in other articles and can extract information from the whole document. This will result in excellent performance of the method in terms of accuracy and precision and can facilitate further processing.

REFERENCES

[1] Anna Huang, "Similarity Measures for Text Document Clustering*", NZCRSC*, pp. 49- 56, 2008

[2] Atika Mustafa, Ali Akbar, and Ahmer Sultan. " Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and and Categorization". *International Journal of Multimedia and Ubiquitous Engineering* ,vol. 4, no. 2, pp. 183-188, 2009.

[3] Bilal Zaka. "Theory and Applications of Similarity Detection Techniques", Phd Dissertation. ***I****nstitute for* ***I****nformation Systems and* ***C****omputer* M*edia* (IICM) Graz University of Technology, pp. 251-260, 2009.

[4] Congnan Luo , Yanjun Li and Soon M. Chung. "Text document clustering based on neighbors". *Journal of Data & Knowledge Engineering*, 69, pp. 1271–1288, 2009.

[5] Cutting D.R., D.R. Karger, J.O. Pedersen, J.W. Tukey " Scatter/gather: a cluster-based approach to browsing large document collections". *in: Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. pp.318–329, 1992.

[6] Eloize Rossi Marques Seno and Maria das Graças Volpe Nunes. "Some Experiments on Clustering Sentences of Texts in Portuguese". A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, © Springer-Verlag Berlin Heidelberg. pp. 133–142, 2008.

[7] Erhan Akpınar, Muşturay Karçaaltıncaba. "Analysis of scientific papers in the field of radiology and medical imaging included in Science Citation Index Expanded and published by Turkish authors". *Diagn Interv Radiol* vol.16, pp. 175–178, 2010

[8] Fasheng Liu and Lu Xiong. "Survey on Text Clustering Algorithm". In Proceeding of International conference of Software Engineering and Service Science (ICSESS), IEEE, pp.196-199, 2011

[9] Hanyurwimfura Damien, Bo Liao, Humphrey Njogu,Eustache Ndatinya, "An automated Cue Word based Text Extraction", *JCIT: Journal of Convergence Information Technology*, vol. 7, no. 10, pp. 421-429, 2012

[10] Jain A.K, R.C. Dubes, "Algorithms for Clustering Data". Prentice Hall, Englewood Cliffs, 1988.

[11] Jiawei Han and Micheline Kamber . "*Data mining Concepts and techniques*", Morgan Kaufmann. San Francisco, Second edition, 2006

[12] Juan José García Adeva and Rafael Calvo, "Mining Text with Pimiento", *University of Sydney*

[13] Warhurst Junsheng Zhang, Yunchuan Sun, Huilin Wang and Yanqing He. "Calculating Statistical Similarity between Sentences". *Journal of Convergence Information Technology*, vol. 6 no. 2, pp. 22-34, 2011.

[14] Kamal Sarkar. "Sentence Clustering-based Summarization of Multiple Text Documents". *International Journal of Computing Science and Communication Technologies*, vol. 2 no. 1, pp. 225-235, 2009.

[15] Li, Y., Luo, C., & Chung, S. M. "Text clustering with feature selection by using statistical data*". IEEE Transactions on knowledge and Data Engineering*, vol. 20, pp. 641–652, 2008

[16] Liping Jing · Michael K. Ng · Joshua Z. Huang. "Knowledge-based vector space model for text clustering*". Knowl Inf Syst*, vol. 25, pp. 35–55, 2010.

[17] Ramiz M. Aliguliyev. "A new sentence similarity measure and sentence based extractive technique for automatic text summarization". *Expert Systems with Applications*, pp. 7764–7772, 2009.

[18] Richard Khoury. "Sentence Clustering Using Parts-of-Speech". *International Journal of Information Engineering and Electronic Business*, vol. 1, pp. 1-9, 2012.

[19] Rosell, M., Kann, V., Litton, J. "Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications*". in: proceeding of International Conference on Natural Language Processing*, pp. 207–216, 2004.

[20] Sicui Wang, Weijiang Li, Feng Wang and Hui Deng. "A Survey on Automatic Summarization". *2010 International Forum on Information Technology and Applications*. pp.193-196, 2010

[21] Steinbach M., G. Karypis, V. Kumar, A comparison of document clustering techniques, *in: KDD Workshop on Text Mining*, 2000.

[22] Tamara Polajnar. "Survey of Text Mining of Biomedical Corpora". pp. 1-21, June, 2006.

[23] Vishal Gupta and Gurpreet S. Lehal. "Survey of Text Mining Techniques and Applications". *Journal of emerging technologies in web intelligence*, vol. 1 no. 1, 2009.

[24] Yunhua Hu, Hang Li, Yunbo Cao and Dmitriy Meyerzon. "Automatic Extraction of Titles from General Documents using Machine Learning". *Information Processing and Management: an International Journal archive*, vol. 42, no. 5, pp. 1276 – 1293, 2006.

