GRAPH BASED RATIONAL TEXT CLUSTERING USING SEMANTIC ONTOLOGY

[1]S. SUBBAIAH, [2]C. CHANDRASEKAR

[1]Asst. Prof., Department of Master of Computer Applications, K. S. Rangasamy College of Technology, Tiruchengode, Namakkal – 637215, TamilNadu, India.
[2]Assoc. Prof., Department of Computer Science, Periyar University, Salem, TamilNadu, India
E-mail: [1]subbaiah.phd123@gmail.com, [2]ccsekar@gmail.com

ABSTRACT

Text clustering methods have been discussed with various factors. We have the goal and problem of making decision with rationality factors about how much information is available or how complete the document about the knowledge. We propose a new rational text clustering algorithm using the semantic ontology. The documents are processed to extract the key terms as feature vectors. Semantic frequency SF and inverse semantic frequency ISF are computed for each document. Using computed SF and ISF we compute semantic weight for each document towards various categories, based on which the document is identified to a class or category. The proposed method produces more accurate clusters with reduced overlap.

Key Words: *Semantic Ontology, Clustering, Text Mining.*

1. INTRODUCTION:

Text clustering is a common topic of research of past 40 years, where the application of text mining is information technology and data mining. The application of text clustering has impact in various domains and application. Whenever the volume of document increases, the process of mining or extracting information becomes complicated. In order to extract or mine exact document or knowledge is completely depend on the kind of clustering methodology used.

Clustering is a way of organizing or indexing the text document, so that the document can be retrieved easier at later stage. The text document contains several paragraphs, each paragraph may contain many statements and a statement is combination of terms. For clustering purpose each documents has to be analyzed and identified about the concept what it is talking about. For example if we give 100 documents, each documents topic has to be identified, so that we can index the document into a cluster. Some of the document may discuss about data mining and few of them may speak about image processing and so on. The problem here is what we do when the volume of documents increases; this is where we start thinking about the clustering algorithms which works on computerized manner.

In any text document it is not necessary that it has to contain exact words from the semantic ontology. While clustering in a term based systems, the terms with relationship are neglected, which reduces the clustering efficiency. All the clustering methods proposed earlier misses the usage of logical meaning of the text or sentence present in the corpus. The semantic meaning of the sentences present in the document has to be used for clustering purpose. So that we have decided to use semantic ontology for the purpose of clustering.

Ontologies can play a crucial role in enabling knowledge processing, sharing, and reuse between applications. Generally defined as shared formal conceptualizations of particular domains, ontologies provide a common understanding of topics that can be communicated between people and application systems. Ontologies are used in e-commerce to enable machine-based communication between buyers and sellers; vertical integration of markets and description reuse between different marketplaces. Search engines also use ontologies to find pages with words that are syntactically different but semantically similar.

Ontology typically contains a hierarchy of concepts within a domain and describes each concept's crucial properties through an attribute-value mechanism. Further relations between concepts might be described through additional logical sentences. Finally, constants (such as "January") are assigned to one or more concepts (such as "Month") in order to assign them their proper type.

Since ontologies can represent the domain knowledge, the semantic relations and hierarchies of the classes in the ontologies may also be seem as knowledge or rules of the domain. In order to take full advantage of ontologies, the semantic relations and hierarchies of the classes in the ontologies should be considered in ontology building.

Here we propose a new clustering algorithm using semantic ontology. We use the relation and the class labels from the semantic ontology to compute various metrics, which represent the relation of the document to a particular category. Using the computed values the document will be assigned to a cluster.

2. BACKGROUND:

General text clustering can be classified as text based, link based and hybrid. The text based method depends on the content of the document where as the link based method are depends on the link structure of the document and the hybrid method uses both present in the document.

The implementation of text clustering has various methodologies as follows.

- Partitional Methods
- Graph Based Methods
- Hierarchical Methods
- Neural Network Based Methods
- Fuzzy methods
- Probabilistic Methods.

There are various inventions which have proposed by various research peoples. We discuss the basic methods here for the better understanding. Hierarchical clustering is proposed earlier, where the documents are clustered hierarchically, in this the time complexity is more. In order to search a document in this way of clustering the processing time is more.

K-means clustering is proposed, where the documents are indexed using the distance between the text documents and the terms in the document. Here the problem of false indexing is present and time complexity is also more. Frequent term based indexing is also proposed, which cluster the document only using the frequency of the terms. For each term in the document the number of occurrence is calculated and total terms are computed and finally frequency of particular term is also computed. Term Frequency TF is computed and Inverse Document Frequency IDF is calculated Based on computed values TF, IDF and Entropy weight is calculated. Based on the calculated weight the document will be indexed to a single cluster.

1. An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection [1] presents a novel ontology-based text-mining approach to cluster research proposals based on their similarities in research areas. The method is efficient and effective for clustering research proposals with both English and Chinese texts. The method also includes an optimization model that considers applicants' characteristics for balancing proposals by geographical regions. The proposed method is tested and validated based on the selection process at the National Natural Science Foundation of China.

2. Document Clustering in Correlation Similarity Measure [2] Space the documents are projected into a low-dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. Since the intrinsic geometrical structure of the document space is often embedded in the similarities between the documents, correlation as a similarity measure is more suitable for detecting the intrinsic geometrical structure of the document space than Euclidean distance. Consequently, the proposed CPI method can effectively discover the intrinsic structures embedded in high-dimensional document space.

3. An Analytical Assessment on Document Clustering [3] analysis of Suffix Tree Clustering (STC) Algorithm and other clustering techniques (K-means, AHC) that are being done in literature survey. The paper also focuses on traditional Vector Space Model (VSM) for similarity measures, which is used for clustering the documents. This paper also focuses on the comparison of different clustering algorithms. STC algorithm improves the searching performance as compare to other clustering algorithms as the papers studied in literature survey. The paper presents STC algorithm applied on the search result documents, which is stored in the dataset.

4. Document clustering and topic discovery based on semantic similarity in scientific literature [4] proposes a modified semantic-based model is proposed where related

terms are extracted as concepts for concept-based document clustering by bisecting k-means algorithm and topic detection method for discovering meaningful labels for the document clusters based on semantic similarity by Testor theory. The proposed method is compared to the Topic Detection by Clustering Keywords method using F-measure and purity as evaluation metrics.

5. Evaluation of text document clustering approach based on particle swarm optimization [5], proposes two techniques for efficient document clustering involving the application of soft computing approach as an intelligent hybrid approach PSO algorithm. The proposed approach involves partitioning Fuzzy C-Means algorithm and K-Means algorithm each hybridized with Particle Swarm Optimization (PSO).

6. Dirichlet Process Mixture Model for Document Clustering with Feature Partition [6] , a novel approach, namely DPMFP, to discover the latent cluster structure based on the DPM model without requiring the number of clusters as input. Document features are automatically partitioned into two groups, in particular, discriminative words and non discriminative words, and contribute differently to document clustering. A variational inference algorithm is investigated to infer the document collection structure as well as the partition of document words at the same time.

7. Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection [7], present an approach that applies document clustering algorithms to forensic analysis of computers seized in police investigations. We illustrate the proposed approach by carrying out extensive experimentation with six well-known clustering algorithms (K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) applied to five real-world datasets obtained from computers seized in real-world investigations. In addition, two relative validity indexes were used to automatically estimate the number of clusters.

8. A Similarity Measure for Text Classification and Clustering [8], proposes a new similarity measure for document clustering. To compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases like the feature appears in both documents, the feature appears in only one document, and the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents.

9. Evolutionary Algorithm for Extractive Text Summarization [9], propose unsupervised document summarization method that creates the summary by clustering and extracting sentences from the original document. For this purpose new criterion functions for sentence clustering have been proposed. Similarity measures play an increasingly important role in document clustering. Here we've also developed a discrete differential evolution algorithm to optimize the criterion functions.

10. An Effective Rule-Based Probabilistic classifier [10] for text mining is discussed. It presents a methodology to generate positive and negative rules. Based on generated rules they calculate probability values to identify the category of document.

11. CCS Classifier for Efficient Text Clustering [11] is proposed which using various measures like Conceptual Strength Measure, Conceptual Depth Measure, and Semantic Similarity Measure to identify the cluster. Finally they compute a combined weight to identify the cluster of the text document.

Here we propose a new way of clustering algorithm of rational method which uses semantic graphs for the computation of different metrics.

3. PROPOSED SYSTEM:

The proposed model consists four stages namely preprocessing where the document content is processed to remove stop word and performed stemming operation to get pure term set. In the second stage semantic graph is generated using the domain ontology, at the third stage rational metric computation is performed and finally the document is clustered.

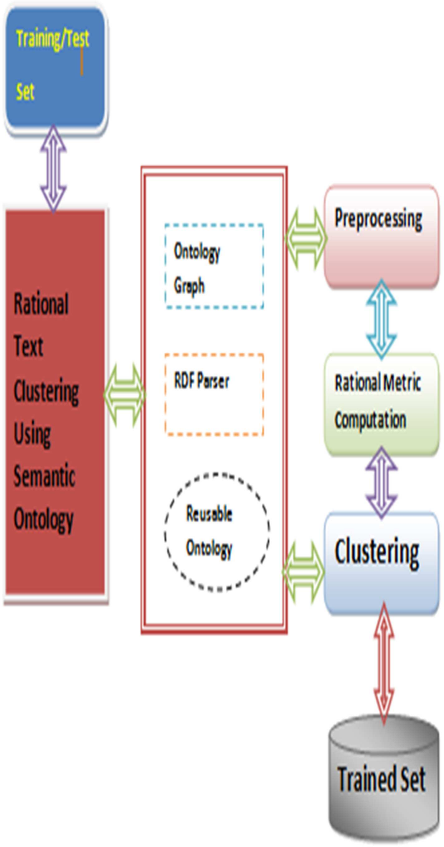


Fig1: Proposed System Architecture

3.1 Graph Generation:

The reusable ontology is loaded and the rdf files are processed by parser to extract the classes, attributes and relations. Using the extracted features the graph is constructed to form semantic graph. The graph consist of nodes connected with relation and each category consist of various graphs all connected to a root node labeled by the name of the category.

```
Algorithm:
Step1: read ontology Os from data base Odb.
Step2: for each ontology O from Os
        Create graph $G_i$ with root node named
              category $C_i$.
        For each label $l_i$ from O
              Create new node $N_i$.
              Identify parent node $P_i$
              Add to parent.
              Assign value the node.
              End
        End.
Step3: end.
```

3.2 Preprocessing:

At the preprocessing stage, our proposed method reads the testing documents and preprocesses the documents from the document set D_s. It reads the textual information from each document D_i from the document set $_{Ds}$ and generates a term set T_s. From the term set T_s , unnecessary words are removed as stop words and verbs also identified using standford part of speech tagger and removed from the term set. With the remaining terms in the term set T_s, stemming process is performed to get pure nouns from the terms in the term set. The selected pure nouns are used to calculate other measures to compute combined weight.

3.2.1 Preprocessing algorithm

```
Step1: Read the documents in the Training set
        $D_{s.}$
Step 2: for each document $D_i$ from Document Set
        $D_s$.
        Extract text content from Document $D_{i.}$
        Split text into Paragraphs.
        Split paragraph into statements.
        Remove punctuation marks.
        Split text into individual terms and
        collect as term set $T_{s.}$
        Remove stop words from the term set
        $T_{s.}$
        For each term in the term set $T_{s.}$
        Use pos tagger to identify verb/noun.
        If verb
              Remove from Term set $T_{s.}$
        Else
              Perform stemming process.
        End.
Step 3. Return Terms Set $T_{s.}$
```

3.3 Rational Metric Computation:

From the extracted feature set T_s , with generated graph set G_s we compute two rational metrics called semantic frequency sf and inverse semantic frequency isf. Semantic frequency represent the semantic relation of the document with a particular document in the cluster and inverse semantic frequency represent the semantic relation with other document in the cluster. We compute both the measures for each term in the term set T_s towards all category of the cluster. Also we compute the semantic deapthness measure , which represents the exact location within the cluster.

Algorithm:

```
Step1: Start
Step2: Initialize set of semantic deapthness set Sd, Depth indicator di, SF, ISF.
Step2: For each term $T_i$ from term set $T_s$.
        For each graph $G_i$ from $G_{s.}$
        For each branch $B_i$ from $G_i$
                Traverse the graph $G_i$
                        If( node value== $T_i$)
                Compute sf =( n/tn)×Log(f+h).
                        SF(i)=sf.
                n-No of node value matches
                tn-Total no of nodes
                f-no of outgoing leafs.
                h-no of incoming leafs.
        compute isf=Log(an/atn).
        ISF(i)=isf.
        an-No of node value matches in other branches
        tn-Total no of nodes in other branches
        if(di<depth of node)
        di=depth of node.
        Sd(i)=di.
        end
                                end
                        end
                end
        end
Step3:  End.
```

3.4 Clustering:

Once the rational metrics are computed then clustering is done using all the values computed. We have a set of semantic frequency and inverse semantic frequency and semantic deapthness measures. Using all these values we compute a semantic weight represents the weightage of document towards a category.

Algorithm:

```
Step1: Start.
Step2: Read all computed values {$T_s$, Sd, SF, ISF}.
Step3: For each category $C_i$
        $Sw=\sum(\sum(SF(T_i)\times ISF(T_i)))\times\sum Sd(T_i))$
        Sw-semantic weight for a category $C_i$.
        End.
Step4: Sort Sw.
Step5: Select top value from Sw.
Step6: Assign document to the category $C_{i.}$
Step7: Stop.
```

4. RESULTS AND DISCUSSION

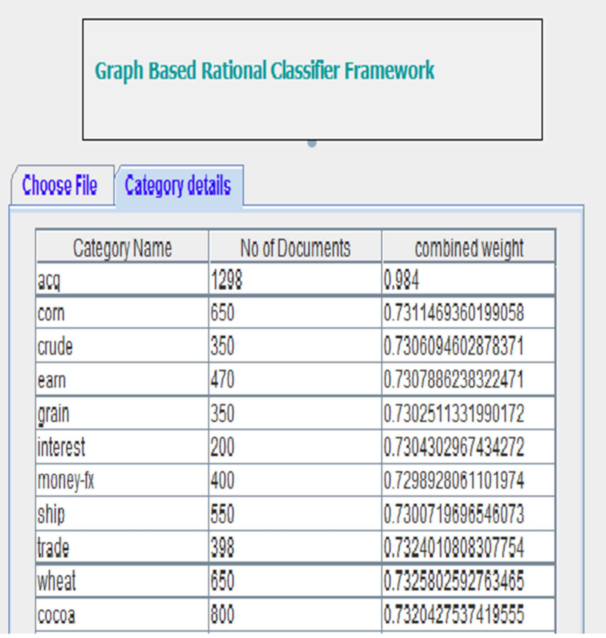
The proposed system produces very good results compare to other algorithms. We used 2 million text documents and 200 concepts to cluster the document. The algorithms used 70 percent of documents for training and 30 percent as testing documents.

We used twenty categories for testing purpose of our algorithm and the following table shows the number of documents used to evaluate the testing.

Category	No of documents	Category	No of documents
acq	793	ship	550
corn	650	trade	398
crude	350	wheat	650
earn	470	cocoa	800
grain	350	Veg-0il	765
interest	200	copper	489
Money-fx	400	housing	567
Money-supply	798	ship	678
coffee	468	cotton	325
sugar	257	carcass	767
reserves	573		

Table1: shows the category names and number of documents used for clustering.

The Table1 shows that number of documents used in each category for training phase. In training phase each document in all the categories are processed for clustering. A part of document in each category is used for testing purpose.



Category Name	No of Documents	combined weight
acq	1298	0.984
corn	650	0.7311469360199058
crude	350	0.7306094602878371
earn	470	0.7307886238322471
grain	350	0.7302511331990172
interest	200	0.7304302967434272
money-fx	400	0.7298928061101974
ship	550	0.7300719696546073
trade	398	0.7324010808307754
wheat	650	0.7325802592763465
cocoa	800	0.7320427537419555

Fig 2: shows the result of generated weight value.

The Figure2 shows the computed weight using our algorithm for a document which is given as input from the category acq. The computed weight shows that the input documents have more weight for the category acq.

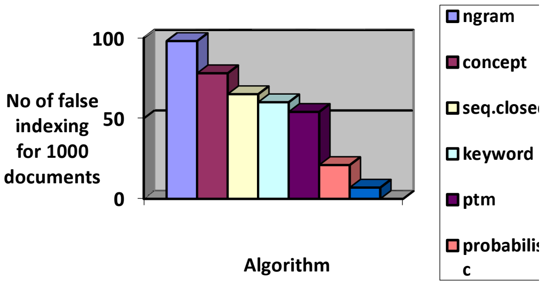


Figure3: Number of false indexing for set of documents by different algorithms

The figure3 shows the comparison of different algorithms and number of false indexing generated by them. It clearly shows that our proposed algorithm reduces the false indexing and increases the efficiency of clustering.

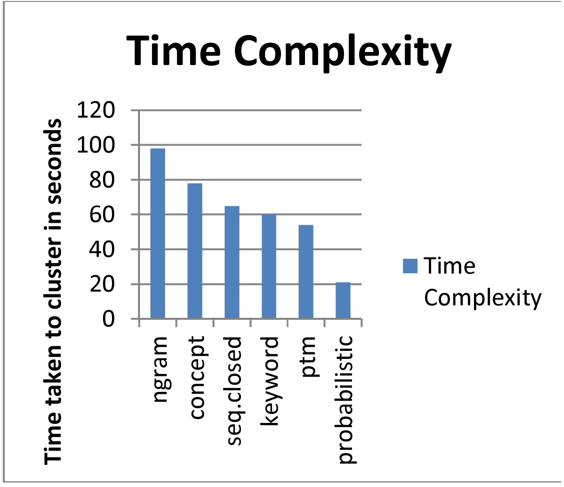


Figure4: Shows The Time Complexity Graph.

The figure 4 shows the time complexity of different algorithms. it shows that the proposed method takes very less time compare to other methodology to cluster 1000 document.

5. CONCLUSION:

We proposed a graph based rational text clustering algorithm which produces very little false indexing, overlap and produces good result. The rational graph based cluster produces good results and the measures computed are very effective and based on the computed measures semantic weight is calculated. We used Reuter's data set and we split each corpus in the data set into ten categories. We used 70 percent of the corpus as training set and 30 percent as testing set. Further we can modify the measures to compute the semantic weight and refine the results produced.

REFERENCES:

[1] Jian Ma, An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection, IEEE Systems, man and cybernetics, volume 42, issue 3, pp 784-790,2012.

[2] Thaiping Zhang, Document Clustering in Correlation Similarity Measure Space, IEEE knowledge and data engineering, volume 24 , issue 6 , pp 1002-1013, 2012.

[3] Pushplata, An Analytical Assessment on Document Clustering, IJCNIS, volume.4, No.5, June 2012.

[4] Jeyabarathy, Document clustering and topic discovery based on semantic similarity in scientific literature, IEEE ICCSN , pp 425-429 , 2011.

[5] Stut Karol, Evaluation of text document clustering approach based on particle swarm optimization, Volume 3, Issue 2, pp 69-90, 2013.

[6] Ruizhang Huang, Dirichlet Process Mixture Model for Document Clustering with Feature Partition, IEEE knowledge and data engineering, volume 25, issue 8, pp 1748-1759, 2013.

[7] Da cruze nassif , Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection, IEEE information forensics and security, volume 8, issue 1, pp 46-54, 2013.

[8] Yung-Shen Lin , A Similarity Measure for Text Classification and Clustering, IEEE Transactions on Knowledge and Data Engineering, 2013.

[9] R. Alguliev , "Evolutionary Algorithm for Extractive Text Summarization,"

Intelligent Information Management, Vol. 1 No. 2, pp. 128-138, 2009.

[10] Yuepeng Cheng, A Document Clustering Technique Based on Term Clustering and Association Rules,” IEEE Trans. Database Technology and Applications (DBTA), pp 1-3, 2010.

[11] S.Subbaiah, Dr. C. Chandrasekar “A Probabilistic Rule Based classifier for Text Clustering”, European Journal of Scientific Research, ISSN 1450-216X Vol. 93 No 4 December, 2012, pp.547-552, Dec 2012.

[12] S.Subbaiah, Dr. C. Chandrasekar “EFFICIENT TEXT CLUSTERING USING C2S CLASSIFIER”, Australian Journal of Basic and Applied Sciences, 7(13) November 2013, Pages: 107-113.

[13] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, “A Two-Stage Text Mining Model for Information Filtering,” Proc. ACM 17th Conf. Information and Knowledge Management (CIKM ’08), pp. 1023-1032, 2008.

[14] Kale, U. Bharambe, "A New Suffix Tree Similarity Measure and Labeling for Web Search Results Clustering", Proc. Second International Conference on Emerging Trends in Engineering and Technology, p.856-861, ICETET,2009.

[15] S.Na,G. yongand L. Xumin, "Research on K-means Clustering Algorithm",IEEE Symposium on intelligent Information Technology and security informatics, pp 63-67, 2010.

[16] Xiuguo Chen, Weighted k-Means Algorithm Based Text Clustering, Information Engineering and Electronic Commerce(IEEC) ,pp 51-55, 2009.

[17] Chen He-Nian , A Text Clustering Method Based on Two-Dimensional OTSU and PSO Algorithm, Computer Network and Multimedia Technology, pp 1-4, 2009