ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Novel Approach for Text Clustering

Simma Tejeswararao
M.Tech (CSE)
Department of CSE
Avanthi Institute of Engineering & Technology.

B.Ramesh Babu, M.Sc, M.Tech
Assistant Professor
Department of CSE
Avanthi Institute of Engineering & Technology.

Dr.A.Chandra Sekhar (Ph.D)
Professor & HoD
Department of CSE
Avanthi Institute of Engineering & Technology.

ABSTRACT

Text clustering is used to group documents with high levels of similarity. It has found applications in different areas of text mining and information retrieval. The digital data available nowadays has grown in huge volume and retrieving useful information from that is a big challenge. Text clustering has found an important application to organize the data and to extract useful information from the available corpus. In this paper, we have proposed a novel method for clustering the text documents. In the first phase features are selected using a genetic based method. In the next phase the extracted keywords are clustered using a hybrid algorithm. The clusters are classed under meaningful topics. The MLCL algorithm works in three phases. Firstly, the linked keywords of the genetic based extraction method are identified with a Must Link and Cannot Link algorithm (MLCL). Secondly, the MLCL algorithm forms the initial clusters. Finally, the clusters are optimized using Gaussian parameters. The proposed method is tested with datasets like Reuters-21578 and Brown Corpus. The experimental results prove that our proposed method has an improved performance than the fuzzy self-constructing feature clustering algorithm.

INTRODUCTION

Document clustering (or text clustering) is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.

The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.

Text clustering is an unsupervised learning process that is independent of the prior knowledge of data collection or input dataset. There is a strong relation between data mining and machine learning or learnarability.The process of learning may be described as the compression of datasets from the mathematical perspective[5,7].

The relation between complexity of datasets and learnability may be stated mathematically as follows. If we can design an algorithm ‘A’ which can compress the input dataset, D to a feasible or reduced dataset D’ then this means that we have learned some important

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

information. [15-17]. This means that there exists a strong relation between datasets and learnarability.

Dimensionality reduction is also one of the key issues in text clustering and text classification. In the perspective of text mining, dimensionality reduction may be viewed as technique applied to learn and deduce new information from existing input datasets or to achieve learnarability. In this context we can state that learnarability and complexity of datasets are strongly related and hence must be handled very carefully and efficiently.

Let F be a set of all features in a given dataset. Let P be a process of applying dimensionality reduction to dataset D. Now after applying the process P on Dataset D, the feature set F is reduced to F' with reduced number of features or predictive components. In this process of reducing a given dataset, there is also a chance of missing valuable features which may have shown significant contribution in decision making. Such a situation must also be taken care when reducing or eliminating predicate components.

The problem of dimensionality reduction using the concept of finding frequent item or itemsets is gaining significant importance from datamining researchers [18] and this forms basis for the current work. However the method of dimensionality reduction by eliminating stop words, stemming words or using TF-ITF methods has been dealt in the previous works [4].

Clustering is usually carried out using two strategies [3,12]. Incremental Strategy or Complete Strategy also called as Static Strategy. In the Incremental Clustering approach, we may need to re-cluster the text files as and when a new text file is considered for clustering. The main requirement for designing any clustering algorithm is to first design a suitable similarity measure. This may be followed by the design of clustering algorithm or using any of the existing approaches.

Existing System:

Text clustering is used to group documents with high levels of similarity. It has found applications indifferent areas of text mining and information retrieval. The digital data available nowadays has grown inhuge volume and retrieving useful information from that is a big challenge. Text clustering has found animportant application to organize the data and to extract useful information from the available corpus.

Disadvantages:

1. Time complexity is more
2. Performance is low

Proposed System:

We have proposed a novel method for clustering the text documents. In the first phase features are selected using a genetic based method. In the next phase the extracted keywords are clustered using a hybrid algorithm. The clusters are classed under meaningful topics. The MLCL algorithm works in three phases. Firstly, the linked keywords of the genetic based extraction method are identified with a Must Link and Cannot Link algorithm (MLCL). Secondly, the MLCL algorithm forms the initial clusters. Finally, the clusters are optimized using Gaussian parameters.

Advantages:

1. It reduce the time Complexity
2. It improves the efficiency and performance

A.1 Algorithm Text_Clustering and Topic Identification (N, frequent items, Text files)

// N – number of text files

A.1.1 Preprocessing Phase

Begin
Check if the input file is in .txt or .doc or .docx format. If not, convert it in to proper format

Step1:
For each text file of the form .txt or .doc do
Begin
Step a. Eliminate Stop words
Step b. Eliminate Stemming words
Step c. Apply any frequent item finding algorithm
Step d. Define feature word size equal to size of all Frequent itemsets obtained in step2

End for
End of Preprocessing Phase

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

B. Processing Phase

Step 2: Form a feature set consisting of m Words consisting of each word in frequent items of each document.

Step 3: Form Binary Matrix with row indicating text file and column each frequent itemset respectively

For each text file in input file set do

Begin

For each frequent item set in feature set do

Begin

If (f_k in feature set W is in text file T_i)

Begin

Define Cell value $M[T_i, F_k] = 1$

// 1 indicates presence of frequent item

Else

Define Cell value $M[T_i, F_k] = 0$

// 0 indicates absence of frequent item

End if

End for

End for

Step 4: Obtain Similarity matrix for each pair of text files applying new similarity Function defined in table 1

Step 5: Count the number of 0's and place the count in the matrix for each pair of text files.

Step 6: Find the cell with maximum value. Group each pair of each such files into a new cluster.

Step 7: Repeat Step6 until no files exist or we reach the stage of first minimum value leaving zero entry.

Step 8: Display all the clusters formed.

Step 9: Identify topics. Give label to each cluster.

End of Processing Phase

Modules:

1. Keyword Extraction
2. Probability crossover
3. Clustering Process
4. Mutation

Keyword Extraction:

The initial phase of the keyword extraction involves pre processing of the document. The terms need to be assigned with a weight that will help in prioritizing them within the population. Once the initial weight is calibrated the genetic procedures are executed to gain a final keyword population. The terms are chromosomes and the weights are the numeric representation of genes. A simple modified arithmetic technique is applied for crossover, trialedby the "Expected Number of Elements in the Population" [10] viewpoint to declare the fitness of the engendered populace. Mutation is alleged only if the fitness utility is not contended for cessation.

Probability crossover

The basic principle behind crossover involves the divide and conquer method. The population is broken into two halves where the first segment contains the better half and the rest holds the weight of the lower probability population. Wiindicates the weight of a word in position "i". Pciis the probability ratio of the most feasible word with respect to the word that has the highest occurrence in the other part of the division.

Clustering Process:

In this phase, the prime attribute that is taken into consideration is the high dimensionality of the document space. The proposed system uses employs three different mechanisms. The first stage is the identification of related words in a document using the MLCL algorithm. The relevant keywords are grouped to form clusters using three mainequations. Thereafter the clusters formed are optimized using Gaussian parameters. The Gaussian parameter identifies the word patterns and standard deviation of clusters. Then the words are grouped in accordance with the Gaussian outputs and colligated into clusters with reference to the documents.

Mutation:

Mutation engrosses the amendment of term weights with a probability mutation "pm". Mutation has the

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

capability to reinstate the mislaid genetic material into the population, thus thwarting the convergence of the solution into a suboptimal region or divergence into an infinite loop. T (i) decides whether the mutation process is to be applied or not. When a word does not lay within the fitness condition the process of mutation is being applied. The fitness value determines mutation. When two consecutive iterations have a similar weight-age for the terms, the ultimate keyword list is generated. If a word is fit, the mutation would not be applied.

CONCLUSION

The Proposed algorithm has the input as similarity matrix and output a set of clusters as compared to other clustering algorithms that predefine the count of clusters. In this work, frequent items are generated using APRIORI approach by following a similar method. We can replace apriori algorithm by any frequent item finding algorithm. The algorithm for clustering considers the set of frequent items generated from all the documents. This gives the commonality between document pairs. The count of frequent items serves as the distance measure.

REFERENCES

1. Congnan Luo, , Yanjun Li, Soon M. Chung. Text document clustering based on neighbors, Data & Knowledge Engineering (68), 2009, 1271–1288

2. Tianming Hu, Sam Yuan Sung, Hui Xiong, Qian Fu. Discovery of maximum length frequent itemsets, Information Sciences (178), 2008, 69–87

3. Wen Zhanga,, Taketoshi Yoshida, Xijin Tang, Qing Wang. Text clustering using frequent itemsets, Knowledge-Based Systems 23 (2010)379–388

4. Wen Zhanga,Taketoshi Yoshida, Xijin Tang. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications 38 (2011) 2758–2765

5. Vincent Labatut and Hocine Cherifi. Accuracy Measures for the Comparison of Classifiers, ICIT 2011, The 5th International Conference on Information Technology

6. Jung-Yi Jiang et.al A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

7. Melita Hajdinjak, Andrej Bauer. Similarity Measures for Relational Databases, Informatica 33 (2009) 143–149

8. R. Agrawal, T. Imielinski, A. Swami. Mining association rules between sets of items in very large databases, Proceedings of the ACM SIGMOD Conference on Management of data, 1993, pp. 207–216

9. F. Beil, M. Ester, X.W. Xu, Frequent term-based text clustering, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 436–442

10. Yung-shen Lin, Jung-Yi Jiang et.al “A similarity measure for text classification and clustering “, IEEE Transactions on Knowledge and Data Engineering, 2013.

11. Yi Peng, Gang Kou. A Descriptive frame work for the field of data mining and knowledge discovery , International Journal of Information Technology and Decision making, Volume 7, No.4, 2008,Pages 639-682,Impact Factor 3.139

12. Kou,G and Lou,C. Multiple Factor Hierarchical Clustering Algorithm for Large Scale Web Page and Search Engine Clickstream Data, Vol 197,Issue 1,Page 123-134, Annals of Operation Search,2012.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

13. Niraj Kumar, Kannan Srinathan. Automatic Key phrase Extraction from Scientific Documents Using N-gram Filtration Technique. Proceedings of the eighth ACM symposium on Document engineering, Pages 199-208, 2008

14. G.Suresh Reddy, Dr.T.V.Rajinikanth, Dr.Ananda Rao Text Clustering Using Frequent Patterns and Jaccard Dissimilarity Function, Proceedings of the second ICACM 2013.

15. Pieter Adriaans and Dolf Zantinge. Data Mining. Eleventh Pearson Education. 2013

16. Daniel Larose. An Introduction to Data Mining. John Wiley Publications

17. Daniel Larose. Data Mining. Models and Methods.2012. John & Wiley Publications

18. K. Ravi Shankar, G.V.R. Kiran, Vikram Pudi. "Evolutionary clustering using frequent itemsets". ACM StreamKDD '10. Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques, Pages 25-30, 2010.