A Review on Document Clustering Using Concept Weight

Sapna Gupta[1], Prof. Vikrant Chole[2], Prof. Dr.A.Mahajan[3]

1 Department of CSE, GHRAET,RTM Nagpur University,
2 Department of CSE, GHRAET,RTM Nagpur University,
3 Department of CSE, PIET,RTM Nagpur University

Abstract

Traditional document clustering techniques are mostly based on the number of occurrences and the existence of keywords. The term frequency based clustering techniques takes the documents as bag-of words while ignoring the relationship between the words. Similarly Phrase based clustering technique only captures the order in which the words appear in a sentence instead of determining the semantics behind the words. Considering the drawbacks of such system this paper proposes a concept based clustering technique. The ideology behind this concept is uses Medical Subject Headings MeSH ontology for extracting the concept and the concept weight calculation is done by its identity and relationship with its synonym. The method used for clustering documents on Semantic is called K-medoid algorithm through which the results are analyzed.

Keywords: *Document Clustering; Ontology; semantic similarity; concept weight*

I. INTRODUCTION

With the increase in popularity of the internet, the accessibility of data is growing day by day.The data is not available in the structured form,so there is a need to manage this large amount of data according to user query. Huge volume, high dimensionality, complex semantics and sparsity make the unstructured text document clustering process as the most difficult. To overcome this problem document clustering technique is used .It groups all the documents so that the one which are similar is under one group and dis- similar documents under one group.

The bags-of-words used for clustering ignores the semantic relation between the words like synonyms, polysemy etc and the results are not satisfactory. So to resolve this issue concept based clustering is used. In order to identify and extract the concepts in a document, core ontologies can be integrated as background knowledge into the process of clustering the plain text documents. It is used to measure the conceptual similarity between the terms. Biomedical ontologies enable us to resolve many linguistic problems when text mining approaches handle biomedical literature. In this paper we apply K-Medoids algorithm. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters.

After the clustering process the page is ranked using page rank algorithm.

In this paper, in section 2 some earlier related work is explained In section 3 the disadvantages of the existing systems are enlisted named as problem definition. In section 4, the objectives are given which may be satisfied in future. Finally in section 5, the conclusion with some future work is given.

II.REVIEWOFRELATEDWORKS

Clustering of documents in a distributed manner has become the main point of attraction of researchers because of demand of scalability and efficiency.The cluster quality is improved by incorporating its background knowledge through the domain ontologies with the plain text documents are proven by the recent works.

In the paper [1] of Sadiq, A.T and Abdullah, S.M shows the text categorization technique in which text documents are grouped into one or more predefined categories based on their contents .it consists of three main steps: text document representation, performance evaluation and classifier construction .In the first step set of pre classified documents is provided which is then pre- processed in order to be split into feautres. These features are weighted based on the frequency of each feature.The non-informative features are then eliminated and the remaining feature are standardized by reducing a feature to its root with the help of stemming process. As large number of features still remains even after the stemming process and non-informative features removal , specific thresholds is applied to the system to extract distinct features which represent that text document.

In the second step, the text categorization model is built by learning distinct feature.They represent all the pre-classified text documents for each sub -categories.This can be achieved by supervised categorization technique which is also called as rough set theory.

Thereafter, the model uses a pair of precise concepts that are called the lower and upper approximations to group any test text document into one or more of main categories and sub- categories.In the final step ,performance is evaluated.
The categorization techniques cannot directly process the text documents in their original form,so each input text document should be converted into compact representation of its content by using the preprocessing steps

In the paper [2] of Hmway Hmway Tar and Thi Thi Soe Nyunt shows the importance of document clustering process as the popularity of internet grows.Clustering process is becoming essential as it is useful in generalizing large amount of information.This technology focuses on term weight calculation.For achieving more accurate document clustering ,more informative feature including concept weight are required.
As the large document is containing irrelevant or reduntant feature ,to reduce them feature selection process is important for clustering process as it may misguide the clustering results.
To resolve this issue the system presents the concept weight for text clustering with the help of K-means algorithm so that importance of words of a cluster can be identified by the weight values.Due to this it resolved the semantic problem in specific areas to a certain extent.The proposed method is effective and showed its practical value when experimented using dessertation papers from google search engine.

By using domain-specific ontology, the proposed system is able to categorize documents on the basis of the concept level. Concept weighting scheme tries to capture some aspect of the Semantic Web. When weighed using concept, the clustering system can improve the accuracy and performance of text documents.

The paper of [3] Rekha Baghe and Dr. Renu Dhir presents a technique of document clustering based on frequent concepts.The FCDC technique works with frequent concepts rather than frequent items. The other clustering deal with documents as bag of words and ignore the important relatonship between words like synonyms.
The Proposed FCDC algorithm utilizes the semantic relationship between words to create concepts.
It exploits the WordNet ontology too create low dimensional feature. It uses a Hierarchical approach to cluster the text documents. FCDC is more accurate, effective and Scalable whencompared with other clustering algorithm like UPGMA and FIHC. It provides high dimentionality and accuracy.Clustering accuracy reduces as cluster size varies by large scale.

In the paper of A. A.Kogilavani, B. Dr.P.Balasubramanie [4] they shows the importance of clustering the document. As the amount of diverse data is growing now a days information and knowledge management has become a real challenge.
Even though larger amount of data are merely available,easier to access but the most appropriate form is still difficult.Particularly the medical field suffers from the problem of the information retrieval.As physicians and researchers in medical and biology needs quick and efficient access to up- to-date information.The proposed system combines both document and text summarization technique.
In the proposed system the user query is mapped with synonyms and semantically related concepts using MeSH ontology knowledge source. Based on the revised query medical documents are retrieved from trustworthy online sources and these documents are then clustered. Ontology can improve document clustering performance with its concept hierarchy knowledge,it requires less space to present main ideas in document.it improves efficiency

surekha and S.C.Punitha[5] in the proposed paper compares the performance of algorithm based on K-Means and DBScan Clustering. Ontology by using a concept weight is introduced which is calculated by considering the correlation coefficient of the word and probability of concept.
The experimental the results showed that the inclusion of ontology increased the efficiency of clustering . The proposed document clustering using ontology combines concept weighting or semantic weighting and two clustering algorithms, namely, K-Means and DBScan algorithm.

The system consists of three steps for clustering, namely, calculating concept weight based on the ontology ,document preprocessing, and clustering documents with the concept weight.

The performance of the two algorithms was analyzed using the precision, recall, F measure and Accuracy. The F-measure is calculated based on two measures, precision and recall, which are derived from four values, namely, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) .

Experimental results shows comparison between K-means and DBScan based on ontology and it is found that DBScan clustering algorithm using ontological weighting scheme produce better result than K-Means algorithm. The performance of ontology-based DBScan algorithm is better than the ontology-based K- Means algorithm

In this paper [7] of michael steinbach , goerge karypis,vipin kumar experimented some of the common document clustering techniques.
In this paper two approaches has been

compared Hierarchical clustering and K- Means.Hierarchical clustering provides better quality clustering approach but is limited because of quadratic time complexity.On the other hand k-means is having time complexity which is linear in the number of documents.Also this technique produce inferior clusters.so results indicates that K- means technique is better than hierarchical approaches tested for a variety of cluster evaluation metrics.

Prof Anil khare [6] proposes that better clustering is achieved by discovering semantic structure of sentences in documents. The proposed model which consists of four components as below:

Text Preprocessing: This is done by separating the sentences, labeling the terms, removing stop words and then stemming the words

Concept Based Analysis: After preprocessing a concept based analysis is carried out under which different metrics like conceptual term frequency(ctf), Term Frequency (tf) and Document frequency (df).

Concept-based Document Similarity: Similarity between the documents is measured with the help of new concept based system.

Clustering techniques: Finally Clustering is achieved by different clustering techniques like – Single pass technique, HAC technique or KNN technique

The above concept based analysis is definitely better than traditional analysis however most of document categorization is based on Vector Space Model and the proposed system is based on Semantic structure modeling.

III.Problem Definition

Clustering is most useful task in data mining process for discovering groups ad identifying patterns in the given data.

The key element of clustering Is concept of similarity which is completely ignored by the conventional document clustering Method. The conventional document clustering methods rely on the classical vector space model using the keywords as the same feature. However these methods ignore the semantic relation among the keywords do not really address the special problems of document clustering which are as below:

- ☐ High dimensionality of the data
- ☐ High computation complexity
- ☐ Lack of semantic consideration

Huge volume, complex semantics and sparsity make the unstructured text document clustering process as the most difficult. The proposed paper eradicates the defects of conventional clustering method.

IV.Objectives and FutureScope

- ☐ To improve cluster quality
- ☐ To improve the performance of the document clustering with its domain knowledge
- ☐ To improve the accuracy and performance of the text document clustering process
- ☐ Concept based model retrieves more documents than the term based model
- ☐ One of the future directions is to use a concept based semantic similarity measure to cluster the documents.

V.Conclusion

The dimensionality of the data gets condensed in this proposed concept based clustering model. Concept based indexing technique with dynamic weight is used to effectively identify the leading concept of the document based on the back ground knowledge provided by the MeSH concept hierarchy. It is an efficient method for retrieving the documents even from very large databases. Concept based clustering exhibits a better performance than the traditional term based clustering.

REFERENCES

[1] Dr. Ahmed T. Sadiq, Sura Mahmood Abdullah,"Hybrid Intelligent Techniques for Text Categorization" (IJACSIT), Vol. 2, No. 2, pp. 23-40,2013.

[2] Hmway Hmway Tar , Thi Thi Soe Nyaunt, " Ontology-based Concept weighting for Text Documents", world Academy of Science, engineering and Technology, no.81, pp. 249-253, 2011.

[3] Rekha Baghel and Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Applications, vol.4, no.5, pp.6-12, 2010.

[4] A. A.Kogilavani, B. Dr.P.Balasubramanie, " Ontology Enhanced Clustering Based Summarization of Medical Documents", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.

[5] V Sureka et al," Approaches to Ontology Based Algorithms for Clustering Text Documents", ,Int.J.Computer Technology & Applications, Vol 3 (5), 1813-1817,2012.

[6] Shehata, Fakhri and Mohamed S.Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering", journal of IEEE Transactions on Knowledge and Data Engineering, Vol.22, pp. 1360-1371, 2010.

[7] Steinbach M, Karypis G and kumar V, " A comparison of document clustering techniques", KDD Workshop on text Mining'00, 2000.

[8] Bo-Yeong Kang, Sang-Jo Lee, " Document indexing: a concept-based approach to term weight estimation", Information Processing and Management, no.41, pp.1065-1080, 2005.

[9] Hmway Hmway Tar , Thi Thi Soe Nyaunt, " Enhancing Traditional Text Documents Clustering based on Ontology", International Journal of Computer Applications, vol.33, no.10, pp. 38-42, 2011.

[10] Shanfeng Zhu, Jia Zeng and Hiroshi Mamitsuka, " Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity", Data and text mining in Bioinformatics, vol. 25, no.15, pp.1944-195, 2009.

[11] Shehata, Fakhri and Mohamed S.Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering", journal of IEEE Transactions on Knowledge and Data Engineering, Vol.22, pp. 1360-1371, 2010.

[12] Steinbach M, Karypis G and kumar V, " A comparison of document clustering techniques", KDD Workshop on text Mining'00, 2000.

[13] Tahayna, B, Ayyasamy, R.K, Alhashmi, S, and Eu-Gene, S., "A Novel Weighting Scheme for Efficient Document Indexing and Classification",journal of IEEE International Conference on Information Technology.Vol.2, pp. 783-788, 2010.

[14] Xiaodan Zhang, Liping Jing, Xiaohua Hu, Michael Ng and Jiali Xia," Medical Document Clustering Using Ontology – Based Term Similarity Measures", International Journal of Data Warehousing and Mining, vol.4, no.1, pp. 62-73, 2008.