Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.1207 – 1210

RESEARCH ARTICLE

Document Clustering Using Concept Weight

Sapna Gupta[1], Prof. Vikrant Chole[2]

[1]Department of CSE, GHRAET, RTM Nagpur University

[2]Department of CSE, GHRAET, RTM Nagpur University

[1] sapna.gupta598@gmail.com, [2] vikrantchole@gmail.com

Abstract— Traditional techniques for clustering the document are mostly based on the number of occurrences and the existence of keywords. Similarly Phrase based clustering technique ignores the semantics behind the words, only captures the order in which the words appear in a sentence. The term frequency based clustering techniques takes the documents as bag-of words while ignoring the semantic relationship between the words. Considering the drawbacks of such system this paper proposes a concept based clustering technique. The ideology behind this concept is, it 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. K-medoid algorithm is used for clustering documents on Semantic through which the results are analyzed.

Keywords: Document Clustering; semantic similarity; Ontology; Concept weight

INTRODUCTION

With the increase in popularity of the internet, the accessibility of data is growing day by day. The data is unstructured, so there is a need to manage this large amount of data according to user query. Huge volume, complex semantics, high dimensionality, 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 concept based clustering is used to resolve this issue. 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.

Document clustering is automatic organization of document, fast information retrieval and, topic extraction or filtering. It is related to data clustering. Document clustering involves the use of descriptors and descriptor extraction. Descriptors are nothing but sets of words that describe the contents within the cluster. Document clustering is a centralized process. Document clustering includes web document clustering for search users. There are two types of document, online and offline. Online applications have efficiency problems when compared with offline applications.

Ontology is a way of specifying relationships among concepts, objects, and other entities belonging to a particular area of human experience or knowledge. A concept based indexing method consider keywords in the document, as well as the concepts based on the background domain knowledge. A document may hold multiple concepts in it. The concepts are extracted from the documents in concept based approach,. After extraction semantic weight is computed for effective indexing and clustering.

ONTOLOGY FOR CLUSTERING

MeSH [12] published by the National Library of Medicine mainly consists of the controlled vocabulary that provides a consistent way of retrieving information that may use different terminology for the same concept. MeSH helps to ensure that searches are comprehensive and allows for searching biomedical literature at varying levels of specificity. The controlled vocabulary consist of several different types of terms, such as subject headings also known as descriptors. short description , links to related descriptors, very similar terms descriptor and a list of synonyms. Descriptors and Entry terms are used in the proposed indexing method. Descriptor terms are the main concepts or main headings in the ontology. Entry terms are the synonyms or the related terms to descriptors.

CONCEPT BASED CLUSTERING

K-medoid is a classical partitioning technique of clustering that clusters the data set of *n* objects into *k* number of clusters [1, 3]. This k: the number of clusters required is to be given by user. This algorithm minimizes the sum of dissimilarities between each object and its corresponding reference point. It randomly chooses k objects in dataset D as initial representative objects called medoids. A medoid is defined as the object of a cluster, whose average dissimilarity is minimal to all the objects in the cluster i.e. it is a most centrally located point in the given data set. It then assigns each object to the nearest cluster depending upon the object's distance to the cluster medoid. After assigning data object to a particular cluster the new medoid is decided.

1) Input

k: the number of clusters.

D: a data set containing n objects.

2) Output

A set of k clusters.

3) Algorithm

1. Randomly choose k objects in D as the initial representative objects;
2. for all objects in the data set D

CONCEPT BASED INDEXING

The weight of the individual term *i* in document *j* is defined as in equation (1).

$$wji = tfji \times idfji = tfji \times log2 (n/dfji) \quad (1)$$

where *tfji* is the number of occurrences of term *i* in the document *j*, dfji is the term frequency in the collection of documents and *n* is the total number of documents in the collection. Concept based indexing improves the weight of the concept because it considers not only the concept word but also all the words that are associated to the concept word by means of the semantic relations.

EXPERIMENTAL RESULTS

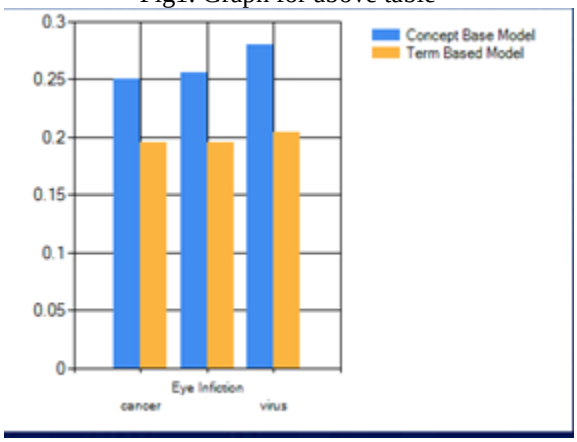
The experiments are conducted for the documents that are collected from MEDLINE based on three categories such as cancer, virus and eye infection. First of all three documents containing data from Medline is uploaded.

Stop word method is applied to remove unwanted or anonymous data from the document. The result of this is that we get formatted document. After that concept search method is applied to all three documents. Particular concept is searched in all the documents. It retrieves the frequency of searched concept from all the documents. Based on total, weight is calculated using above formula define in eqn (1) .Similarly term method is called, repeating same process as mentioned, again weight is calculated.

Table 1: Term Based Weight Vs Concept based Weight

Document Name	Method	Frequency	Weight
Cancer	Concept	302	0.250
Eye Infection	Concept	308	0.258
Virus	Concept	338	0.280
Cancer	Term	246	0.195
Eye Infection	Term	246	0.195
Virus	Term	256	0.203

Fig1: Graph for above table



The Graph shows the ratio or frequency of each method. It is clear from the graph that concept based technique retrieves more document than term based technique.

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 huge databases. Concept based clustering perform better than the traditional term based clustering. Concept bases technique retrieves more data than term based technique. One of the future directions is to use page ranking algorithm for ranking the pages.

REFERENCES

[1] 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 andMining, vol.4, no.1, pp. 62-73, 2008.
[2] Jiawei Han and Micheline Kamber,, “ *Data Mining Concepts and Techniques*”, Second Edition, Morgan Kaufmann Publishers, 2006.
[3] Dr. Ahmed T. Sadiq, Sura Mahmood Abdullah,”*Hybrid Intelligent Techniques for Text Categorization*” (IJACSIT),Vol. 2, No. 2, pp. 23-40,2013.
[4] V Sureka et al,“ *Approaches to Ontology Based Algorithms for Clustering Text Documents*”, ,Int.J.Computer Technology & Applications, Vol 3 (5), 1813-1817,2012.
[5] 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.
[6] Steinbach M, Karypis G and kumar V, ” *A comparison of document clustering techniques*”, KDD Workshop on text Mining’00, 2000.

[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] Steinbach M, Karypis G and kumar V, " *A comparison of document clustering techniques*", KDD Workshop on text Mining'00, 2000.
[12] 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.
[13] 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.
[14] 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.
[15] 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.

