DYNAMIC CLUSTERING OF DOCUMENTS

Nilam A. Sinari[1], Rachel Dhanaraj[2]

[1]*Computer Engineering Department, Goa College of Engineering, Farmagudi-Ponda, Goa, India*
[2]*Computer Engineering Department, Goa College of Engineering, Farmagudi-Ponda, Goa, India*

Abstract

Due to increasing number of documents in the World Wide Web, which requires faster processing and efficient methods to handle them retrieval has become a major issue. Clustering process is used to retrieve the document and automatically group them into clusters. Here the issue is scalability. Therefore to handle the scalability problems documents are clustered dynamically when they add up, without the need to re-cluster the formulated clusters which are clustered by static document clustering. The proposed method will dynamically process the newly arrived document and assign it to the meaningful cluster without the need to disturb the formed clusters. In order to exploit semantic relation between terms within the document WordNet ontology is been used to add hypernyms.

Keywords— *Document Clustering; WordNet; Static document clustering; Dynamic Document clustering; Semantic Similarity*

----------***----------

1. INTRODUCTION

Most of the information available on the World Wide Web is stored in text databases, which consist of documents from various sources, such as digital libraries, news articles, research papers, books, e-mail messages and web pages. Text documents are increasing day by day increasing the amount of information available in electronic and digitized form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web. The information is overloaded and at the same time more and more people use the World Wide Web to retrieve their queries. This huge information present makes retrieval a tedious process for users. Some methods have been developed to make the task of user quick and retrieve data much faster and in efficient way. The problem is basically not to find the text document but to find relevant document. Document clustering plays an important role to achieve this objective. Clustering methods are used to automatically group the retrieved document into meaning clusters. Most of the current methods for text clustering are based on the similarity between the text sources. The similarity measures work on the syntactically relationships between these sources and neglect the semantic information in them. By using the vector-space model in which each document is represented as a vector or 'bag of words', i.e., by the words (terms) it contains and their weights regardless of their order. Vector space model is a popular model for document representation however, semantic relations between terms are not taken into account. Two terms with a close semantic relation and two other terms with no semantic relation are both treated in the same way. This unconcern about semantics could reduce quality of the clustering result.

So most of the research work is based on clustering with semantic relation. However due to huge data adding up everyday scalability becomes a big issue. In this paper a novel approach where documents are clustered dynamically without the need to disturb the formulated clusters is been presented. Clustering it in this way saves the time and efforts taken and processes the new document and assigns them to appropriate clusters instead of re-clustering the entire document set. Incorporating semantic feature from WordNet helps in increasing clustering efficiency but at the same time leads to Word Sense Disambiguation problem. Computer program cannot automatically select which sense to choose. If the Word Sense Disambiguation problem is not taken into account clustering accuracy will decrease and will lead to inefficient clusters. So Word Sense Disambiguation needs to be taken care of if efficient clustering is needed.

2. LITERATURE REVIEW

Vector Space Model (VSM) is the widely used data representation for clustering and is being used by existing document clustering algorithms. The terms in VSM is represented as a feature vector of terms in document. In [1] the author presented a survey which states that most of the clustering algorithms are based on concept based on TF-IDF method. The issues of these methods are:

- The semantic importance of each term in a document is not considered;
- It does not consider the semantic of the word and assigns weight.
- Synonyms, polysemous, hypernyms are not considered

In order to improve the clustering results WordNet is used in most of the clustering techniques. Semantic features are incorporated from the WordNet lexical database to improve the accuracy of clustering results. When synonyms and hypernyms are added and disambiguation of the word is done by their PoS and WordNet concepts may not be successful in improving clustering efficiency because due to the addition of extra terms may let to introduction of noise which may be

generated by incorrect senses extracted from WordNet. Since the synset may contain may hypernym and synonyms addition of all of them may increase noise as in [2]. In [3] the author presented a brief survey on various clustering techniques based on semantics. In [4] the authors proposed a semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network. In the proposed work the author generates documents vectors using the lexical category mapping of WordNet after preprocessing the input documents. They applied three different clustering algorithms, SOM neural network, k-means, and bisecting k-means to the generated documents vectors. The results showed that SOM neural network achieves higher clustering quality than other two clustering algorithms k-means, and bisecting k-means.

As per the study it shows that most of the clustering techniques are based on term frequencies. Some clustering techniques use concepts (synonyms and hypernyms) from WordNet. There is not much research based on dynamically clustering the documents but few once are discussed below.

2.1 Phrase Semantic Similarity Histogram (PSSM)

In [5] the author has proposed Phrase Semantic Similarity Histogram (PSSM) model which is an incremental algorithm. Incremental clustering process is carried out by integrating text semantics. The semantic similarity within each cluster is measured by using semantic histogram representation. This method is based on single word as well as phrase analysis, assigning and adjusting the term weight based on similar terms that occur in a document. The semantic histogram ratio is calculated as soon as the new document is added. The insertion order problem is taken care by eliminating inefficient documents that reduce the cluster efficiency, and reassign them to a more appropriate cluster.

2.2 Concept-Based Mining Model (CBM)

The proposed Concept- based Mining Model is used to Enhance Text Clustering. This model consists of sentence-document and concept based corpus analysis and concept-based similarity measure. This model combines the factors affecting the weight allowing concept matching and similarity calculations of documents in an accurate way [6]

2.3 An Incremental Algorithm for Clustering Search Results (ICA)

In [7] the author has proposed an incremental clustering algorithm. This algorithm was based on Cluster Average Similarity Area (CASA). It is used to score the degree of coherency of a cluster. CASA was used to compute cohesiveness quality information of a cluster. It incrementally assigns data object to respective clusters one at a time.

3. PRE-REQUISITES

3.1 TF-IDF

"Term Frequency, Inverse Document Frequency" is important step in frequency weighing. There are many words which will be present in the document now considering all the words for clustering may lead to increase in dimensionality and also noise. So in order to find important words in a document TF-IDF method is used. This method works on how frequently a word appears across multiple documents. Term frequency (TF) gives you whether a word is important within a document giving you the count. The count is found by finding how many times that particular word appears in the document. Inverse document frequency (IDF) gives count of how many document contains that word.

Term Frequency, Inverse Document Frequency is given by

$$tfidf_{ij}=0.5+0.5*\frac{f_{ij}}{\max_{t_j\in d_i}(f_{ij})}\times log\left(1+\frac{|D|}{|\{d_i|t_j\in d_i, d_i\in D\}|}\right)$$

$tfidf_{ij}$ = term frequency inverse document frequency

f_{ij} = frequency of term t_j in d_i

$\max_{t_j\in d_i}(f_{ij})$ = maximum frequency of all terms in d_i

D = Total number of documents in the corpus

3.2 WordNet

WordNet has been widely used to improve the performance of information retrieval (IR) systems. The research problem which is faced by information retrieval system is handled by using WordNet. In term based representation semantic of the term is not taken into account so to incorporate semantic i.e. meaning of the term an ontology-based representation has been proposed which exploits the hierarchical is-a relation among concepts, i.e., the meanings of words. For example, to describe a term-based representation with documents containing three words: “fruit”, “apple”, and “mango” a vector of three elements is needed; with an ontology-based representation, since “fruit” subsumes both “apple” and “mango”, it is possible to use a vector with only two elements, related to the “apple” and “mango” concepts, that can also implicitly contain the information given by the presence of the “fruit” concept. A set of independent concept that covers the whole ontology by defining ontology base allows use of fixed sized document vector. [8]

3.3 Relations in WordNet

The most widely used relation among all the relations in synsets is the super-subordinate relation (ISA relation). It links general synsets like {car} to increasingly specific ones like {vehicle}. Thus, WordNet states that the category car includes vehicle and conversely, concepts like vehicle make up the category car. They contain hierarchies which go to

root node ultimately. Hyponymy relation is transitive. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules, it does not include prepositions, determiners etc. [8]

4. PROPOSED MODEL

In the proposed model we first begin with preprocessing the documents. Fig 1 shows the flow of the proposed model.

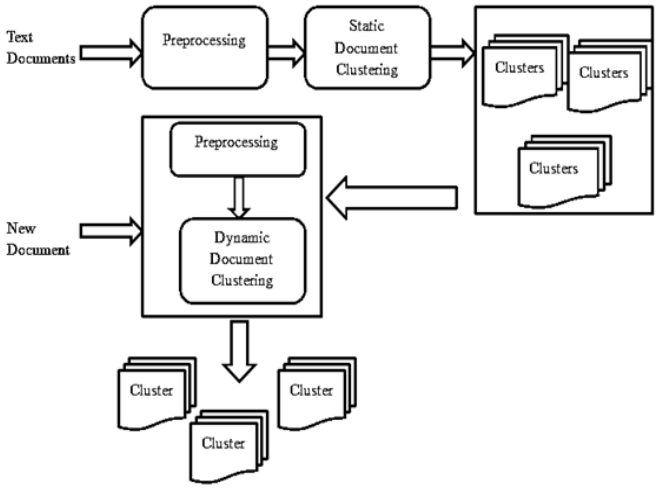

Fig 1. Proposed Model

4.1 Preprocessing stage

4.1.1 Generation of Key Term Set

There are two stages in this module, namely Key Term Extraction and Key Term Selection, for dimensionality reduction of the document set. Fig 2 shows the preprocessing stage.

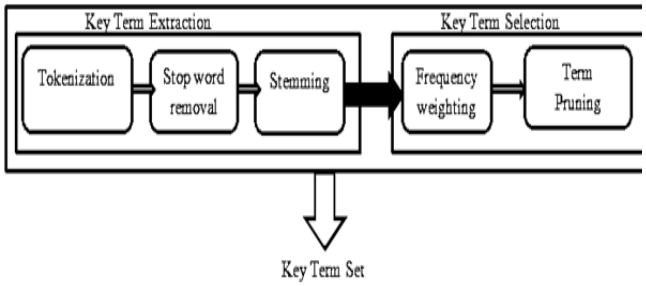

Fig.2. Preprocessing

4.1.1.1 Key Term Extraction

This step basically consists of three stages. Tokenization, stop word removal and stemming. In Tokenization we simple split the sentences into tokens substituting a sensitive data element with a non-sensitive equivalent. Stop words are the words which are occurring frequently but individually do not contribute any meaning so it is better to remove them. Predefined stop word list is maintained and the terms that appear in it are removed retaining the other terms. After the documents are processed through stop word list, stemming is performed with the words getting converted to their base form. Porter stemming algorithm is used as in [9]

4.1.1.2 Key Term Selection

Usually the terms with low frequencies are useless for identifying appropriate cluster and are treated as noise so it is appropriate to remove this noise from the document set. For frequency weighing the tf–idf (term frequency × inverse document frequency) method is applied which choose the key terms for the document set. A fixed tf–idf threshold γ will be maintained and the terms whose weight is less than the threshold it will be discarded. The weight of each term in each document is calculated and those which satisfy the pre-specified minimum tf–idf Threshold is retained and the rest terms are pruned off. These retained terms form a set of key terms for the document set D. The key term set of a document is the set containing only meaningful terms which are not appearing in the stop word list and satisfy pre-specified minimum threshold.

4.1.2 WordNet Based Processing

The basic objective of this module is to enrich the document set by using WordNet. In this study we basically use the hypernyms provided by the WordNet whose relationship of relevant terms has been pre-defined in WordNet as a useful feature for document clustering. Once the key terms are extracted from the document they are looked up into WordNet and organized based on the hierarchical (IS–A) relationship of WordNet [8] to construct term trees. Construction of the term tree is done by matching a key term in WordNet and checking its hypernyms up to five levels. A term tree [10] of term t, is a 4-tuple denoted by F = (W, H, I, t) where W represents the WordNet and H links the set of hypernyms up to five levels in W. A set of hypernyms I= $\{h_1,\ldots, h_r\}$ of a key term $t_j \in W$, together with their reference function $H:2^{|W|}\rightarrow 2^{I}$ in W. Hypernym is usually denoted by $h_1 \leq h_2$ where h_2 is the hypernym of h_1 defined in WordNet. A term forest of a set of key terms $\{t_1,t_2,\ldots,t_i, \ldots,t_m\}$, denoted by $F=\{F_1, F_2,\ldots, F_m\}$, is a set of term trees, where m is the total number of key terms in D. Fig 3 shows WordNet based processing.

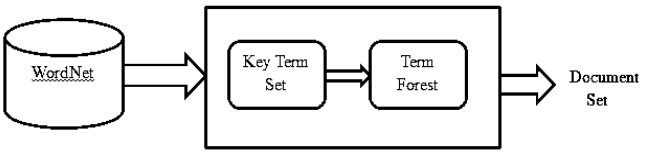

Fig.3. WordNet Based Processing

By using hypernyms from WordNet magnifies hidden similarities to identify related topics which lead to better clustering results. Hence to find semantically-related documents we enriched the representation of each document with hypernyms based on WordNet. When using WordNet to enrich the semantic relation the main problem is which sense (meaning) to choose from WordNet. For Example “bank” can be a financial institute as well as a sloping land as a human being we very well know which sense to choose for a particular context based on the surrounding content but

computer program can't handle it directly this problem is called as Word Sense Disambiguation. Which is this present research work is handled with the help of adapted Lesk algorithm [11].

4.2 Static Document Clustering

The document set will be converted to document vector which will be given to bisecting k-mean clustering algorithm which will lead to the formation of clusters. Bisecting K-mean works as follows the algorithm divides the dataset recursively into clusters. The k-means algorithm is used by setting k to two in order to divide the dataset into two subsets. Then the two subsets are divided again into two subsets by setting k to two. The recursion terminates when the dataset is divided into single data points or a stop criterion is reached. Fig 4 shows static document clustering.

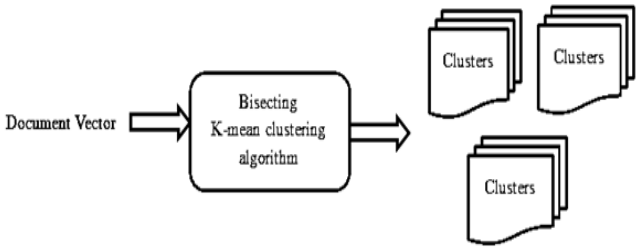

Fig.1. Static Document Clustering

4.3 Dynamic Document Clustering

In this module the newly arrived document is inserted to the appropriate existing cluster without the need to disturb the formulated cluster. When the new document is available it is preprocessed right from tokenization, stop word removal, stemming and WordNet processing and then it is dynamically clustered. Following issues should be handled

- Effectiveness: when the new documents arrive how accurately they are inserted to the existing clusters.
- Order of Insertion: care should be taken on how the new document arrives. The arrival of new documents should not affect the clustering results.

Recursively each new document are added to the existing cluster. Dynamically at run time the new clusters are added and no need to re-cluster them back. The result of which is that updating of existing clusters takes place and final clusters are obtained.

Algorithm
Input Description

- Input the formed k clusters which are formulated by bisecting k-mean algorithm C = $\{C_1, C_2 \ldots C_k\}$
- Let S=$\{S_1, S_2 \ldots S_K\}$ where S_i is a sample of cluster C_i
- S_1=$\{ d_1, d_2 \ldots d_m\}$ set of document in sample set S_i

//choose samples around cluster centroid – (1/3 in size)

- Let D = $\{ doc_1, doc_2 \ldots doc_p \ldots doc_q\}$ //doc_p is the newly arrived document

For all documents in D
1. Begin

- Preprocess the newly arrived document
  a. Perform tokenization, stop words removal, stemming and select equivalent or hypernyms terms from WordNet
- Let the DV (document vector) contains terms (hypernyms/concepts)

2. For every sample set S_i begin
3. For each document in sample set
For new document and document in sample set semantic similarity [12] is computed
4. As the number of document in each sample varies Normalize the semantic similarity so as to calculate maximum resemblance given by similarity measure by number of document in each sample
5. Repeat the above computation for all samples and compute normalized concept importance ratio
6. Check for the normalized concept importance ratio i.e. which one is maximum and if it is above the threshold.
7. Include the newly arrived document to the appropriate cluster where the normalized concept importance ratio is maximum.
8. If the normalized concept importance ratio is less than the threshold forms a new cluster.

5. CONCLUSION

A different approach to document clustering is proposed which is motivated based on the idea that most of the clustering algorithms have to re-cluster the documents when new documents arrives which results in computational cost and scalability is also an issue. So by dynamically clustering the document at run time we are looking to improve the clustering efficiency hence no need to re-cluster the documents which are already clustered. Also hypernyms are added to the document up to five levels. A term forest is constructed to revel semantic of a term. Addition of hypernyms and checking for Word Sense Disambiguation would improve the clustering efficiency.

REFERENCES

[1] Prathima, Y, & Supreethi, KP."A survey paper on concept based text clustering" .International Journal of Research in IT & Management, 1(3), 45–60 (2011).

[2] L. Jing, M.K. Ng and J.Z.Huang, "Knowledge-based vector space model for text clustering", Knowledge and Information Systems, 25 (1), pp. 35-55 (2010).

[3] Shah, N & Mahajan, S. ."Semantic based Document Clustering: A Detailed Review". International Journal of Computer Applications (0975 – 8887). Vol. 52–No.5 (2012).

[4] Gharib, T., Fouad, M., Mashat, A. & Bidawi, I. "Self Organizing Map -based Document Clustering Using WordNet Ontologies", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2 (2012).

[5] Gad WK & Kamel, "Incremental Clustering algorithm based on phase- semantic similarity histogram". Proceedings of Ninth International Conference on

Machine Learning and Cybernetics, 11(14), 2088-2093(2010).

[6] Shehata S." An efficient concept based mining model for enhancing text clustering" Journal of Knowedge and Data Engineering 22(10)(2010)

[7] Li, Y, Chung, SM, & Holt, JD."Text document clustering based on frequent word meaning sequences". Journal on Data & Knowledge Engineering, 64(1), 381–404 (2008).

[8] G.A. Miller, "WordNet: a lexical database for English", Communications of the ACM 38 (11) (1995) 39–41.

[9] Porter, MF. "An algorithm for suffix stripping program".14(3), 130–137 (1998).

[10] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang."An integration of WordNet and fuzzy association rule mining for multi-label document clustering". Journal on Elsevier Data & Knowledge Engineering 69 (2010) 1208–1226 (2010).

[11] Satanjeev Banerjee & Ted Pedersen "An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet" Springer-Verlag Berlin Heidelberg 2002.

[12] Shehata, S. "A WordNet-based Semantic Model for Enhancing Text Clustering". IEEE International Conference on Data Mining Workshops(pp. 477–482). 6 Dec. 2009.