www.ijecs.in

International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 8 Aug 2015, Page No. 13704-13709

Efficient Incremental Clustering of Documents based on Correlation

A.Devender, B.Srinivas, A.Ashok

M.Tech(SE), Department of CSE,
KITS Warangal,T.S, INDIA
devenderarecse@gmail.com

Asst Professor, Department of CSE,
KITS Warangal, T.S,INDIA
Srinu1032@gmail.com

MCA,Department of CSE,
Vagdevi Degree & PG clg Warangal,T.S,INDIA
Aare.ashok05@gmail.com

Abstract: With this project, a few dynamic file clustering algorithms, namely: Term consistency based Greatest Resemblance Doc Clustering (TMARDC), Correlated Concept primarily based MAximum Resemblance Document Clustering (CCMARDC) and Correlated Notion based Quickly Incremental Clustering Criteria (CCFICA) usually are proposed. From the aforementioned three suggested algorithms this TMARDC algorithm will be based upon term consistency, whereas, the CCMARDC and CCFICA are based on Correlated conditions (Terms and their Associated terms) notion extraction protocol..

Keywords: — Clustering, Document Analysis

1. Introduction

Tremendous growth inside the volume of text documents available through various sources such as Internet, electronic digital libraries, reports sources, and company-wide intranets has led to an increased desire for developing methods to help users to help effectively get around, summarize, along with organize data, with a good ultimate objective of helping the users to find what they are seeking. In this particular context, fast along with high-quality doc clustering algorithms play a significant role, because they have shown to provide each an user-friendly navigation/browsing process, by organizing large amounts of data into few meaningful groupings, as well as to greatly improve the retrieval effectiveness either by means of cluster-driven dimensionality lowering, term-weighting Tang et 's. (2005), or maybe by issue expansion Sammut along with Webb (2010). As today's google search does merely string matching, documents retrieved is probably not so tightly related to the user's issue. Thus, an excellent document clustering technique if available and implemented help you in coordinating the doc corpus automatically into a meaningful chaos hierarchy for efficient exploring and routing. Further, it will help to help overcome the inherent deficiencies regarding traditional data retrieval methods.

Document clustering has become investigated for use in a number of different regions of text exploration and data retrieval. In the beginning, document clustering had been investigated for improving the precision or maybe recall within information collection systems and since an efficient technique of finding the nearest neighbours of any document Vehicle Rijsbergen (1989 along with Kowalski along with Maybury 2002, Buckley along with Lewit 1985). Then clustering was used in browsing a collection of documents or maybe in organizing the outcomes returned by a search engine in reply to a user's issue Cutting et 's. (1992; Zamir et 's. 1997). Document clustering was already been used to help automatically generate hierarchical groupings of documents Steinbach et 's. (2000). By way of example, a web google search often returns thousands of pages in reply to a wide-ranging query so that it is difficult for users to help browse as well as to identify related information.

Clustering methods may be used to automatically class the reclaimed documents into a directory of meaningful classes, as is actually achieved by means of Enterprise Search engines like yahoo such since: Northern Light and Vivisimo Andrews along with Fox (2007). However, in this particular case scalability becomes a major issue as the number of documents raises day-by-day, thereby necessitating the necessity to cluster documents dynamically, without disturbing the formulated groupings. By clustering documents dynamically, the commitment taken for clustering is actually drastically decreased, as active algorithms processes the new document along with assigns it in to the meaningful groupings directly, rather than re-clustering the entire document inside the corpus. Though many document clustering methods exist for clustering documents in a very dynamic surroundings which provide terms Wang et 's. (2011) or maybe Synonyms along with Hypernyms Nadig et 's. (2008), there're not suitable for documents which have been technically connected. To defeat to earlier mentioned limitations, a design for active document clustering determined by Term consistency and Linked Terms (Terms along with their connected terms) since concepts within Scientific literary works and Newsgroups information set, is proposed within this paper.

I. Previous Work

The steady and amazing progress of computer hardware technology in the last few years has led to large supplies of powerful and affordable computers, data collection equipments, and storage media. This technology provides a great boost to the database and information industry and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis. So we can say that this technology provides a tremendous growth in the volume of the text documents available on the internet, digital libraries, news sources and company-wide intranets. With the increase in the number of electronic documents, it is hard to manually organize, analyze and present these documents efficiently. Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. [1]

Lexical chains have been proposed that are constructed from the occurrence of terms in a document. Problem to improve the clustering quality is addressed where the cluster size varies by a large scale. They have stated that variation of cluster size reduces the clustering accuracy for some of the state-of-the-art algorithms. An algorithm called frequent Itemset based Hierarchical clustering (FIHC) has been proposed, where frequent items i.e. minimum fraction of documents have used to reduce the high dimensionality and meaningful cluster description. However, it ignores the important relationship between words.

The benefits of partial disambiguation of words by their PoS is explored. They show how taking into account synonyms and hypernyms, disambiguated only by PoS tags, is not successful in improving clustering effectiveness because of the noise produced by all the incorrect senses extracted from WordNet. A possible solution is proposed which uses a word-by-word disambiguation in order to choose the correct sense of a word. CFWS has been proposed. It has been found that most of existing text clustering algorithms use the vector space model which treats documents as bags of words. Thus, word sequences in the documents are ignored while the meaning of natural language strongly depends on them.[1]

Most of document clustering algorithms use the vector space model (VSM) alone for document representation. VSM represents documents as vectors in the space of terms and uses the cosine similarity between document vectors to estimate their similarity. VSM, however, ignores any semantic relations between terms. Sometimes by matching only query terms to document, it is not possible to retrieve relevant documents. This motivates the proposed work to prepare document clusters based on nearest neighbors cluster similarity so that it will not only retrieve the documents which contain query terms but also retrieve those documents which are similar to retrieved document. First the n nearest neighbors of all points are found. If two data points are similar enough, they are considered as neighbors of each other. Every data point can have a set of neighbors in the data set for a certain similarity threshold. The documents are ranked only on the basis of the term frequency. This motivates the proposed model to give importance to the information content of the document under consideration. Thesaurus or Ontology as background Knowledge has been applied to various text mining problems but very few attempts have been made to utilize it for document clustering. Here, in this paper, we utilize the online encyclopedia Conservapedia, to retrieve the synonyms of the query term so that from the retrieved documents of the dataset the correlated semantic terms of the specified query term are identified and finally more similar documents are ranked based on semantic correlation similarity. This improves the accuracy of the retrieved relevant documents without much increasing time.

Some work has been proposed on using Latent Semantic Indexing (LSI) for document clustering. However, most of dimension reduction techniques are computationally very expensive for large data sets, and they suffer from the problem of determining the intrinsic dimensionality of the data. Other models for document representation are based on analyzing the semantics of terms using a lexical database, such as WordNet. Hotho et al for instance, proposed an approach in which terms are augmented or replaced by vectors based on related concepts from WordNet. Document clustering algorithms are then applied to these concept vectors. Recent work on document clustering constructs concept vectors using Wikipedia.

These methods, however, are computationally complex and produce high-dimensional concept vectors. Some related work for document clustering is based on explicitly grouping related terms first, and then grouping documents into clusters based on term clusters. Simultaneous clustering of terms and documents is a related approach which is based on spectral partitioning of a bipartite graph of documents and terms. These methods, however, do not scale well to handle large data sets. Numerous documents clustering algorithms appear in the literature. The two most common techniques used for clustering documents are Hierarchical and Partitional (K-means) clustering techniques. [2]

Most existing text clustering approaches rely on the bag-of-words representation. Using words as features, each document is represented in a high dimensional vocabulary space as a vector of (normalized) word frequency counts. The sparsity (most documents contain less than 5% of the vocabulary terms) and the noise (text data extracted from internet pages, chat logs or e-mails may often contain spelling errors and abbreviations) in this representation indeed affect the final clustering performance.

These difficulties have motivated the development of dimensionality reduction techniques as a pre-processing step to determine a more compact and relevant document representation. Examples of such approaches are the singular value decomposition used in the Latent Semantic Indexing (LSI) or other matrix factorization approaches like random projections or non-negative matrix factorization.

Matrix factorization approaches have successfully been applied to the clustering of text data, including web access log pages. Other approaches for dimensionality reduction of text data include probabilistic models and co-clustering approaches. The former include the popular Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation. Those two models have successfully been used for the task of topic discovery. Co-clustering approaches aim at simultaneously clustering documents and words.

Recent advances include formulations in the bipartite graph framework and in the matrix factorization framework. Non informative words can also be removed by simple heuristics based for example, on their document frequencies. These unsupervised approaches though less efficient than supervised feature selection methods allow to find less noisy representation space than the initial bag-of-words space. [4]

The proposed mining model is an extension of the work. The proposed 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 Figure above. A raw text document is the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on the PropBank notations. After running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence.

The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document, and corpus levels.

In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled term either word or phrase is considered as concept. [5]

II. Proposed System

A. *Preprocessing*

In this module the preprocessing involves; tokenization, removing stop words and stemming. Tokenization, is the process of splitting the sentences into separate tokens. For example, this is a paper about document clustering is split as: this\is\paper\about\document\clustering. Stop words are frequently occurring words that have little or no discriminating power, such as: \a", \about", \all", etc., or other domain-dependent words. Stop words are often removed. Stemming is the process of removing the affixes in the words and producing the root word known as the stem. Typically; the stemming process is performed to transform the words into their root form. For example: connected, connecting and connection would be transformed into ‘connect’.

B. *Static document clustering*

In this module document vectors are computed and the processed documents are clustered with the use of cosine similarity, using a K-means clustering algorithm in order to group similar documents. Cluster analysis or clustering is the assignment of a set of observations into subsets called clusters so that observations of the same cluster are similar in some sense. The K-means method will split a large cluster into sub-clusters and this step will be repeated for several times, until the K numbers of clusters are formed with high similarity.

C. *Term frequency based maximum resemblance document clustering*

This algorithm adopts the core concept of MARDL i.e. Maximum Resemblance technique. This algorithm is purely based on a bag of words representation. This dynamic algorithm starts with the set of clusters which is obtained as the result of K-Means clustering. Initially, the sample set is constructed for each cluster set. One third of the documents are chosen randomly as samples from the set of documents in each cluster. The samples chosen should be unique and should not be replica’s of documents in samples. The new documents are preprocessed first which includes stop word removal process and stemming process. The new documents are stemmed using

DOI: 10.18535/ijecs/v4i8.15

a stemming algorithm. After preprocessing of the new document, the new document is compared with samples based on Sentence Importance computation (SIC), Cluster set Importance computation (CIC) and the influence of the new document in each cluster termed as Frequency Value (FV) is calculated. The CIC should be normalized to obtain the FV, because the number of documents in each sample may vary. Then the dynamic algorithm assigns the new document to the cluster with the high FV, provided, the FV is within the threshold value. The threshold value is maintained for clustering process to make a document to form a new cluster or assigning a document to the appropriate cluster. If all the clusters result in FV less than the threshold value, then, the new document forms a separate cluster. The threshold value is calculated through a series of experiments on all worst, average and best case inputs and it is termed as Threshold value (Tmax). A newly arrived document, if it's FV falls less than the Tmax it forms a separate cluster, thus ensuring that no document goes without clustering, even it doesn't patches with any of the existing clusters.

D. *Correlated concept based maximum resemblance document clustering*

Incorporation of semantic features, improve the quality of document clustering and also the accuracy of information extraction techniques. In this study, concept extraction algorithm, which itself is a modification of the existing semantic-based model proposed has been adopted. The model aims to cluster documents by meaning. The semantic-based similarity measure is used for the two CCMARDC and CCFICA algorithms. In order to extract concepts, a domain-specific dictionary consisting of scientific terms and terms related to newsgroup tracks are created, where in Word Net lexical database Miller was used for Synonyms/Hypernyms extraction. Domain-specific dictionary for scientific and Newsgroups are used for concept extraction, as it eliminates the need for word sense disambiguation (WSD). Considering the extraction of Synonyms/Hypernyms as concepts degrades the efficiency of the results in the case of scientific literature and news group dataset because of the fact that the documents speak more about scientific or technical terms. Concept extraction is based on Correlated concepts are nothing but the terms and their related terms. For Concept extraction, domain specific dictionary is used where terms related to each domain is kept along with the meaning of the term.

E. *Correlated concept based fast incremental clustering algorithm*

In this module Fast Incremental Clustering Algorithm (FICA) an increment data clustering algorithm for mushroom data set. The main objective of this algorithm is to cluster the categorical data into the K number of clusters using incremental method. The existing algorithm uses dissimilarity measure for finding the distance between the new object and the existing cluster. The core idea of the above algorithm is considered in the CCFICA proposed here. The FICA algorithm is modified for clustering the documents for dynamic document corpuses, based on semantic similarity. For every cluster, the top correlated concepts from each document are extracted and are maintained as a concept pool. Instead of computing the dissimilarity between document clusters and the new document, the semantic similarity between the new document and the concept pool is computed, which reduces the computation overhead.

III. Results

The concept of this paper is implemented and different results are shown below, The proposed paper is implemented in Java technology on a Pentium-IV PC with minimum 20 GB hard-disk and 1GB RAM. The propose paper's concepts shows efficient results and has been efficiently tested on different Datasets.

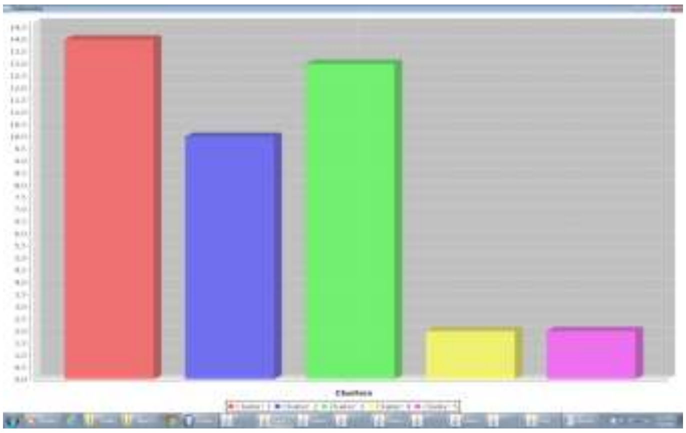

Fig. 1 Graph for Static Clustering

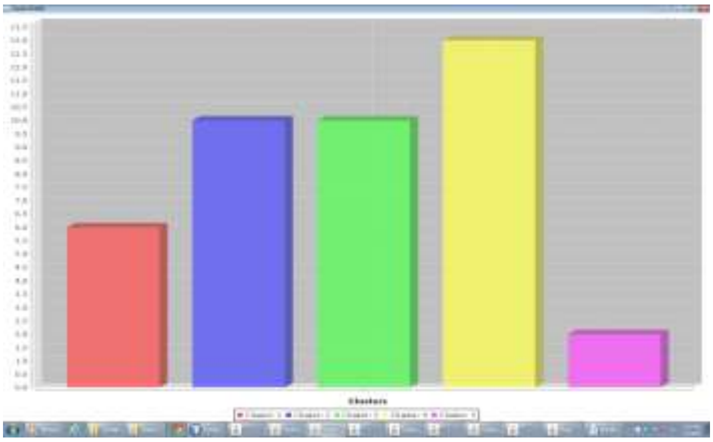

Fig. 1 Graph for TMARD Clustering

DOI: 10.18535/ijecs/v4i8.15

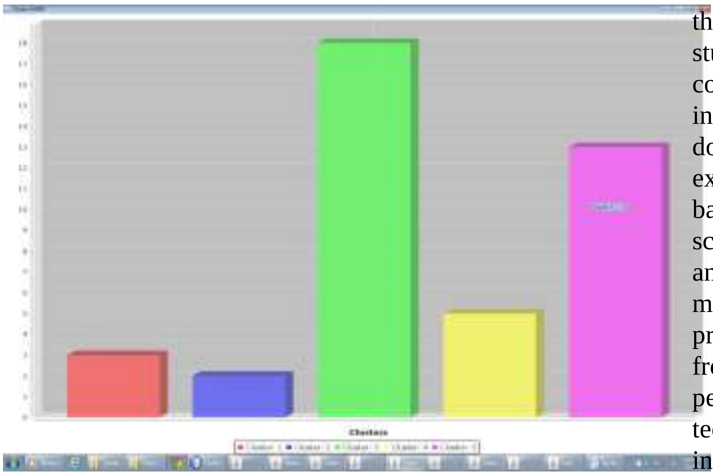

Fig. 3 Graph for CMARD Clustering

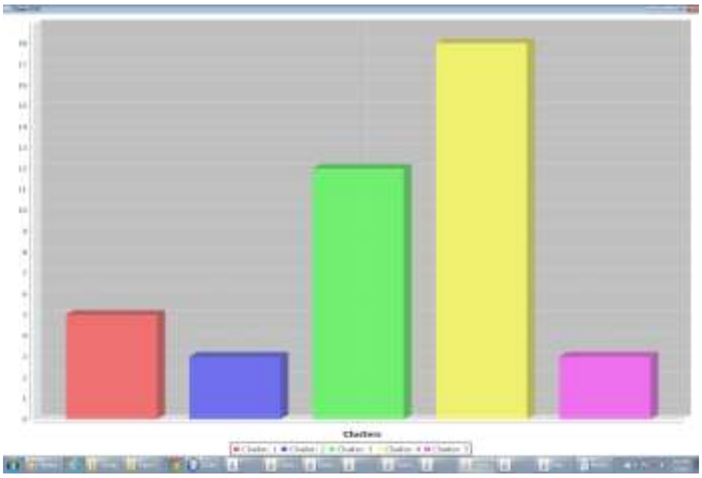

Fig. 4 Graph for CCFIC Clustering

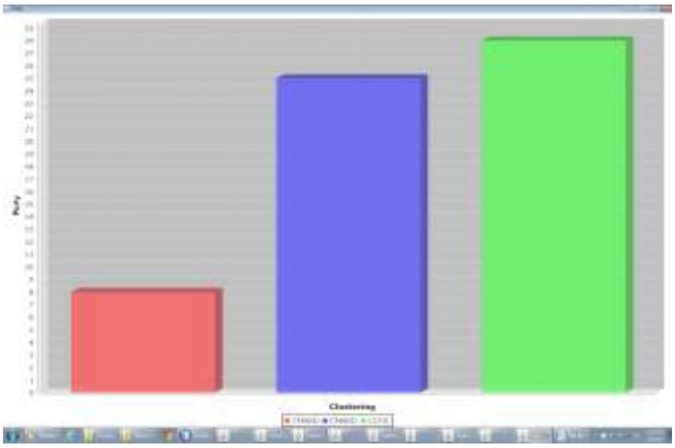

Fig. 4 Graph for Accuracy Evaluation

IV. Conclusions

The emphasis in the present work is Active Document Clustering based on Term rate of recurrence and Related based Principle algorithms, making use of semantic-based similarity measure. The core concept of Data mining algorithms MARDL and also FICA can be adopted with the proposed algorithms TMARDC, CCMARDC and also CCFICA. Generally speaking the docs are manifested as TF-IDF, while, in this particular study the documents tend to be represented by way of correlated expression vector (crtv). This manifestation helps an individual to record the technological correlation between your documents. The suggested algorithms are compared with the existing term rate of recurrence and synonyms/ hypernyms based incremental document clustering algorithms thinking of scientific novels and newsgroup dataset. On the comparative analysis it could concluded that will considering crtv manifestation for energetic document clustering results in promising results for scientific novels. Sometimes the outcomes from the Newsgroup dataset aren't promising, due to the need pertaining to relatively a lot more English literary terms, quite technical words. In foreseeable future, it can be proposed to increase concept extraction based on significant terms in docs, and also by including semantic associations like hyponymy, holonymy, and also meronymy.

References

[1] Baghel, R, & Dhir, R. (2010). A frequent concept based document clustering algorithm. International Journal of computer Applications, 4(5), 0975–8887.

[2] Bharathi, G, & Vengatesan, D. (2012). Improving information retrieval using document clusters and semantic synonym extraction. Journal of Theoretical and Applied Information Technology, 36(2), 167–173.

[3] Danushka, B, Yutaka, M, & Ishizuka, M. (2011). A Web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge And Data Engineering, 23(7), 977–990.

[4] Pessiot, JF, Kim, YM, Amini, MR, & Gallinari, P. (2010). Improving document clustering in a learned concept space. Journal of Information Processing and Management, Elsevier, 26, 182–192.

[5] Shehata, S, Fakhri, K, & Mohamed S, S. (2010). An efficient concept-based mining model for enhancing text clustering. IEEE Transactions On Knowledge And Data Engineering, 22(10), 1360–137.

[6] Yan, J, Liu, N, Yan, S, Yang, Q, Fan, WP, Wei, W, & Chen, Z. (2011). Trace-oriented feature analysis for large-scale text data dimension reduction. IEEE Transactions on Knowledge and Data Engineering, 23(7), 1103–1117.

[7] Zhang, T, Member, YY, Tang, BF, & Xiang, Y. (2011). Document clustering in correlation similarity measure space. IEEE Transactions on Knowledge And Data Engineering, 24(6), 1002–1013.

Author Profile

A . DEVENDER : is currently pursuing his M.Tech Computer Science & Engineering Department in Kakatiya Institute of Technology and Science,

B. Srinivas is currently working as Asst. Professor at Kakatiya Institute of Technology & Science, Warangal, A.P, INDIA. He has completed his M.Tech from JNTU, pursuing Ph.D from K.L University main research area in information retrieval ,Data Mining, text Mining, Algorithm Analysis and Design, . He has been involved in the organization of a number of conferences and workshops.His area of interest include Big Data.

A.Ashok : is currently working as Asst.professor at vagdevi degree and PG college, Warangal.He has completed his MCA Computer Science in Vagdevi Degree and PG College, Warangal. his area of interests includes Data mining and operating system.

Warangal. he received his B.Tech in Computer Science and Engineering from Warangal institute of technology and science, oorugonda, warangal. Her area of interests includes Data mining and operating system.