Journal of Theoretical and Applied Information Technology
31st May 2017. Vol.95. No 10
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMMON SENSE BASED TEXT DOCUMENT CLUSTERING ALGORITHM BY COARSE AND FINE GRAINED CLUSTERING TECHNIQUES

[1]G. LOSHMA, [2]DR. NAGARATNA P HEDGE

[1]Research Scholar, Jawaharlal Nehru Technogical University, India

[2]Professor, Vasavi Collge of Engineering, Hyderabad, India

E-mail: [1]loshmagunisetti15@gmail.com

ABSTRACT

Text documents occupy the major source of data and hence it is important to keep the data in an organized fashion. Clustering is one of the ways for data organization, which tends to group similar documents together. In spite of the presence of numerous existing clustering algorithms, still there is an emergent need for accurate clustering algorithms. Additionally, most of the clustering algorithms work by distance based measures, which is the reason for lack of accuracy. In order to overcome these issues, this work presents a double layered text document clustering algorithm. The entire system is categorized into phases such as document pre-processing, representation, clustering and cluster labelling. The document pre-processing phase prepares the document in such a way that it is suitable for the forthcoming processes. The document representation phase is to standardize the structure of the document and this is done by Document Index Graph (DIG) model. The documents are then clustered by cosine similarity and rough set of clusters are formed. The second level of cluster refinement is achieved by ConceptNet, which works on the basis of common sense reasoning. Finally, the clusters are labelled by picking the top ranked key-phrase. This work is tested over BBCSport and 20 NewsGroup dataset and the proposed approach proves better results in terms of F-measure, purity and entropy.

Keywords: *Document clustering, DIG model, Sense based clustering, Distance based clustering.*

1. INTRODUCTION

Today's world thrives with data and the growth of data is aggressive. IBM has estimated that over 2.5 quintillion bytes of data are generated everyday and claimed that the today's world data is generated in the past couple of years [1]. Globally, the term data can refer to audio, video, image and text. Among all these data, the text data usage is in practise prevalently, as the memory consumption is minimal. Besides this, no special requirements are needed for generating the text data. The growth of text data is magical and thus effective text data management scheme is necessary for coping up with the voluminous data. The data has to be organized such that the data search and retrieval can be made easier.

The utilization of the data is better when the available data is properly organized with metadata. Data clustering is one of the popular strategies for organized data storage. The main goal of a data clustering scheme is to group similar data together, by measuring the similarity ratio. Text document clustering is a separate research area, which intends to group similar documents together on the go. Some of the noteworthy applications of text document clustering are data search and retrieval and building document taxonomy. Data clustering improves the usability and accessibility of the user, as the similar documents are grouped together.

A standard document clustering algorithm consists of four significant phases and they are document representation, similarity computation, document grouping based on similarity measure and cluster labelling. The document representation phase is to standardize and to impose uniformity over the documents being involved in the clustering process. The similarity computation phase intends to calculate the level of similarity between the documents. The related documents are grouped by taking the level of similarity into account. The main challenges faced by document clustering algorithms are clustering accuracy and speed. Taking these challenges into account, this article intends to address them.

This research article aims to present an effective document clustering algorithm, which is

comprised of Document Index Graph (DIG) based document representation, similarity computation, document clustering and labelling. The DIG based document representation follows the principle of graph theory, in which each word of the document acts as a node and all the words are linked together. The flow of the document is preserved and thus it is easy to proceed further.

The next step is to cluster the documents, which is achieved by two levels of clustering, such as coarse and fine clustering. The coarse-grained clustering process is carried out by different distance based similarity measures, so as to compare the potentiality of different measures. In order to sharpen the clustering algorithm, the fine-grained clustering employs 'ConceptNet' which takes the sense of the word into account. Based on the detected sense, the clusters are refined further. This step improves the accuracy of the clustering process. Finally, this work gives labels to the clusters not by keywords but by key-phrase. The cluster labelling step is equally given importance, as it plays a vital role in determining the nature of the cluster. Some of the merits of this work are listed below.

- The documents are represented by DIG model, which preserves the flow of the document and makes the forthcoming processes easier.
- Two levels of clustering such as coarse and fine grained clustering are enforced to attain accurate clusters.
- The coarse level of clustering is achieved by distance based similarity measures such as Euclidean, Pearson, Jaccard and Cosine.
- The so formed clusters are refined by the fine-grained clustering, which is carried out by ConceptNet.
- ConceptNet perfectly suits text data clustering, owing to the ability of contextual reasoning.
- This work labels the clusters with the key-phrase, which improves the readability of the clusters.
- The performances of distance based similarity measures are also analysed.
- The proposed work shows accurate results, owing to the employment of two levels of clustering.

The rest of the article is organized as follows. Section 2 presents the recent review of literature with respect to text document clustering and the motivation of the proposed approach. The proposed approach along with the overview is elaborated in section 3. The performance of the proposed system is analysed in section 4. Finally, the concluding remarks are presented in section 5.

2. BACKGROUND

This section reviews the recent literature with respect to ConceptNet and semantics based text document clustering.

In spite of the presence of several semantic knowledges, commonsense knowledge is the most familiar kind of knowledge. As per definition, the common sense knowledge involves the idea about the spatial, physical, social and psychological facets. The major objective of ConceptNet is to impart common sense to the machine, as humans do. ConceptNet is simple to use and it takes the semantic relations between words into account. ConceptNet may sound similar to WordNet however, WordNet focuses on literal meaning and the ConceptNet focuses on the sense of the word. The semantic relationship between words are represented by links such as is-a, part-of, effect-of, capable-of, location-of, property-of and so on.

On this front, ConceptNet is capable of predicting the emotion of the story/comment, recognizing the correct meaning of the word irrespective of multiple meanings. For instance, the word 'break' has several meanings such as fracture, time interval, separation, opportunity, to convey and so on. The ConceptNet catches up the exact meaning of the word, even though the word has got multiple meanings. This is achieved by the contextual reasoning ability of the ConceptNet. Owing to the numerous advantages of ConceptNet, this work employs ConceptNet to recognize the correct sense of the word [2]. ConceptNet is the largest common sense source with almost two lakh and fifty thousand relations.

Text clustering algorithms based on distance measures compute the degree of closeness of documents. Though there are numerous distance measures, cosine similarity distance is popular for text document clustering. Most of the text clustering algorithms based on distance measures, compute the relationship between the query and the documents rather than the relationship between documents. Finally, the obtained results are ranked on the basis of the similarity level.

Semantic based clustering is employed in several applications such as text clustering, summarization, data search and retrieval. The semantic similarity measures are classified into four classes, which are based on path length,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information, features and hybrid of the already mentioned classes. In [3], the authors improved the performance of the data retrieval by incorporating MeSH ontology. The semantic relationship is computed by taking the edge count between the words into account. Later on, a similarity measure is proposed which takes the path length into consideration [4]. Another measure is proposed in [5], which computes the total count of links between two different terms that emphasize on the same concept. The enhanced version of [5] is presented in [6], which additionally includes Roget's thesaurus into the measure.

In [7], a semantics based document clustering algorithm is presented. This work computes the semantic relationship between the terms of documents, which is followed by the computation of cosine similarity. The semantic analysis is achieved by WordNet. The work presented in [8], presents a text clustering algorithm based on WordNet and Artificial Bee Colony (ABC) algorithm. A text clustering algorithm which incorporates Ant Colony Optimization (ACO) algorithm and WordNet is presented in [9].

Motivated by the above works and concepts, this work strives to present an accurate Double Layered Text Document Clustering Algorithm (DLTDCA) which is based on semantics. To the best of our knowledge, this work is the first to employ two layers for text document clustering for improved performance. The forthcoming section intends to elaborate the proposed text document clustering algorithm.

3. PROPOSED TEXT DOCUMENT CLUSTERING ALGORITHM (DLTDCA)

This section elaborates the proposed approach along with the overview of the entire work.

3.1 Overview of the Work

Text document clustering is always an evergreen research area, because of the skyrocketing growth of the textual data. The available data is beneficial only when it is organized in a proper fashion. The role of text document clustering comes into picture, at this juncture. However, clustering huge volume of data in an efficient way is highly challenging. There are several means to achieve document clustering. Traditionally, the document clustering process is done by following several approaches such as partitional, hierarchical, probability, density based clustering and so on. However, these clustering operations are blind and thus, the accuracy cannot be expected for all the test cases. Today's world operates on artificial intelligence and thus, a reasonable accuracy rate has to be achieved by the clustering algorithm. The following figure 1 depicts the overall flow of the proposed work.

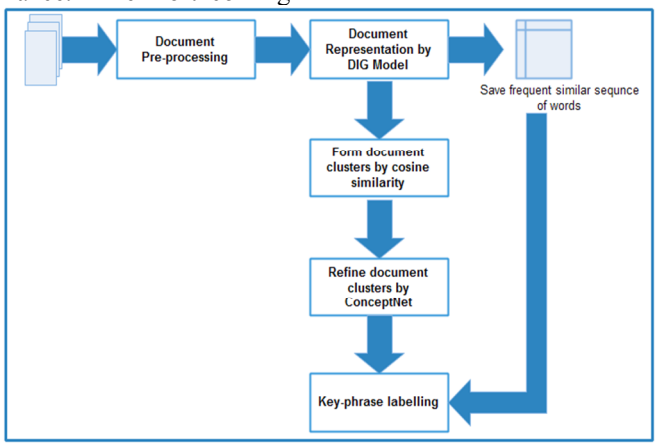

Figure 1:Overall flow of DLTDCA

Taking this as a challenge, this paper presents a text document clustering algorithm, which takes the semantic relationship between the terms into account. In order to achieve reasonable accuracy rate, this work employs double layered text document clustering algorithm, which involves coarse and fine grained clustering. The coarse grained clustering is achieved by distance based similarity measures. The obtained clusters are then refined by ConceptNet in the fine grained clustering phase. In an overall v iew, this work is segregated into data pre-processing,

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

representation, application of clustering algorithm and cluster labelling.

The data pre-processing step aims to weed out the unnecessary components of a document. Usually, the stop and stem words are removed from the documents. This is to speed up the process and to save memory. After the completion of pre-processing step, the documents are standardized such that uniformity can be achieved. This work represents the documents by DIG model. The double layered clustering process is the next step, in which the clusters formed by the coarse grained clustering scheme are refined by the fine grained clustering scheme. Finally, all the clusters are given suitable label with key phrase. Usually, the cluster labels are formed by single keyword however, this work chooses key phrase for enhanced readability. The following subsections explain all the sub-phases of the proposed approach.

3.2 Text Document Pre-Processing

In the text document pre-processing phase, the stop and stem words being present in the documents are intended to be removed. Stop words are the words that do not have meaning on their own. These stop words can be preprositions, articles, pronouns and so on. It is better to filter out these stop words for improved performance. All the search engines remove the stop words, before proceeding further. The removal of stop words conserve both time and space. The stop words are removed from the input query and the document library as well. Some of the sample stop words are presented in table 1.

Table 1: Sample list of stop words

Sample list of stop words			
Each	In	This	Because
Is	He	And	Been
Of	She	Or	Before
A	It	Us	Being
An	Her	About	Between
Was	By	Above	Both
The	Below	Again	But
With	For	Against	By
After	Can	All	Could
On	Be	Am	Cannot
At	To	Any	Did
His	As	Are	During
From	Further	Have	Has
Had	Into	More	Most
Than	That	There	Those
This	Then	Under	Until

These stop words do not influence over the meaning of the text. Besides this, these stop words are repeatedly present in the text. Thus, it is pointless to process these stop words and the stop words are removed in the pre-processing step. For the same concern of saving the time and space, the words in the document are clipped. This means that the tense of the statement is not a matter for clustering and thus, the words that represent the past, continuous, plural extensions (ed, ing, s, es) are removed. The goal of stemming is to obtain the root word, which is enough for performing clustering operation and it is achieved by Porters Stemmers algorithm [10]. Hence, the pre-processing step removed the stop and stem words.

3.3 Text Document Representation by DIG Model

DIG model follows the principle of graph theory, which contains the nodes and edges. This work prefers to utilize this model for document representation, as the flow of the document is preserved. The unique terms of the documents act as the nodes and the edges are the links between the unique terms and thus it forms a directed graph. The directed graph is represented as

$$Gr = (V, E) \quad (1)$$

$$V_k = \{v_{k1}, v_{k2}, v_{k3}, \dots, v_{kn}\} \quad (2)$$

$$E_k = \{e_{k1}, e_{k2}, e_{k3}, \dots, e_{kn}\} \quad (3)$$

where V and E denote the vertices and edges of document k respectively. $\{v_{k1}, v_{k2}, v_{k3}, \dots, v_{kn}\}$ represent the unique terms of document k. Every edge $\{e_{k1}, e_{k2}, e_{k3}, \dots, e_{kn}\}$ links two unique words of the document. By this way, the vertices and edges are constructed for all the documents. The basic idea of DIG model is to construct a graph for every document and connect all the graphs together, such that each document appears as a sub-graph. The process of document representation is depicted in figure 2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

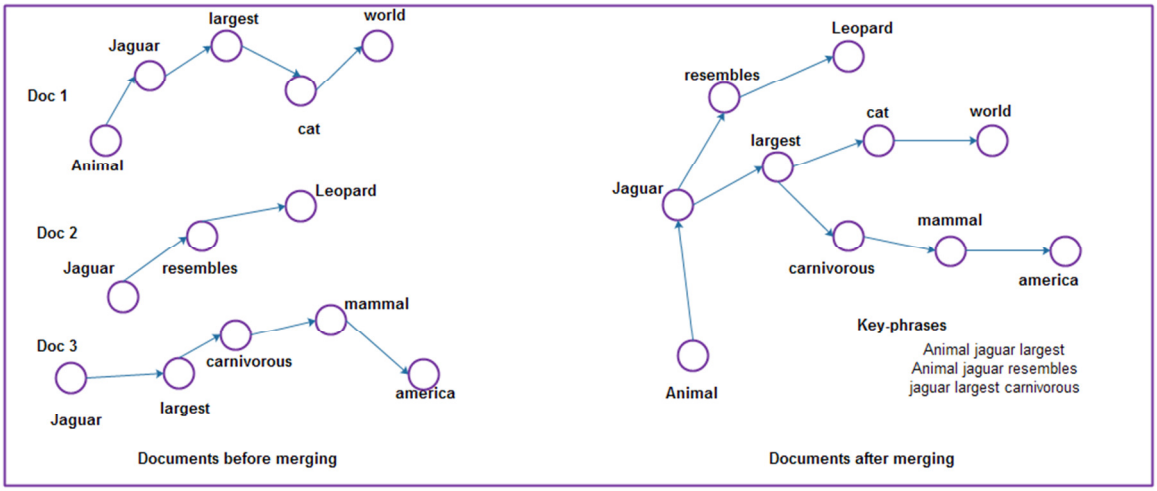

Figure 2:Document Representation Model

For instance, each document doc_i is connected to the graph GR_i and the complete graph contains many sub-graphs (gr_i). Each sub-graph represents an individual document. Hence, the DIG model is meant to meld all the documents together in incremental fashion and it is represented as follows.

$$GR_i = GR_{i-1} \cup gr_i \quad (4)$$

Consider the documents doc_a, doc_b; the similar sequence of words are found by the intersection operation. The intersection is carried out over the sub-graphs of doc_a, doc_b, which are gr_a, gr_b and is represented by

$$SSoW_{a,b} = gr_a \cap gr_b \quad (5)$$

By following the same rule, it is easy to track the repeated similar sequence of words by comparing the new and the existing documents and is represented in the following equation.

$$SSoW_i = gr_i \cap GR_{i-1} \quad (6)$$

Hence, this way of document representation supports in extracting the similar sequences of words from all the documents. This phase forms the platform and acts as the backbone for the forthcoming clustering process.

3.4 Double Layered Document Clustering

This work promotes double layered text document clustering, in which the initial clustering operation is done by distance based similarity measure and the formed clusters are refined by the semantics based clustering done by ConceptNet. As the clustering process deals with the DIG model, it is easy to measure the distance and also the semantic relationship. The overall algorithm is presented below.

DLTDCA – Algorithm

Input: Set of documents (1,2,..n) ;
Output: Document clusters with label;
Begin
For (i=1 to n)
Do
// Document Pre-processing
Remove stop words;
Remove stemming words by Porter-Stemmer algorithm;
Save the data;
// Document Representation
Represent document by DIG model;
Find similar sequence of words by eqn.6;
Store the data;
//Document Clustering
Compute distance between the sub-graphs;
Form the document clusters;
Refine the clusters by ConceptNet;
//Cluster labelling
Check the database for similar sequence of words;
Find the frequently occurring key-phrase;
Declare it as the cluster label;
End;

After successful document representation, the documents with high degree of intersection are stored in database for local reference. The proposed double layered text document clustering algorithm is found to be accurate, owing to the incorporation of DIG, distance based similarity and ConceptNet. The following explains the initial clustering method.

3.4.1 Distance based Clustering

In this initial level of clustering, the distance between the sub-graphs is found out. The distance

based clustering gives first priority to the high degree intersection documents. After this, the distance between all the documents is computed. However, in case of maximum high degree intersection documents (say half the amount of all documents), then the computation of distance among all documents is dropped. This idea saves time and computational power as well.

The most common distance based similarity measure for text documents is the cosine similarity measure. However, this work tries with different distance measures such as Euclidean, Jaccard and Pearson. The performance of the distance measures is discussed in the experimental analysis section. The cosine similarity computation between two different documents is presented as follows.

$$\cos(p,q) = \frac{p \cdot q}{||p||\,||q||} \quad (7)$$

In the above equation, $p \cdot q$ represents the vector dot product and $||p||\,||q||$ is the length of the vector. The similarity level of documents increases as the distance between the documents decreases. Thus, the distance and the similarity between the documents are inversely proportional to each other. The value of cosine similarity falls between 0 and 1. By this way, the cosine similarity is computed for all the sub-graphs of the complete graph. Thus, the initial layer of clustering is over and the next subsection discusses about the semantics based clustering.

3.4.2 Semantics based Clustering by ConceptNet

The initially formed clusters are refined by this layer of clustering. This kind of clustering is more powerful, as it takes the semantics into account. This work employs ConceptNet, as it is superior than the WordNet. The seeds of WordNet are sown in the year 1985. The WordNet is a repository of nouns, verbs and adjectives, which are organized by discrete sense and little semantic relationship is imposed [11]. Mostly, the semantic relationship is in the form of is-a. The basic reason for the success of WordNet is its simplicity and usability.

WordNet is developed by knowledge engineers, whereas ConceptNet is developed by Open Mind Common Sense (OMCS) corpus. The OMCS submitted a help request to the public. Hence, the common sense knowledge is imparted from several individuals. One of the members of OMCS created a web page and declared it as World Wide Web (WWW) based collaborative project. Numerous volunteers logged in and entered uncountable sentences. These sentences are processed and extraction rules are applied to create the semantic knowledge base.

The ConceptNet is known for its contextual commonsense reasoning and it considers a rich set of semantic relations such as Is-a, Part-of, Property-of, Capable-of, Location-of, Subevent-of, Made-of, Used-for and so on. ConceptNet possesses far more knowledge, when compared to WordNet. This strong reason made this research to choose ConceptNet in the place of WordNet. This work employs ConceptNet version 5, which is the most recent one and is downloaded from [12].

The ConceptNet is employed on the already formed clusters. The sense of documents is learnt by the clustering algorithm and the clusters are again refined. As ConceptNet comprise the real world knowledge, the clusters are refined perfectly. To exemplify the concept, the distance based clustering roughly clusters all the documents, whichever possesses the term Jaguar. However, ConceptNet refines the formed clusters by creating different clusters. For instance, one cluster represents the animal Jaguar and the other cluster represents the luxury car. Thus, a single cluster is splitted into furthermore clusters, which is performed by sense based clustering. Finally, the clusters are labelled with the keyphrase, which is presented in the next section.

3.5 Cluster Labelling

This phase intends to label the clusters with a meaningful phrase, being extracted from the cluster. Initially, the frequently occurring phrases of words are stored in database for local reference. These stored phrases of words play a vital role in cluster labelling. The frequently occurring phrases in the documents are organized as candidate phrases for the process of labelling. From the candidate phrases, the most frequently occurring phrase is picked as the key phrase.

The reason for utilizing key phrase instead of keyword is that employment of key-phrase is more meaningful than the keyword. For example, when a set of documents contains data about India and consider that the term India is repeated multiple times. Consider that most of the documents highlight the available natural resources in India. In such situation, a keyword based cluster labelling process select India, as its label. The keyword based cluster labelling just chooses a single key-term, from which the user cannot infer anything. On the other hand, consider the proposed work that employs key-phrase based

labelling. The label of the proposed work is 'India natural resources'. The key phrase is more meaningful and it describes the content of the cluster effectively. Moreover, finding the key phrase is simple, as this work follows a graph based structure.

4. RESULTS AND DISCUSSION

This section analyses the performance of the proposed text document clustering algorithms by utilizing several performance measures. The datasets being exploited for testing the performance of the proposed work are 20 NewsGroups [13] and BBCSport [14]. The 20 NewsGroups dataset contain 18,828 documents in 20 different categories. The BBCSport dataset contains 737 documents which are related to sports. The total class labels of this dataset are 5. The performance of the proposed algorithm is tested by creating multiple scenarios. For this purpose, standard performance metrics such as F-measure, purity and entropy are employed. Initially, the experiments are carried out to emphasize the importance of semantic analysis. In order to prove this, the experiments are executed with ConceptNet, WordNet, distance based similarity measures and the proposed DLTDCA. Similarly, the distance measures are varied to highlight the best performing measure. The distance based similarity measures being employed in this work are Euclidean, Jaccard, Pearson and cosine similarity measure. The performance metrics are explained below.

F-Measure: F-measure depends on the precision and recall values. The precision and recall values of the cluster (cl_j) by taking the class topic (ct_i) into account are presented below.

$$\mathrm{PR}(ct_i, cl_j) = \frac{N_{ct_i cl_j}}{N_{cl_j}} \quad (8)$$

$$\mathrm{RE}(ct_i, cl_j) = \frac{N_{ct_i cl_j}}{N_{ct_i}} \quad (9)$$

In the above equations, $N_{ct_i cl_j}$ is the count of documents of class ct_i which are in cluster cl_j. N_{cl_j} is the total count of clusters and N_{ct_i} is the total entities of class ct_i. The F measure is computed as follows.

$$FM(c_{t_i}) = \frac{2PR\,RE}{PR+RE} \quad (10)$$

On the whole, the F-measure of the clusters is computed by taking the weighted average of the F-measure of every class ct_i.

$$FM_c = \frac{\sum_i N_{ct_i} \times FM(c_{t_i})}{\sum_i N_{ct_i}} \quad (11)$$

Where N_{ct_i} is the total count of entities in class ct_i. The maximum F-measure improves the accuracy of the algorithm.

Purity: This performance metric measures the cluster coherence. The purity of the clustering algorithm for a cluster cl_i with size cn_i is measured as follows.

$$Pur(cl_i) \frac{1}{cn_i} \max_h cn_i^h \quad (12)$$

Where $\max_h cn_i^h$ is the total count of documents which belong to a primary category in the cluster cn_i and cn_i^h is the total count of documents of cluster cn_i, which are allotted to the category h. For instance, if the purity value of the cluster is claimed as 1, then the cluster contains documents that belong to a single category. Hence, the purity value is directly proportional to the quality of the cluster.

Entropy: Entropy is the next performance metric that measures the overall allocation of categories in a cluster. The entropy of a cluster cl_i with size cn_i is computed by

$$En(cl_i) = -\frac{1}{log\ ctg} \sum_{h=1}^{v} \frac{cn_i^h}{cn_i} log\left(\frac{cn_i^h}{cn_i}\right) \quad (13)$$

Where ctg is the total count of categories being present in the dataset. cn_i^h is the total count of documents from the h^{th} class which are allotted to the cluster cl_i. The entropy measure computes the overall allocation of categories in a cluster. The entropy value of a cluster with all the documents that belong to a specific category is 0. Hence, the quality of clusters is improved, as the entropy value decreases. The overall entropy value that considers all the formed clusters is given by

$$En = \sum_{i=1}^{v} \frac{cn_i}{n} En(cl_i) \quad (14)$$

The experimental results of the proposed work are compared with some analogous techniques and the results are presented through figures 3 to 6. We analyse the performance by varying the similarity measure and clustering technique, with respect to BBCSport and 20NewsGroups, in terms of F-measure, purity and entropy.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

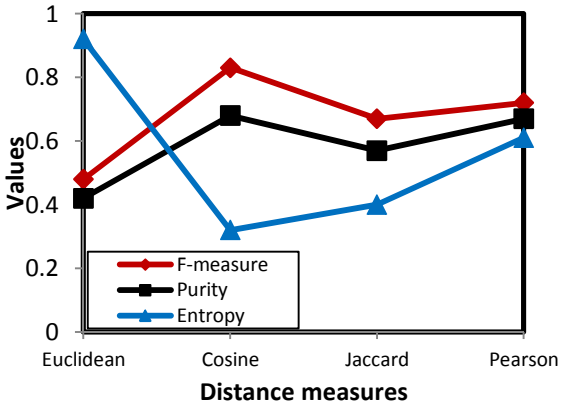

Figure 3:Analysis by varying distance measures for BBCSport

The above presented graph depicts the effectiveness of clustering results over BBCSport dataset. We perform the clustering operation by varying the distance measures such as Euclidean, Cosine, Jaccord and Pearson. The F-measure, purity and entropy of the results being presented by these distances are measured. From the experimental results, we conclude that the performance of cosine distance is satisfactory, as it yields maximum F-measure and purity along with minimal entropy. The forthcoming graph (fig.4) illustrates the performance of the same similarity measures with 20 NewsGroups dataset.

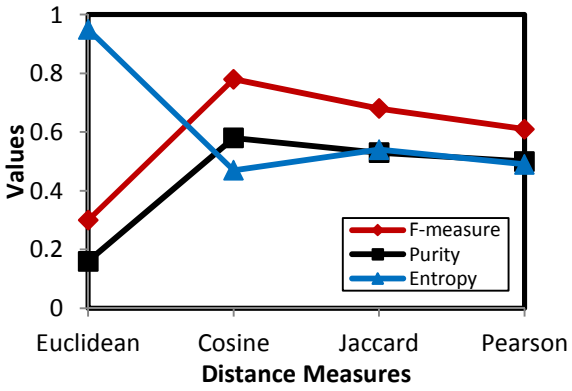

Figure 4:Analysis by varying distance measures for 20NewsGroup

Cosine similarity measure performs well when compared to Euclidean, Jaccard and Pearson even with respect to the 20 NewsGroup dataset also. The cosine similarity measure registers its performance with the F-measure, purity as 0.78 and 0.58 respectively. Similarly, the entropy of cosine similarity is the least, when compared to all the analogous measures and the value is 0.47. Thus, the cosine similarity is proved to be the best and thus, this work intends to club cosine similarity measure with the ConceptNet. The following figures illustrate the performance of the clustering techniques such as ConceptNet, WordNet, Cosine similarity and the combination of ConceptNet and cosine similarity.

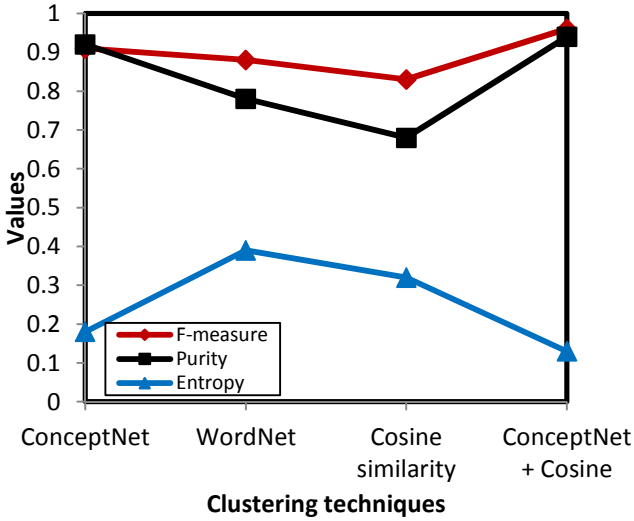

Figure 5:Analysis by varying clustering technique w.r.t BBCSport

The above presented figure presents the experimental results of several clustering techniques, in order to check the level of competency. The data present in the BBCSport are organized under atheletics, cricket, football, rugby and tennis. As ConceptNet works on senses, it easily relates documents and yield better results. WordNet renders limited semantics when compared to ConceptNet however, the results are comparable with the ConceptNet. When the ConceptNet is clubbed with cosine similarity, the performance of the clustering technique is far better. Here, the cosine similarity measure act as the first line of clustering and is followed by the ConceptNet. The combination of ConceptNet and cosine similarity shows high F-measure and purity with lesser entropy rates. Again, the same techniques are applied over 20 NewsGroups and the results are presented in fig 6.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

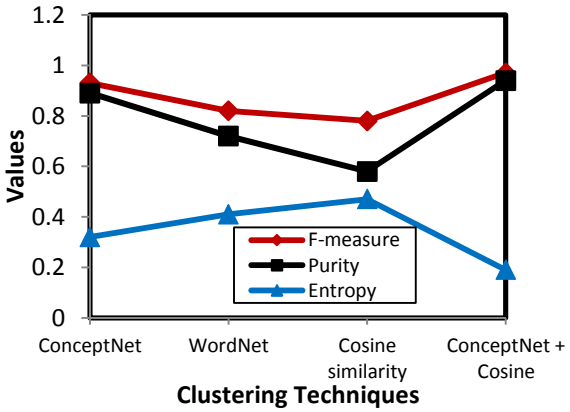

Figure 6:Analysis by varying clustering technique w.r.t 20NewsGroup

Again, the results are found to be convincing. Initially, the ConceptNet alone is implemented and the results are noted and for proving the potential of ConceptNet, the results are compared with the WordNet. Even though ConceptNet has generated good results, we felt eager to combine it with the cosine similarity, which is the root of the proposed work. ConceptNet itself yields better results and performs amazingly when it is combined with the cosine similarity. Thus, the double layered text document clustering is performed.

The double layered clustering technique generates outstanding results with the greatest F-measure and purity along with the least entropy. This proves the quality of clustering. The main underlying reasons for this result starts from the document representation itself. The DIG model represents the document in graphical form, which makes it easier for distance measurement between the documents. Besides this, the ConceptNet can easily figure out its semantic relationship between the documents. All these factors help out in achieving better results. Thus, the main goal of this research is attained by presenting an accurate double layered text document clustering algorithm.

5. CONCLUSION

This article presents a novel accurate double layered text document clustering algorithm. A double layered clustering algorithm is proposed, in which the first line of clustering is achieved by cosine similarity measure and the second line of clustering is done by ConceptNet. ConceptNet is popular for its contextual common sense and it has multiple elements, which can determine the relationship between different documents easily. This common sense based clustering boosts up the clustering accuracy. Additionally, the ConceptNet refines the initially formed clusters, which is done by cosine similarity measure. This work utilizes two different datasets such as BBCSport and 20 NewsGroup for proving the performance of the proposed approach. The experimental analysis is done by comparing the proposed algorithm with other analogous techniques. The results are observed to be satisfactory, in terms of F-measure, purity and entropy. In future, we plan to focus on reducing the time complexity of the algorithm.

REFERENCES

[1] https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[2] H Liu, P Singh, "ConceptNet — a practical commonsense reasoning tool-kit", BT Technology Journal, Vol 22, No 4, pp: 211-226, 2004.

[3] [3] Rada, R., Mili, H., Bicknell, E., & Blettner, M., "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, No. 1, pp. 17–30, 1989.

[4] Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In Paper presented at the proceedings of the 32nd annual meeting on association for computational linguistics.

[5] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database, 49(2), 265–283.

[6] Jarmasz, M., & Szpakowicz, S. (2012). Roget's thesaurus and semantic similarity. arXiv preprint arXiv:1204.0245.

[7] Loshma G., Dr. Nagaratna P. Hedge, "Semantic analysis based text clustering by the fusion of bisecting k-means and UPGMA algorithm", ARPN Journal of Engineering and Applied Sciences, Vol. 11, N0. 3, pp. 1803-1810, 2016

[8] Loshma G., Dr Nagaratna P Hegde, "SABC: Semantic Analysis based Artificial Bee Colony Algorithm for Effective Text Document Clustering", International Conference on Advances in Sciences, Engineering and Management ICASEM 2015, Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem, 4th-5th

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

December 2015,ISBN-978-2-642-24819-7, pp 80-86

[9] Loshma G., Dr Nagaratna P Hegde, "SAACO: Semantic Analysis based Ant Colony Optimization Algorithm for Efficient Text Document Clustering", International Journal of Recent Advances in Engineering & Technology, Vol.3, No.12, pp.21-26, 2015.

[10] Porter, Martin F. "An algorithm for suffix stripping." Program 14.3 (1980): 130-137.

[11] Fellbaum C (Ed): 'WordNet: An Electronic Lexical Database', MIT Press (1998).

[12] http://conceptnet5.media.mit.edu/

[13] http://qwone.com/~jason/20Newsgroups/

[14] http://mlg.ucd.ie/datasets/bbc.html