International Journal of Control and Automation
Vol. 13, No. 2s, (2020), pp. 123-128

Comparative Analysis of Unstructured Text Clustering Techniques

K.V. Kanimozhi
Assistant Professor, Computer Science and Engineering
Saveetha School of Engineering, Chennai, Tamilnadu, India.
kani.kalai4@gmail.com,kanimozhikv.sse@saveetha.com

Abstract

In recent times due to the increase of internet growth and technological advancement most of the data generated are of unstructured text data. Text mining plays an important role in information retrieval. Therefore the unsupervised learning technique like clustering is mainly applicable for unstructured text data. To improve the quality of information retrieval text clustering gives precise mining by organizing clusters of similar documents from huge collection. It is a current challenge to explore meaningful and compact insights from massive collections of the unstructured documents. Although there exists a lot of text clustering technologies, most of them are not scalable and clusters provided are not that much efficient to give desired result, and it leads to huge computation time. This paper gives the detailed comparative analysis of different clustering techniques which improves the precision of retrieval accuracy and leads to efficient clusters.

Keywords: *Unstructured Text , Clustering, Similarity, Frequent Item, Clustering*

1 Introduction

Since the data grows in rapid manner because of the increasing globalization, the contemporary challenge is to extract useful hidden information from those huge unstructured text. In Text mining the text clustering or document clustering is one of the important concepts which cluster the similar documents with similar contents mainly to enhance reliability and efficiency of different text mining applications such as text categorization, document classification, and information retrieval. The main challenge of the clustering is huge dimension and automatic grouping of clusters numbering based on the similarity with précised clusters and topic modeling. The main theme about the clusters is that objects within the valid clusters are more similar to one another.

Requirements of Clustering:

The different requirement of cluster analysis is
1. To process with various attributes types.
2. To find out the clusters with arbitrary shape.
3. Scalability
4. Huge Volume
5. To interpret and usage
6. Capability of handling noisy data
7. Due to unsupervised less domain knowledge is required to find input parameters.
8. To support incremental clustering
9. Constraint based clustering.

2 Literature Survey

The paper proposed the different clustering algorithm implementations like K-means, K-mediods, Single Link, Complete Link, Average Link, and CSPA approach for various real time datasets. From implementations the clustering results using Average Link and Complete Link algorithms provides better solutions.[1].

ISSN: 2005-4297 IJCA

Aim of the proposed work is to categorize different jobs in unstructured text documents which solves the problem of data sparseness to extract more knowledge and takes care of cluster accuracy in better way and achieves good accuracy.[2].

This paper presents a detailed survey on mining the interesting knowledge from weblogs. [3].

To construct the Chinese domain ontology efficiently the new suggested method works well for unstructured text documents by episode based ontology construction.[4]

This paper works mainly deals with two different clustering algorithm based on frequent word sequence and clustering algorithm based on frequent word meaning sequences and compared the work with bisecting k-means clustering algorithm and proved that the proposed methodologies achieves better cluster accuracy than bisecting k-means.[5]

The proposed work uses k-means clustering algorithm with the concepts of neighbors and link in three various views, Depend on the ranks of candidate textual documents proposed method finds the initial centroid for the cluster. Secondly cosine with link functions is to provide better similarity measures between the text documents. Thirdly it focuses mainly on the novel heuristic method based on neighbors for cluster selection and proved the proposed methods achieves increased performance and enhances the accuracy with less processing time. [6]

The proposed methodologies analyze social media tweets considered the features from it and measured the performance using two parameters like relevance and representativeness. The output obtained is of informative and useful insights from the languages from humor and irony.[7]

From the huge volume of time stamped web document streams they proposed a efficient technique fouses on event detection from various social perceptions where the results proves rich representations and improves the accuracy of detected events and methodology proposed is linear for online facet versus total number of text documents present in data stream. [8].

The paper implements and discusses about various real time datasets to solve real time problem which principal component analysis is done and compared and the results shows an improved performance with reduced computation time.[9].

A new strategy is implemented using DBSCAN clustering algorithm tested with real and synthetic datasets and compared the accuracy with the conventional methods. The main objective of the work is to find the regional hazard regions present in the Japan Seismic region. [10].

The implemented work concentrates on ontology and developed a new model on learnable focused crawling framework using artificial neural network concept to categorize the web pages. And the paper mainly compares the results of proposed model with breadth-first search crawling approach, and the output shows better result for proposed model which use only domain specific ontology. [11].

In order to evaluate the performance of hypernymy, hyponymy, holonymy, and meronymy, the very effective method is chosen and compared with word net based clustering. And the implementation output shows improve effectiveness from given order or semantic representation of clustering and finally proves that noun phrase based shows enhanced clustering. [12]

The paper discusses about the comparative execution of backmarking method with forward evaluation method with respect to precision and recalland the result proves that forward evaluation methods performs better than backmarking and more efficient for the input datasets.[13]

3 Problem Statement and Proposed Solution

Text processing mainly faces a major problem due to massive volume. ie. scalability , moreover to improve the cluster quality and increase the efficiency of clustering and minimize the processing time .there are various important techniques mainly applicable in the area of text mining are proposed and proves that these techniques gives better results based on the input dataset using tools like hadoop.

The different techniques are majorly applicable as follows, like in Partitioning based method, given a larger database of textual data, this method partitions the data values where every partition denotes a cluster,

In Hierarchical based method from the given input database this hierarchical method performs hierarchical decomposition where the method may be divisive based on how the hierarchal decomposition is done. In Density based method as long the increasing volume f data in the neighborhood which exceeds some threshold the cluster density is also grows. In Grid based method the main clustering operations are done on grid structure. To quantize the object space into finite number of cells this creates the grid structure.
In Model based method to calculate the best fit of the input data the model based methods hypothesize a model for each of the clusters. In Clustering high dimension**,** clustering the high dimension is very significant task in clustering analysis as many applications benefit from this method where it contains huge number of features or dimensions. Secondly in Constraint based clustering, based on user specified constraint or application oriented constraint the constraint based clustering is performed.

Vector space model:

Here in vector space model in unstructured data most of the text documents are analyzed by vector space model, where every text document d is represented as vector in term space, n is the number of documents in the dataset and usually considered as term frequency vector (tf). From the document the individual terms are taken and in total represented by D and the frequency of the term i in the unstructured text document is considered as tf_i,

$$D_f = [tf_1, tf_2, \ldots tf_D]$$

And the term frequency and the inverse document frequency is represented by

$$D_{tf\text{-}idf} = [tf_1 \log(n/df_1), tf_2 \log(n/df_2) \ldots tf_D \log(n/df_D)]$$

and the centroid vector c_j is calculated by taking the document vector s input by

$$C_j = \frac{1}{|Cj|} . \sum_{di \in cj} d_j$$

Similarity Measure:

Although there are lot of similarity measures available in most of the text documents analysis for calculating the similarity between the text documents, cosine similarity is best suited and mainly taken to measure similarity between two documents.

$$\text{Sim}(\vec{d}_i, \vec{d}_j) = \cos(\vec{d}_i, \vec{d}_j) = \frac{\vec{d}_i \vec{d}_j}{|\vec{d}_i| . |\vec{d}_j|}$$

To find the similarity between the clusters, we use,

$$\text{Sim}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum d_i \in cluster_i, d_j \in cluster_j . \cos(\vec{d}_i, \vec{d}_j)}{|cluster_i| * |cluster_j|}$$

Where the size of the clusters are represented by $|cluster|_i$ and $|cluster|_j$.

Topic Modeling:

Using probabilistic latent semantic technique combined with cosine similarity proves best for finding out he efficient clusters, by comparing the huge volume of documents with input from search phrase and related are retrieved immediately and topics can be labeled using this method in comparison with singular

value decomposition, Latent semantic indexing .the dimension reduction is efficient using this probability based cosine similarity method.

Frequent item based method:

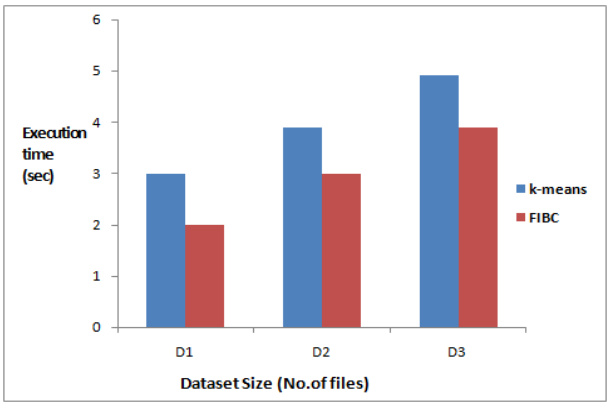
Recent challenges mainly focuses on frequent pattern mining which find outs the frequently occurring pattern from huge volume of data. It mainly discovers the correlation and association between the data values[14,15].
Eg. frequent term based cluster analysis. Here the individual terms are identified and extracted by tokenization method, then the unnecessary words are eliminated by stop word removal method, From the generated terms from huge document collection are used to calculate the frequent items occurred in the dataset.

Applications: The major applications where the unstructured text document plays main role in

- Telecommunication industry.
- Retail Industry.
- Financial data analysis
- Biological data analysis.
- Scientific data analysis
- Intrusion detection.
- Fraudulent analysis.
- Forensic analysis etc.

4 Experiments and Results Analysis:

The different subsets of various datasets from 20Newsgroup real datasets are extracted as D1, D2 and D3 which contains around 50, 100 and 500 documents respectively. The K-means algorithm using vector space model is implemented and evaluated and secondly frequent item based clustering (FIBC) is implemented. The result proves for this datasets as frequent item based clustering works better for huge dimension reduction compared to the k-means algorithm and shows better clustering when compared to K-means by reduced execution time.



5 Conclusion

This papers gives the most significant techniques used for unstructured text documents specifying the important requirements of clustering and suggested the two major algorithm suited for textual clustering like K-mean and the frequent item based clustering, and compares the two methodologies with varying newsgroups dataset for scalability check and proves that frequent item

based method works better by reduced dimension and execution time and yields good cluster quality than the k-means algorithm hence it leads to increase in precision of retrieval accuracy. Hence Frequent item based algorithm is more efficient than vector space technique like K-means algorithm. And also we suggested the probabilistic based model using cosine similarity best suits for topic modeling.

References:

[1] Laxmi Lydia, E., Sharmili, N., Nguyen, P. T., Hashim, W., & Maseleno, A. (2019). Automatic document clustering and indexing of multiple documents using KNMF for feature extraction through hadoop and lucene on big data. Test Engineering and Management, 81(11-12), 1107-1130. Retrieved from www.scopus.com

[2] Reddy, S. S. R., Malathi, P., & Mahalakshmi, D. (2019). Efficient datacenter clustering in map reduce framework using cache index algorithm. Test Engineering and Management, 81(11-12), 5418-5422. Retrieved from www.scopus.com

[3] Vanitha, R. (2019). BOVW classification method with particle swarm optimization in big data. Test Engineering and Management, 81(11-12), 4529-4535. Retrieved from www.scopus.com

[4] Kumar, R., & Bhardwaj, D. (2020). An improved moth-flame optimization algorithm based clustering algorithm for VANETs. Test Engineering and Management, 82(1-2), 27-35. Retrieved from www.scopus.com

[5] Kumudha, & Shanmuga Prabha, P. (2020). Effective combination of biclustering mining and adaboost learning for breast tumor analyzation. Test Engineering and Management, 82, 2060-2063. Retrieved from www.scopus.com

[6] Lee, J. -. (2020). Distance-based 2-hop clustering algorithm in WSNs (wireless sensor networks). Test Engineering and Management, 83, 4299-4306. Retrieved from www.scopus.com

[7] Madhumitha, S., & Ashwini, S. (2020). Foresight-based and locality-aware task setup for complementing video transcoding over deep learning clustering. Test Engineering and Management, 82, 2015-2019. Retrieved from www.scopus.com

[8] Amru, M., & Bhavani Sankar, A. (2019). Analytical study on design and development of herichal clustering using adhs algorithm. International Journal of Advanced Science and Technology, 28(20), 1208-1213. Retrieved from www.scopus.com

[9] Bakeyalakshmi, P., & Mahendran, S. K. (2019). Integrated clustering based intrusion detection model (IC-IDM) in mobile ad-hoc networks. International Journal of Advanced Science and Technology, 28(12), 306-318. Retrieved from www.scopus.com

[10] Dinesh, G., Rajinikanth, C., Maheswara Venkatesh, P., & Vanitha, T. V. (2019). An optimized Dijkstra's SPF algorithm used multihop clustering for scalable IoT systems. International Journal of Advanced Science and Technology, 28(14), 298-306. Retrieved from www.scopus.com

[11] Joseph, J., & Kesavaraj, G. (2019). Evaluation of clustering algorithms for credit card data set using WEKA. International Journal of Advanced Science and Technology, 28(17), 392-400. Retrieved from www.scopus.com

[12] Joshi, R. R., Mulay, P., Lohia, A., Singh, A., & Nagar, A. (2019). Mapreduce4CFBA: Distributed incremental closeness factor based clustering algorithm (DICFBA) for analysis of chronic diseases on hadoop mapreduce. International Journal of Advanced Science and Technology, 28(1), 241-253. Retrieved from www.scopus.com

[13] Kiruthika, M., & Sukumaran, S. (2019). Multi-modal approach with deep embedded clustering for social image retrieval. International Journal of Advanced Science and Technology, 28(17), 946-995. Retrieved from www.scopus.com

[14] Madhu, T., Sarma, S. S. V. N., & Murthy, J. V. R. (2019). Fault containment based self stabilized fuzzy relevance clustering algorithm. International Journal of Advanced Science and Technology, 28(14), 502-512. Retrieved from www.scopus.com

[15] Goar, V. K., Kuri, M., & Tanwar, P. K. (2019). Energy efficient base clustering algorithm to achieve optimum value of energy consumption. International Journal of Control and Automation, 12(5), 581-587. Retrieved from www.scopus.com
[16] Kaur, N., Verma, S., Kavita, & Yadav, A. L. (2019). Survey on classification and clustering schemes in big data using image processing. International Journal of Control and Automation, 12(4), 170-188. Retrieved from www.scopus.com

ISSN: 2005-4297 IJCA