A Survey of Text Document Clustering Methodologies based on Similarity Measure

R.Saranya[1] V.Sharmila[2] P.Balamurugan[3] R.Latha[4]
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]K.S.R College of Engineering, Tamilnadu, India

Abstract— Text data mining is research technologies to mine valuable knowledge from massive collections of documents and to improve a system to offer information and to support in decision making. Clustering is an automatic wisdom methodology aimed to combine a set of objects which are similar to each other. Text clustering has turned into a challenging process in recent centuries because of the massive quantity of unstructured data is presented in several formations. In Text document clustering grouping of text documents arises based upon their similarity. There are many rapid and high-excellence document clustering algorithms available which play a main role in efficiently establishing the information. In this paper we are going to discuss various methodology of clustering which is based on the document similarity.

Key words: Data mining, Text mining, Document clustering, Similarity measure

I. INTRODUCTION

Knowledge mining is the process of determining valuable patterns and knowledge from massive amount of data. Data sources can include databases, data warehouses, web, and other information sources that are streamed into the system dynamically. Knowledge extraction consist of an amount of technical tactics, such as clustering, data summarisation, classification, finding addicted networks, analysing fluctuations and spotting anomalies [4].

Natural Language Processing (NLP) is field of computer science, artificial intelligence concerned with the communication between computer and human language. Text mining is the analysis of data contained in NLP. Text mining discovers fresh, formerly unknown information by applying various procedures from natural language processing and knowledge mining [1].useful information can be derived from the unstructured data is called text mining.

Clustering is classical knowledge extraction process which is unsupervised learning model that means training data (previous work) and response variables are previously unknown. By using statistical concept, dataset can be divided into sub datasets. It tries to find the internal structure in unlabelled data. In the set of clusters, similarity degree is high in the intra cluster and low in the inter cluster [4].

Most documents can be represented by Vector Space Model (VSM) [1], [2], which is a broadly used data depiction for text clustering. Document can be represented by feature vector of the word/phrase in VSM. Feature vector holds term frequencies of the terms in the document. The similarity among documents can be calculated by lot of similarity methods that are based on the feature. Document similarity is the process of finding relationship among the documents. It takes presence and absence of the features into consideration to find out the similarity. If the presence and absence feature increases then the similarity degree could be decreases.

The remaining part of this paper is structured as follows: Section 2 contains literature survey of document clustering based on similarity measure, In Section 3 describes the identification of the problem, and the last section summarizes the conclusion and future works.

II. LITERATURE SURVEY

A. Measuring Sentence Similarity for Text Summarization

According to Ramiz M.Aliguliyev [5], Text document summarization provides compressed version of original document. It plays a major role in the process of retrieving the information. Summarization can helps to understand the entire document. Multi document summarization is the process of producing summary from various documents (document set).In the projected model considers sentence clustering that is useful for the document summarization. The result can be depends on optimized function and also similarity measure. Sentence clustering is formed by decomposing the document into sentences and calculating the number of words present into it where documents are represented by vector space model. Number of clusters is estimated for the efficient document summarization.

B. Mining Model based on Concept

Shady Shehata, Fakhri Karray, and Mohamed S. Kamel [6] proposed method is different from the traditional technique which is based on statistical analysis of term (either word or phrase). Term frequency considers the reputation of the term inside particular document only. Though, dual terms can contain identical frequency in their particular document but one term develops the sentence than the former term. In this situation, mining can consider terms that defines the concept of the sentence which results creation of topic. In the projected, concept based mining can takes sentence semantics and terms that grasp the concept. Non-significant terms are ignored. In proposed text data mining model contains four analysis schemes such as

- Concept Analysis based on Sentence,
- Concept Analysis based on Document,
- Concept-Analysis based on Corpus,
- Concept-Based Similarity Measure

This similarity degree takes full benefit of consuming the analysis of concept actions on the sentence, document, and corpus levels in computing the similarity among documents.

C. Semantic similarity of short text

According to [7] [10], depicts similarity of short text in different kind of processing. Jesus Oliva, Jose Ignacio Serrano, Maria Dolores del Castillo, and Angel Iglesias [7] suggested that measuring the short text and the sentences by using syntax based measure. Classical methods are used to

find the similarity short text by the semantic methods. That method is not suitable for grammatically complete sentences and long text similarity. The projected SySTSS method considers that the significance of sentence is depends upon not only the meaning of its distinct words, but also the structural way the words are pooled. The method is based on syntactic and semantic information which are obtained from lexical database and deep parsing process. Bojan Furlan, Vuk Batanovic, and Bosko Nikolic [10] suggested that short text are broadly used in the internet in many forms such as product description, image news lines, tags of another web page, etc. The projected measure is capable of depicts the similarity between two short texts. That short text semantic similarity measure is independent to any language (LInSTSS).It can be suitable in the case when no bulky, freely available, automated linguistic assets can be found for particular language. Modularity is the key advantage of this method because it can regulate any language easily.

D. Clustering Proposal for Research Project Selection based on Ontology

Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu [8] suggested that proposed method is based on the selection of research project. It is a main process for government and private funding agencies. Grouping of research project proposal is difficult task due to group them based on their similarity in particular disciplines. Similar research discipline can be conceded by the manual matching. The accuracy of the cluster is not high in this matching. Because there may be a chance to ignore the related proposal, and this method is suitable for English language. The proposed ontology mining model is efficient to use English as well as Chinese research proposal. Ontology is an information warehouse where the terms are defined. Also it consists of collection of idea, axioms and associations. It contains four phases, they are: Ontology encloses the projects funded in last few years, new proposals are classified, clustering of new proposal based on the self-organized mapping algorithm. Finally, each cluster can hold large quantity of proposal then decompose into sub clusters.

E. Multiple Viewing Point used to Find Similarity Degree for Clustering

According to Duc Thang Nguyen, Lihui Chen, and Chee Keong Chaan [9] suggested that outmoded dissimilarity/similarity degree uses simply a only one viewpoint, that is called origin point, while the later exploits many diverse viewpoints, which are objects estimated not to be in the equivalent cluster with the two objects being dignified. Cosine similarity is based on the origin that means single viewing point. It takes the cosine of angle between two vectors, which starts from the origin. Using multiple viewpoints, more useful valuation of similarity could be accomplished. The two objects to be dignified essential to be in the identical group, though the points from where to launch this ability must be external of the cluster. This scheme is known as multiple viewing points used to find similarity degree. Multiple viewing point similarity degree is highly appropriate for text documents than compared with other similarity degrees. The aim of this similarity degree is to measuring similarity among data objects in any domain such as sparse and highly dimensional documents. Based on this benchmarks clustering algorithms are framed, that are rapid handling and accessible like k-means but proficient of providing great excellence.

F. Word Indexing Model based on Context For Document Summarization

Pawan Goyal, Laxmidhar Behera, and Thomas Martin McGinnity [11] suggested that document summary is created by using important ideas in the document or document set. It is shortened version of the particular document. Traditional summarization methods are used similarity among the sentences, where the document and sentences are indexed. In those kind of summarization do not considers the context. The proposed (Context sensitive document indexing) method is based on the Bernoulli model of randomness, where the co-occurrences of two terms probability in the corpus can be found. This model uses lexical relationship among the terms to provide sensitive weight to the terms which is presented in the document. Sentence similarity matrix can be formed by using the indexing weight of the every term in the document. Sentence extraction can be done by combining similarity measure for sentence with the graph based ranking model.

G. Hybrid XOR for document clustering

Vangipuram Radhakrishna, C.Srinivas, and Dr.C.V.Guru Rao [12] suggested that software reuse is nothing but creating new software component by using existing software component from previous systems. Reuse eliminates the cost of production. If clustering is applied to software components then it will group similar feature into one group and dissimilar feature into another group. Hybrid XOR function defined to govern the degree of similarity between set of documents or software components. In the clustering algorithm takes input as similarity matrix and produces output as set of clusters that formed dynamically. The cluster count is pre-defined similarity matrix order n-1 by n for collection of n document sets or software component, Where the rows of matrix depicts documents and the columns of matrix depicts unique set of frequent items. The computation of this method is very simple and produces high accuracy clusters comparing to other techniques.

H. Grouping of Documents based on Similarity Assessment

Tanmay Basu and C.A.Murthy [13] suggested that clustering of documents is the process of grouping similar documents and separating similarity documents. The task of classifying corpus is very difficult because it consists of huge documents and the documents are based on high dimensional domain and it contains sparse data. The proposed hierarchical clustering method is the combination of traditional k-means clustering and new hierarchical methodology. In this distance function is used for discover the distance between the hierarchical clusters. Nature of corpora can be easily identified by using the distance function. The optimal cluster partition can be done by the proposed method. Traditional k-means requires the number of clusters before executing the algorithm. Initial random selection of the k-centroids can be ignored. F-measure and normalized mutual information are used to estimate the efficiency of the model.

III. PROBLEM IDENTIFICATION

Document similarity is the process of finding relationship among the documents. It takes presence and absence of the features into consideration to find out the similarity. If the presence and absence feature increases then the similarity degree could be decreases. The similarity degree should increases when the difference among two non-zero significance of particular feature reduces. Severe challenges for measuring similarity are high dimensionality and sparsity. High dimensional data are data characterized by few dozens to many thousands of dimensions. Data on medicine, health care, biology are examples of high dimensional domains. The high dimensionality of data is a vastly critical factor for clustering task. When dimensionality is high, the volume of space rises so fast that the accessible data becomes sparse. This problem is known as curse of dimensionality. High dimensional data could likely contain unrelated features, which may obscure the effect of appropriate one. Because of unrelated features the accuracy of clusters could be affected.

IV. CONCLUSION AND FUTURE WORK

This survey paper is based on various similarity measures for document clustering. Severe challenges for measuring similarity are high dimensionality and sparsity. That can be highly critical factor for clustering task. By eliminating high dimensionality we can improve the quality of clusters. As a future work, accuracy of cluster which was not sufficient in the previous works will be improved by reducing the high dimensionality of the domains. And also we will concentrate on the sematic associations between the feature vectors and optimal clustering procedures.

REFERENCES

[1] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.
[2] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[3] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.
[4] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.
[5] Ramiz M.Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," ELSEVIER Expert Systems with Applications, pp. 7764-7772, 2009.
[6] Shady Shehata, Fakhri Karray, Mohamed S. Kamel, "An Efficient Concept-Based Mining Model For Enhancing Text Clustering," IEEE Transactions On Knowledge And Data Engineering, vol. 22, no. 10,pp.1360-1371, October 2010.
[7] Jesus Oliva, Jose Ignacio Serrano, Maria Dolores del Castillo, Angel Iglesias, "SyMSS: A syntax – based measure for short-text semantic similarity," ELSEVIER Data and Knowledge Engineering,pp. 390-405,January 2011.
[8] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, Ou Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection," IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, vol. 42, no. 3, pp. 784-790,May 2012.
[9] Duc Thang Nguyen, Lihui Chen, Chee Keong Chaan, "Clustering with Multiviewpoint-Based Similarity Measure," IEEE Transactions on Knowledge And Data Engineering,vol. 24, no. 6, pp. 988-1001, June 2012.
[10] Bojan Furlan, Vuk Batanovic,Bosko Nikolic, "Semantic similarity of short texts in languages with a deficient natural language processing support," ELSEVIER Decision Support Systems, pp. 710-719, February 2013.
[11] Pawan Goyal, Laxmidhar Behera, Thomas Martin McGinnity, "A Context-Based Word Indexing Model for Document Summarization," IEEE Transactions On Knowledge And Data Engineering, vol. 25, no. 8,pp. 1693-1705, August 2013.
[12] Vangipuram Radhakrishna, C.Srinivas, Dr.C.V.Guru Rao, "Document Clustering Using Hybrid XOR Similarity Function for Efficient Software Component Reuse," ElSEVIER Information Technology and Quantitative Management, pp. 121-128, 2013.
[13] Tanmay Basu,C.A.Murthy, "A Similarity assessment technique for efficient grouping of documents," ElSEVIER Information Sciences, pp. 149-162,March 2015.

