ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

An efficient document clustering by using adaptive k-means clustering algorithm

R. Umamaheswari[1], Dr.N. Shanthi[2],S.Lalitha[3]

[1,3]Assistant Professor, [2]Professor

Department of Computer Science and Engineering,

[1,3]GnanamaniCollegeofTechnology, Namakkal, T. N, INDIA

[2]NandhaEngineeringCollege, Erode, T. N, INDIA

umait1978@gmail. com , shanthimoorthi@yahoo.com , lalitha@gct.org.in

ABSTRACT- Now-a-days the web content increases day-by-day and searching relevant information from web increases lots of overhead. The text documents are in various formats like text, audio, video etc. The text documents are in the form of structured, unstructured and semi structured. The clustering is the process of grouping the text documents into meaningful clusters which gives better search results. To retrieve relevant useful documents the new technique is introduced for efficient clustering of unstructured documents.

Keywords: Document Clustering, Information retrieval, K-Means algorithm, Web Search.

1. INTRODUCTION

Web Search is the process of extracting information from World Wide Web (WWW). Text mining research includes several statistical machine learning algorithms for classifying the documents. Due to the huge existence of web based information, it is difficult to identify the exact and appropriate information. Relationship among user query and document matching is measured by using similarity scores. A similarity computation is an important part in information retrieval and text mining. Many analyses are made on gathering the information from unstructured data. Most of the web documents are does not having proper structure. Therefore, information retrieval holds huge challenges while collecting information from the web. Hence, clustering algorithms and classifiers are utilized for creating new classes from unstructured documents. In web search process, the difficulties are arises from information management, searching and retrieval of relevant documents, etc. It is necessary to develop the methods to organize large amount of web data into smaller number of meaningful clustering which will help to solve most of the problems. Clustering and classification are the most general methods for grouping the documents in a successful manner. K-Means Algorithms gives good clustering results for large and unstructured documents. So in this work the adaptive K-Means algorithm is proposed for cluster the unstructured documents.

2. RELATED WORK

Convolutional Neural Network (CNN) model is introduced by Peng Wang et al. (2016). Practically, the semantically associated words are generally nearer to each other in embedding regions. Therefore, a semantic group through fast clustering is initially identified. Euclidean distance among semantic units and semantic groups are calculated for identifying accurate semantic units.

A new representation technique is called WordNet-based lexical semantic Vector Space Model (VSM) is introduced by Long

Jun et al. (2015) for handling text document errors. In lexical semantic document representation, WordNet is used to organize a data structure of semantic component information for representing lexical semantic contents. In lexical-semantic space of corpus, the lexical-semantic eigenvector of document representation is constructed through measuring the weight of each synset. Moreover, Neighbor-Weighted K-Nearest Neighbor (NWKNN) algorithm is applied for classifying text corpus. Integrated forms of WordNet through lexical chains are introduced by Tingting Wei et al. (2015) for text classification. But, the use of integrated form, improves the difficulties of document representation. Therefore, a modified WordNet with lexical chains approach is designed for developing ontology hierarchical structure. Fuzzy based methods are the established to interpret the uncertain information. The combination of Fuzzy and Ontology depended information retrieval model by Balasubramaniam (2015) produces better results for handling semantics and the ambiguity of information. A methodology for clustering using WordNet and lexical chains is designed for enhancing the text classification. A modified WordNet-based semantic similarity calculation is planned for discovering meaningful word. The WordNet-based semantic measure is used solve document clustering problems such as transferring suitable description, disambiguating the uncertain and synonymous words. However, clustering using WordNet and lexical chains method only addresses above issues.

A unified framework designed by Peng Wang et al. (2016) increases the short texts depending on word embedding clustering and Convolutional Neural Network (CNN).In embedding spaces, the limited Nearest Word Embeddings (NWEs) of semantic units are selected with expanded matrices where semantic cliques are employed as supervision information.An integrated method of WordNet with lexical chains is designed for solving clustering problems in text clustering and proving improved classification accuracy. The merged form of explicit and implicit semantic relationships in WordNet pays an important role for evaluating word sense similarity. In addition, WordNet-based similarity measure is applied to evaluating the true number of clusters by observe the obtained results. Moreover, the lexical chain features is employed to effectively increase the quality with minimized number of features in the document clustering process. However, WordNet-based similarity model is attempts to examine the possible impact of few searches in lexical chains on text clustering.

2.1 K-MEANS CLUSTRING

Clustering group unordered text documents into meaningful and coherent clusters. Document clustering is an essential technique for data analysis. Document Clustering using K-Means algorithm is designed by Irwan Bastian et al. (2016). The documents similarity is computed through the Winnowing algorithm and Cosine algorithm. Topic modeling is a statistical model used for determining the latent topics in documents through machine learning. Latent Dirichlet Allocation (LDA) is a modeling approach by Chyi-KweiYau et al. (2014). LDA and its extensions are studied for dividing set of scientific publications into many clusters. A collection of documents are created with academic papers from various fields and checks whether clustered documents are from same field.Text clustering is an essential application for data mining. Many models designed by Sumayia Al-Anazi et al. (2016) groups the capstone project documents by three clustering methods like k-means, k-means fast, and k-medoids. Three similarity measures are verified like cosine similarity, Jaccard similarity and Correlation Coefficient. The quality of the obtained models is evaluated and compared. A hybrid document clustering technique is designed by

TanmayBasu and C. A. Murthy (2015) with hierarchical and k-means clustering techniques. A distance function identifies the distance between hierarchical clusters. A gram-based framework is designed by Haoji Hu et al. (2014) with maximum filter performance. The framework selects the high-quality grams as prefix of query consistent with ability to filter candidates. A method designed by Mario Beauchemin (2015) constructs the affinity matrices for spectral clustering from density estimator depending on K-means with subbagging process. A partitioned k-means clustering (PKM) scheme designed by Shikui Wei et al. (2012) creates large and unbiased vocabulary with small training set. Original space is divided into many subspaces and executes separate k-means clustering process in all subspace. A complete visual vocabulary is framed by joining many cluster centroids from subspaces. But, it failed to form the matrix.A new parallel algorithm is designed by Yanjun Li et al. (2015) for text document clustering depending on concept of neighbor. When two documents are same, they are taken as neighbors of each other. But it increases time for clustering.

The various techniques were analyzed and the new technique is proposed for text document clustering.

3. PROPOSED WROK

An efficient clustering approach is introduced to construct a good classifier model for analysing new documents. Umamaheswari R et al. (2015) introduced Binary Term Frequency Matrix (BTFM) for constructed input document by using WEKA Tool. Term Frequency Matrix is constructed with the help of binary weighting. If a Term is present in the document, then the value is 1, otherwise the value is 0. BTFM is fed into the K-Means algorithm which is a machine learning algorithm used to construct good classification model. After performing K-Means clustering, resultant class label is converted into neural network input class label matrix. Neural network accepts the original BTFM and class label matrix and hence produces good classification results in various iterations. These results of clustering are used for classifying new text documents.In this work the clustering process efficiency is analysed with various existing methods.

Adaptive K-Means clustering algorithm to achieve betterclustering results with minimum computation for large datasets. In adaptive K-Means clustering, each cluster's center is described by the mean value of objects in that cluster. This mean value is obtained by using squared Euclidean distance which helps in finding the distance between documents while constructing similarity. Algorithmic representation of adaptive K-Means clustering is shown in figure 3.1 as follows.

Input: Data set 'D' containing 'n' text documents '$X_1, X_2 \ldots X_n$', Input query with number of terms
Output: A set of 'k' clusters
Step 1: Begin **Step 2:** Select 'k' documents randomly from 'D' as the initial cluster **Step 3:** Measure cluster centers for each cluster **Step 4:** Calculate similarity of an text documents by using equation (3.1) and (3.4) **Step 5:** Measure Euclidean distance of two documents 'X_1' and 'X_2' $d(X_1 \; and \; X_2) = \sqrt{(X_1 - X_2)(X_1 - X_2)^T}$ **Step 6:** Reassign each document to a cluster 't' whose center isclosest according to the measured distance **Step 7:** Update each cluster means of text documents **Step 8:**Go to Step 3 until there is no change in the cluster **Step 9:** Obtained a set of 'k' clusters of text documentsfrom 'D' **Step 10: End**

Figure 3.1 Adaptive K-Means clustering algorithm

As shown in figure 3.1, adaptive K-Means algorithm is implemented in this proposed method to cluster the input text documents based on their similarity. This clustering is achieved by utilizing the default random selection of cluster with a centroid (i.e.,cluster center). Euclidean or cosine distance measure (i.e., from a text document to a cluster center) is used for obtaining term similarity weight of the documents. Finally, the proposed BTFM method updates all the clusters until there is no change in cluster movements. The output of adaptive K-Means algorithm is given as input to the Back Propagation Neural Network pattern recognition which is illustrated as follows.

Qinbao Songet al. (2013) performed fast clustering-based feature subset selection by initially separating the features into clusters. Then features related with target classes are chosen from each cluster to obtain a subset of independent features. But different types of correlation measures are not implemented for clustering. Therefore proposed method is utilized for effective information retrieval with the help of adaptive K-Means algorithm whereEuclidean and Cosine distance measures are employed for finding the distance between terms in text documents.

4. RESULTS AND DISCUSSION

Twenty newsgroup dataset that contains unstructured text documents are taken from UCI repository for performing text analysis. In the proposed approach, ten newsgroup dataset are considered where initial analysis is carried out with adaptive k-means algorithm.

The proposed method is compared with the existing methods such as Clustering and Convolutional Neural Network developed byPeng Wang et al. (2016) and Dimensionality Reduction method with Hidden Markov Model (DR-HMM) developed by A. Seara Vieiraet al. (2016) for analyzing the performance.

4.1 Analysis clustering efficiency

Clustering efficiency is defined as the measure of efficiency occurred during cluster formation using similar words from different text documents. Clustering efficiency is mathematically formulated as follows.

$$CE = \frac{number\ of\ clustered\ relevant\ documents}{Number\ of\ text\ documents} * 100 \qquad \text{Eq. (4.1)}$$

From equation (4.1), Clustering Efficiency 'CE' is calculated as the ratio of clustered relevant documents based on user query to the number of available text documents. Clustering efficiency is measured in terms of percentage (%).When clustering efficiency is higher, the method is said to be more efficient for classification.

Table 4.1 Tabulation forclustering efficiency

Number of input queries	Clustering efficiency (%)		
	Existing CCNN	Existing DR-HMM	Proposed BTFM
10	68.1	60.9	74.6
20	69.4	63.2	76.9
30	71.5	66.1	79.1
40	74.2	67.3	83.2
50	75.9	70.8	84.7
60	78.4	71.2	86.3
70	81.4	72.9	87.4
80	82.6	74.1	89.1
90	84.3	75.8	91.2
100	85.9	77.6	92.5

Table 4.1 demonstrates clustering efficiency for proposed method and various existing methods. From table 4.1, it is clear that clustering efficiency is comparatively

improved in proposed clustering method when compared to other existing methods.

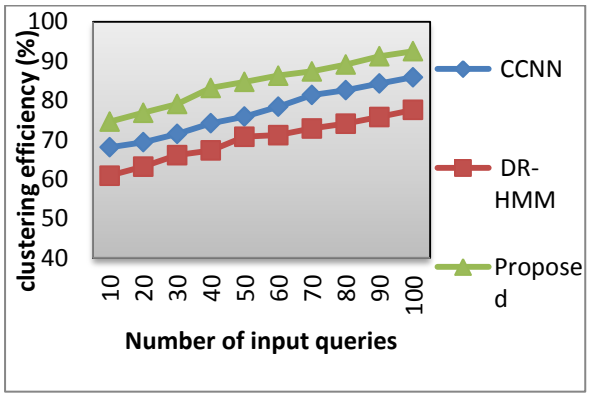

Figure 4.1 Measure ofclustering efficiency

Figure 4.1 shows clustering efficiency with respect to different number input queries using proposed method and existing methods such as CCNN and DR-HMM. From figure 4.1, it is clear that clustering efficiency is improved using proposed method when compared to other existing works. This efficient improvement in clustering efficiency in the proposed method is achieved by utilizing document collection and preprocessing where the stop words are removed earlier. In addition, adaptive K-Means clustering algorithm is employed to achieve better clustering results for input text documents. Therefore clustering efficiencyof proposed method is increased by 9% when compared to existing CCNN method and 17% when compared to existing DR-HMM method respectively.

5. CONCLUSION AND FUTURE WORK

The proposed adaptive K-Means algorithm gives efficient clustering of unstructured documents. Small clusters gives better search results while accessing information from large text corpus. In this work the clustering results were compared with the existing methods namely CCNN and DR-HMM. The proposed methods increases overall 13% improved clustering results when compared to existing methods. The future work may be ontology based clustering and classification gives better search results. The various K- Means distance measures can be implemented for further process.

6. REFERENCES

[1]. Balasubramaniam K 2015, 'Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web',
Elsevier, Procedia Computer Science, vol. 50, pp. 135-142.

[2]. Chyi-KweiYau, Alan Porter, Nils Newman, ArhoSuominen, "Clustering scientific documents with topic modeling", Scientometrics, Springer, Volume 100, Issue 3, September 2014, Pages 767–786.

[3]. Haoji Hu, Kai Zheng, Xiaoling Wang, and Aoying Zhou, "GFilter: A General Gram Filter for String Similarity Search", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 4, August 2014, Pages 1005 – 1018.

[4].Irwan Bastian, Rozaliyana and MettyMustikasari, "Web Document Clustering System Using K – Means Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 8, August 2016, Pages 181-186.

[5]. Long Jun, Wang Lu-da, Li Zu-de, Zhang Zu-ping & Yang Liu 2015, 'WordNet-based lexical semantic classification for text corpus analysis', Springer, vol. 22, no. 5, pp. 1833-1840.

[6]. Mario Beauchemin, "A density-based similarity matrix construction for spectral clustering",Neurocomputing, Elsevier, Volume 151, March 2015, Pages 835–844.

[7]. Peng Wang, Bo Xu, JiamingXu, GuanhuaTian, Cheng-Lin Liu and HongweiHao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification", Elsevier, Neuro computing, Volume 174, January 2016, Pages 806–814.

[8]. Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Volume 25, Issue 1, January 2013, Pages 1 – 14.

[9]. Seara Vieira A, L. Borrajo and E.L. Iglesias, "Improving the text classification using clustering and a novel HMM to reduce the dimensionality", Computer Methods and Programs in Biomedicine, Volume 136, November 2016, Pages 119–130.

[10]. Shikui Wei, Xinxiao Wu, and Dong Xu, "Partitioned K-Means Clustering for Fast Construction of Unbiased Visual Vocabulary", the Era of Interactive Media, Springer, August 2012, Pages 483-493.

[11]. Sumayia Al-Anazi, Hind AlMahmoud, Isra Al-Turaiki, "Finding similar documents using different clustering techniques",ProcediaComputer Science, Volume 82, 2016, Pages 28 – 34.

[12]. TanmayBasu and C.A. Murthy, "A similarity assessment technique for effective grouping of documents", Information Sciences, Elsevier, Volume 311, 2015, Pages 149–162.

[13]. Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou &XianyuBao 2015, 'A semantic approach for text clustering using WordNet and lexical chains', Elsevier, Expert Systems with Applications, vol. 42, no. 4, pp. 2264-2275.

[14]. Umamaheswari R &Shanthi N 2015, 'An Efficient Hybrid Information Retrieval Approach for Unstructured Document Classification', International Journal of Applied Engineering Research vol. 10, no. 24, pp. 44504-44508.

[15]. Yanjun Li, CongnanLuo and Soon M. Chung, "A parallel text document clustering algorithm based on neighbors", Cluster Computing, Springer, Volume 18, Issue 2, June 2015, Pages 933–948.