Analysis of Similarity Measures with WordNet Based and Enhanced Feature Selection in Text Document Clustering

Venkata Nagaraju Thatha[1], *A. Sudhir Babu*[2], *D. Haritha*[3]
[1]*Research Scholar* [2] *Professor* [3] *Professor*
[1]*Department of CSE, JNTUK -Kakinada*
[2]*Department of CSE, PVPSIT-Vijayawada*
[3]*Department of CSE, JNTUK-Kakinada*

Abstract

In the current scenario, a large amount of text is generated from different resources like digitized libraries and world wide web. With these continuous growing of text, it is necessary to organize the text documents depending on the need of the user. Based on the need text document clustering comes into play. Document clustering partitions the large amount of data into smaller manageable clusters. Traditional approaches use word formsand statistical features to cluster the documents. So, the documents with in a cluster are not conceptually similar to one another. To overcome these problems, we model the document clustering in such a way that documents grouped together based on the similarity of the concepts. For that purpose, Ontology is involved in the document clustering. The proposed model first identifies conferences present in each document. By using Semantic similarity and WordNet the problems of synonymy, polysemy is handled as by identifying suitable meaning of the word based on the condition. In this paper, three most popular similarity measures such as cosine, Jaccard coefficient and Pearson correlation coefficient are compared by using enhanced k- means algorithm in both frequency count and DFS-EM representations of the documents.

Keywords: *similarity, DFS-EM, k-means, ontology, WordNet, semantic similarity.*

1. Introduction

At the earlier stage documents are clustered to improve the performance in information Retrieval Systems [1]. Now a days, Clustering is used for browsing the data or to optimize the search results in search engines [2]. Document clustering is an unsupervised learning. This unsupervised document clustering groups the given documents in to clusters in such a way that the similarity between the documents in the cluster is high when compared to the documents in other clusters. Document clustering is used in many applications like recommendation systems and web search engines etc... Suppose in web search engine whenever a user requests a query it produces a large number of results. The results are displayed in sorted order and they are ranked based on the relevance to the query. Users have to check each result until the relevant result is found and it takes lot of time. To overcome this problem clustering is used. clustering groups the results into meaningful fashion to simply the task of searching. i.e. user have to search a few groups rather than all the results [3].

For clustering of documents in a meaningful way, the semantic relationship between the words in the documents are observed. Words are ambiguous. Consider a word ‘orange’ it is referred as colour in one document and fruit in another document depending on the context. i.e., meaning of a word is different in different documents. This is called polysemy problem. So, the context has to be learned and make into different clusters.

Another situation is different words in the document produces same meaning. Consider a sentence ‘food is tasty’ is same as ‘food is nice’. This is called synonyms problem. Before applying feature selection, the synonyms in the document have to be identified for feature count.

Weights are also assigned for each feature in the document to identify the importance of that feature in clustering. For example, if the context is about tennis and the word tennis appeared only once then that word is ignored because count is only one but it is very important word. Hence before remove the unnecessary words in the document first identify the conferenced words and assign the weights accordingly.

2. Related Work

Many clustering algorithms are developed for clustering of text documents.Mainly clustering algorithms are classified into Partitioning and Hierarchical [4][5]. The procedure of hierarchical clustering is it clusters the data in a tree format. Hierarchical clustering is divided into two types based on clustering is formed in top-down or bottom-up fashion. The two types are divisive and agglomerative clustering.The most used Partitioning clustering is k-means and its variants [6].

Before applying clustering to the documents first identify conference resolution and then perform semantic relationships among the words by handling synonymy and polysemy using semantic similarity and Wordnet. WordNet is the most powerful lexical tool to process text data [7]. WordNet groups the words verbs, nouns, adverbs and adjectives into synsets. The synsets are interrelated by using lexical relations and semantics of the concept. i.e. The synonyms that express same concept are grouped in a synset [8]. The words which are having similar sense are placed in a same synset and the word is placed in different synsets if it has different sense. In wordNet synset is the basic building block. This type of structure is very convenient to predict words similarity [9]. This model uses enhanced feature selection for clustering text documents.

3. Proposed model

In proposed model document clustering is performed in 5 steps, namely conference resolution,pre-processing, sense disambiguation, synonymy identification and feature selection [10].

3.1 Coreference Resolution

If two or more than two sentences in a document refers the same concept then conference occurs.The method of finding conferenced words present in the document is called conference resolution. Consider an example, Raju is very good at studies. At the same time, he plays cricket very well. In this example Raju and he are referred as conferenced words. After applying conference resolution, the sentence changed as Raju is very good at studies. At the same time Raju plays cricket very well.

3.2 Pre-processing Module

After conference resolution the data is to be pre-processed. The performance of clustering is increased by adding Part of speech tag and use of WordNet for synsets. The important technique in pre-processing the documents by using wordNet is to enhance term vectors with their concepts from ontology.WordNet covers both lexical and semantic relations between word meanings and word forms.

3.2.1 POS Tagging: In this by using part of speech tagger identify all the parts of speech such as verb,noun,adverb,adjective etc... present in the document and assign a POS tag to each word [11]. The main advantage of identifying parts of speech for each word is it helps in misusing the information from WordNet properly.

3.2.2 Stop word elimination: The data corpus contains many numbers of documents and each document consists of many number of words. Among these many number of words most of the words does not specify any relation to the context and these types of words are called as stop words [12]. To improve the performance of clustering and to reduce the feature space size these stop words are to be removed. The most commonly used stopwords are 'the','a','an','on','since','for' etc..

3.2.3 Stemming: Stemming is the process of converting more than one word which are having same prefix into their root form. For example, consider the words programming,programmer,programmable etc are stemmed to their root word program. Thus, stemming reduces feature space size by replacing the words with their root form. WordNet contains base forms for all the words. morphology functions in wordnet is used to stem the words into their base form.

3.3 Sense Disambiguation and Synonym Identification

In this step the problems of synonymy and polysemy are solved. Different words which are having same meaning are called synonyms and a word having many meanings is called polysemy. All polysemy's are identified and replaced by their synset IDs. For disambiguation of senses and synonyms identification wordnet is used. In this the sense which has maximum similarity value is chosen as correct sense of the word. If a same word is used with different senses in different places are replaced by different synset ids. i.e. the words are segregated to count the frequency efficiently. The synonyms present in the documents are replaced by using same synset Id hence they are merged for counting the frequency.

3.4 Feature Selection

Document frequency specifies the number of documents in which a term appears. The terms that are below a specified threshold are removed from the feature space. Generally, for unrelated features low scores are assigned and for distinctive features high scores are assigned. In normal conditions terms are ranked by considering the following requirements.

- If a term which appears more frequently in one class and does not present in another class a high score is given
- If a term rarely present in one class but present in another class and it is considered as irrelevant then low score is given to that term.
- If a term present more frequently in all the classes and considered as irrelevant then a low score is assigned
- If a term appears only in some classes and if it is considered as a relatively distinctive then a high score is assigned.

$$DFS=\sum_{i=1}^{N}\frac{p(c_i|\mathrm{t})}{p(\mathrm{t}'|c_i)+p(\mathrm{t}|c_i')+1}$$

Where N specifies number of classes and $p(c_i|t)$ specifies conditional probability of class c_i for presence of a term t and $p(t`|c_i)$ is the conditional probability for absence of a term t given a class $c_i$ and $p(t|c_i`)$ is the conditional probability of a term t given the classes other than c_i

The most important step in clustering is feature selection .so, to improve the performance of feature selection DFS approach is combined with entropy method.DFS-EM is an extension of DFS and Entropy[13]. The measure is shown below

$$\mathrm{DF\,S{=}EM} = \sum_{i=1}^{N} \sum_{\partial \in c_i}^{|c_i|} \mathrm{DF\,S}_i(\mathrm{w}).\mathrm{TF}\ (\mathrm{w}|\ \partial)$$

For skewed datasets TF (w| ∂) produces poor results with regards to term frequency so the classes which are having high term frequency among all the classes, TF (w| ∂) term is combined with DFS. For efficient results the equation is modified as

$$\mathrm{DF\,S{=}EM} = \sum_{i=1}^{N} \sum_{\partial \in c_i}^{|c_i|} \mathrm{DF\,S}_i|(\mathrm{t}) * \mathrm{ImpTF}_i(\mathrm{t})$$

$$\mathrm{ImpTF}_i(\mathrm{t}) = \frac{\mathrm{TF}_i(\mathrm{t}) * \mathrm{ATF}_i(\mathrm{t})}{\mathrm{N}_i}$$

$$\mathrm{ATF}_i(\mathrm{t}) = \frac{\sum_{\partial=1}^{k} \mathrm{t}f_{\mathrm{t},\partial}}{\mathrm{M}_i}$$

$$\mathrm{N}_i = \frac{\sum_{\partial \in c_i} \mathrm{termCount}(\partial)}{\mathrm{D}_i}$$

Where M_i, D_i,N_i denotes normalization factor of each class i, number of dissimilar words and number of documents respectively.TermCount(∂) specifies number of terms present in the document ∂ and ATF_i denotes average frequency of a term in a class.

4. Similarity Measures

Document clustering clusters similar documents into a group with high intra cluster similarity and low inter cluster similarity. Generally, weather to documents are similar or not depending on the problem context. To perform clustering efficiently, a distance or similarity measure has to be identified such that it clusters the similar documents. Many number of distance or similarity measures are proposed. Some of them are cosine similarity, Pearson Correlation Coefficient,Jaccard coefficient [15][16]. The description of each of these similarity methods is shown below.

4.1 Cosine Similarity Measure

The most popular and commonly used similarity measure is cosine similarity. If documents are characterized in term vectors then the cosine angle among two vectors is specified as similarity between the documents. Generally, these types of similarity used in text mining to compare documents.

$$cosine - similarity(\vec{t_x}, \vec{t_y}) = \frac{\vec{t_x} \cdot \vec{t_y}}{|\vec{t_x}||\vec{t_y}|}$$

where t_x and t_y represents vectors of n-dimensions over the term set C = c_1, ..., c_n. Heren specifies dimension space of each coordinate. Each dimension specifies a term present in the document along with weight and is non- negative. Cosine similarity is bounded in the range of [0,1]. The value of cosine similarity is either 1 or 0 depending on the similarity between the documents. If documents are more similar value is 1 and nothing is common between the documents value is 0 otherwise depending on measure a value in the range[0,1] is obtained. The main advantage of using cosine similarity is it does not depend on length of the document.

4.2 Jaccard Coefficient

In Jaccard coefficient similarity measured in terms of ratios. In the ratio the numerator is intersection of objects and denominator is union of objects. Here both in intersection and union we are considering same pair of objects. Jaccard coefficient is also referred as Tanimoto coefficient. For documents that contains text, Jaccard coefficient compares sum of weights of all shared terms in both documents to sum of weights of the terms that occur in any one of the documents but not the common terms.

The formulae representation of Jaccard Coefficient's is shown below

$$Jaccard - similarity(\overrightarrow{t_x}, \overrightarrow{t_y}) = \frac{\overrightarrow{t_x} \cdot \overrightarrow{t_y}}{|\overrightarrow{t_x}^2| + |\overrightarrow{t_y}^2| - \overrightarrow{t_x} \cdot \overrightarrow{t_y}}$$

Jaccard coefficient is bounded in the range[0,1]. If two documents are equal then 1 is similarity value and if two documents are disjoint then 0 is the similarity value.

4.3 Pearson Correlation Coefficient

Pearson correlation coefficient is used to measure the correlation between two variables and relation among a pair of vectors. Pearson Correlation Coefficient is measured as ratio of covariance between two variables to product of standard deviation between the two variables.

$$PCC - similarity(a, b) = \frac{covariance(a, b)}{[stddev(a) \cdot stddev(b)]}$$

$$Pearson - similarity(\overrightarrow{t_x}, \overrightarrow{t_y}) = \frac{n \sum_{j=1}^{n} w_{jx} \times w_{jy} - TC_l \times TC_k}{\sqrt{[n \sum_{j=1}^{n} w_{j,x}^2 - F_x^2][n \sum_{j=1}^{n} w_{j,x}^2 - F_y^2]}}$$

$$F_x = \sum_{j=1}^{n} w_{j,x} \qquad F_y = \sum_{j=1}^{n} w_{j,y}$$

Pearson correlation coefficient ranges from +1 to -1. In positive correlation both the variables are either decrease or increase and the value is +1. In negative correlation if one variable decreases another variable increase or vice versa and the value is -1.If no relation between two variables means that there is no correlation and the value is 0.

The cosine similarity, Jaccard coefficient and Pearson correlation coefficients are similarity measures. Now consider the most popular distance measure, Euclidean distance. By applying simple transformation similarity measures are converted into distance values. In cosine similarity and Jaccard coefficients are bounded to [0,1] and in these two cases distance measure is dis=1-similarity but Pearson correlation coefficient ranges from +1 to -1 ,in this case we consider dis=1-similarity if similarity≥0 and dis=|similarity| otherwise.

5. Clustering Algorithm

K-means is the most popular partitioning algorithm for clustering of text documents. This algorithm forms different clusters with the documents by maintaining intra cluster similarity between the documents. The working of k-means algorithm is, the points are assigned to the closest cluster depending on the centroid. Centroid of the cluster is computed by considering average of all data points with in the cluster. Based on assignment of points to the cluster the quality of clustering changes. By using the calculate similarity measures the data points are assigned to the closet cluster. After completion of a step, again the partitions are recomputed and the data points ae transformed from one cluster to another cluster.

5.1 Enhanced k-means

To improve the speed and accuracy of k-means algorithm, the standard k-means is combined with a special initialization technique. The final clusters mainly depends on choice of initial seed sets. Vassilviskii and Arthur proposed k-means++ algorithm [16] that uses randomized seeding technique.

1. Initially, choose a centroid randomly from data set D in uniform fashion.
2. In the next step select a centroid C_j = d′ ∈D with probability

$$\frac{similarity(d')^2}{\sum_{d\in D} similarity(d)^2}$$

Here similarity (d) is the maximum similarity among the document D and already selected centroids.
3. Step 2 is repeated until k number of centroids are selected.
4. Assign each and every point in the document to its closest cluster.
5. After assigning a point again recompute centroid for each cluster.
6.repeat steps 4 and 5 until no change in the centroids of the clusters.

6. Experimental results

6.1 Data sets

For evaluation of results a popular data set for text documents Classic data set is used and it is retrieved from uci.kdd repository. Classic data set contains a total of 7095 documents and it is categorized into 4 classes as [17],

Class name	Number of documents
computer Science (CACM)	3204
Information Science (CISI)	1460
Aeronautics (CRAN)	1398
Medicine (MED)	1033

The dataset consists of single word documents also so, there is no use to consider such type of documents for evaluation purpose. So, by applying some file reduction technique on each category remove the unwanted documents and only the documents which satisfies average length in each category are returned. File reduction is performed by constructing Boolean matrices for all the documents in category wise and then calculate average length of words in each category and the documents which are not satisfy the average length are removed. Once file reduction is applied the data set contains only valid documents and from these valid documents a total of 800 documents such that 200 from each category are selected for evaluation purpose.

Along with classic data set one more benchmark data set Abstract data set is also used for evaluation of results. The Abstract data set contains abstracts of four different areas such as Image processing,Network security, Data Mining, Natural language processing. For experiment purpose a total of 400 documents, 100 from each category are collected.

6.2 Evaluation measure

Entropy is used to measure the quality of the clusters. In entropy first calculate the distribution of data to a class in each cluster i.e., compute p_{ij} for cluster j specifies that a member in cluster j is belong to class i. By using this concept of class distribution, entropy of cluster j is computed by using the formula

$$Entropy_j = -\sum p_{ij}\log(p_{ij})$$

Here we calculate the sum for all the classes. The entropy of entire data set is obtained by adding entropies of each cluster biased by size of cluster.

$$Entropy_S = \sum_{j=1}^{k} \frac{m_j \times entropy_j}{m}$$

Where k denotes number of clusters, m_j denotes size of a cluster j, m denotes total data points.

6.3 Results Analysis

Enhanced k- means algorithm is used for document clustering. In this we analyse the clusters for both classic and Abstract datasets by using frequency count called term frequency and DFS-EM approach with cosine similarity, Jaccard similarity and person correlation coefficient. The results are shown in the below tables. By observing the table 1 Pearson correlation coefficient yield good results with both frequency count and DFS-EM representations and cosine similarity performs well with DFS-EM representation. By observing the results cosine similarity has NAN values means the clusters are empty. Over all person correlation coefficient and Jaccard measures are better for constructing clusters with intra similarity means the clusters with entropy score is low. Table2 and Table 3 shows the partitions of the data set with each class using both frequency count and DFS-EM for Jaccard coefficient and Pearson correlation coefficient respectively.

Table 1: Entropy results of both frequency count and DFS-EM representation

Classic Data set	Cosine		Jaccard		PCC	
	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM

ISSN: 2005-4238 IJAST

cluster 0	0.376	0.212	0.124	0.045	0.131	0.149
cluster 1	0.285	0.014	0.289	0.267	0.0673	0.082
cluster 2	0.189	0.079	0.0325	0.049	0.0256	0.026
cluster 3	NAN	0.042	0.2515	0.179	0.1812	0.169
Total	NAN	0.101	0.1892	0.152	0.11	0.103

Table 2: Clustering results usingJaccard for both Frequency count and DFS-EM

Classic Data set	CISI		CACM		MED		CRAN	
	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM
cluster 0	18	5	16	16	54	49	193	197
cluster 1	68	22	170	179	11	6	3	1
cluster 2	111	173	6	2	0	0	0	0
cluster 3	3	0	8	3	135	145	4	2

Table 3: Clustering results using PCC for both frequency count and DFS-EMrepresentation

Classic Data set	CISI		CACM		MED		CRAN	
	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM
cluster 0	92	154	184	190	4	2	2	0
cluster 1	5	2	8	7	5	3	193	187
cluster 2	99	40	1	2	0	1	0	6
cluster 3	4	4	7	1	191	194	5	7

Now consider Abstract Dataset. By observing the following table4 Pearson correlation coefficient perform good compare to cosine and Jaccardin both frequency count and DFS-EM representations. By considering total entropy Jaccard coefficient performs better with DFS-EM representation. Over all Jaccard coefficient is better for constructing clusters with high intra similarity means the clusters with entropy score is low. Table 5 and Table 6 shows the partitions of the Abstract data set with each class using both frequency count and DFS-EM with Jaccard coefficient and Pearson correlation coefficient.

Table 4: Entropy results of both frequency count and DFS-EM representation

Abstract Data set	Cosine		Jaccard		PCC	
	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM
cluster 0	0.3654	0.221	0.2621	0.2212	0.384	0.32433
cluster 1	0.0298	0.3211	0.0596	0.1187	0.0296	0.059
cluster 2	0.1956	0.2747	0.2619	0.2645	0.2101	0.2109
cluster 3	0.3591	0.0312	0.1899	0.1793	0.312	0.324
Total	0.2882	0.2278	0.2212	0.1896	0.2474	0.2344

Table 5: Clustering results using Jaccard coefficient for both frequency count and DFS-EM representation

Abstract Dataset	NS		IP		NLP		DM	
	FC	DFS-	FC	DFS-	FC	DFS-	FC	DFS-

		EM		EM		EM		EM
cluster 0	2	2	0	1	3	3	93	94
cluster 1	95	96	2	2	4	4	3	2
cluster 2	2	2	1	1	90	92	3	4
cluster 3	1	0	97	96	3	1	1	0

Table 6: Clustering results using Pearson correlation coefficient for both frequency count and DFS-EM representation

Abstract Dataset	NS		IP		NLP		DM	
	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM	FC	DFS-EM
cluster 0	94	97	2	1	1	2	4	2
cluster 1	2	1	0	0	0	0	82	86
cluster 2	3	2	2	2	97	96	14	12
cluster 3	1	0	96	97	2	2	0	0

The result of the clustering is measured in terms of clustering accuracy, Clusteringaccuracy is measured as

$$ac = \frac{\sum_{j=1}^{4} n_i}{m}$$

Where n_i denotes number of times present in both the cluster j and their corresponding class. N specifies number of instances present in the dataset. The accuracy of clustering is more in DFS-EM representation Jaccard and Pearson correlation coefficient measures. With DFS-EM representation, the classic data set has accuracy is more than 92 percent and for Abstract dataset the accuracy is more than 94 percent.

7. Conclusion and Future work

Document clustering is enhanced by using WordNet. An enhanced feature selection method is applied to the data sets and observe three similarity measures among two data sets with both frequency and DFS-EM representations.

In this analysis we observe that all three similarity methods produces good results with Partitional clustering of documents. Among those three measures, Pearson correlation coefficient is better as the formed clusters are more balanced and is almost similar to the manually constructed clusters. With the usage of WordNet. The clustering produces consistent accuracy of more than 93% with DFS-EM representation. The person and Jaccard coefficients form more coherent clusters. Finally, the DFS-EM representation produces good results with Pearson and Jaccard measures. By observing the cluster analysis there are four factors that effect the results and those are, Document representation, Similarity measure, enhancement of ‘bag of words’ and clustering algorithm. In our future we apply the similarity measures with hierarchicalclustering’s and more exploration of WordNet for document clustering.

References

[1] P. Prabhu, “Document clustering for Information Retrieval – A General Perspective”, Indian streamsresearch journalVol.1, Issue.VIII/Aug 11pp.1-4.

[2] Sumathi Rani Manukonda and Nomula , “Efficient Document Clustering for Web Search Result“, International Journal of Engineering & Technology 7 (3.3)Jan (2018) .

[3] PankajJajoo, "Document clustering ", M.Tech thesis, TIT, Kharagpur, 2008.

[4] Xin Jin1 and Jiawei Han, “Partitional Clustering “, Springer Science +Business Media New York 2016.

[5] Koller D and Sahami M, “Hierarchically classifying documents using very few words.”,14th International Conference on Machine Learning (ML), pp. 170– 178.

[6]Duong TrongHai, Cuong Duc Nguyen, ” K-means** - a fast and efficient K-means algorithms”, International Journal of Intelligent Information and Database Systems 11(1):27 · January 2018.

[7] LubomirStanchev , “Semantic Document Clustering Using Information from WordNet and DBPedia”, 12th IEEE International Conference on Semantic Computing-2018.

[8] Sujata R. Kolhe ,Dr. S. D. Sawarkar, “ A Concept Driven Document Clustering Using WordNet”, International Conference on Nascent Technologies in the Engineering Field IEEE- 2017 .

[9] T. Slimani, "Description and Evaluation of Semantic Similarity Measures Approaches", International Journal of Computer Applications, vol. 80, no. 10, pp. 25-33, 2013.

[10]Sneha S. Desai, “WordNet and Semantic Similarity based Approach for Document Clustering”, International Conference on Computational Systems and Information Systems for Sustainable Solutions.

[11]Michael Lamar, “SVD and clustering for unsupervised POS tagging”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010,.

[12] Venkata Nagaraju Thatha,Dr.A.SudhirBabu,Dr.D Haritha, “Research of Clustering Algorithms using Enhanced Feature Selection”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019.

[13] Venkata Nagaraju Thatha,Dr.A.SudhirBabu,Dr.D Haritha, Text Categorization Using Various Advanced ML Techniques in Text Mining Jour of Adv Research in Dynamical & Control Systems, Vol. 11, Issue-08, 2019.

[14] Kavitha Karun A, Mintu Philip, Lubna K, “Comparative Analysis of Similarity Measures in Document Clustering” 2013 International Conference on Green Computing, Communication and Conservation of Energy December 2013.

[15] Baskar Subramanian, “Efficient Text Document Clustering With New Similarity Measures”, International Journal of Business Intelligence and Data Mining January 2018.

[16] Arthur, D., Vassilvitskii, S., “K-means++ the advantages of careful seeding”,Symposium on Discrete Algorithms (2007).

[17] V. TUNALI, "Classic3 and Classic4 DataSets", Data Mining Research. 20IO. Available: http://www.dataminingresearch.com/index.php/20IOI09/c1assic3-classic4-datasets/. [Accessed: 10-Feb- 2016].