International Journal of Engineering & Technology, 7 (2.9) (2018) 102-105

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A semantic approach for text document clustering using frequent itemsets and WordNet

Harsha Patil [1, 3] *, Ramjeevan Singh Thakur [2, 3]

[1] *Research Scholar, Department of Computer Applications*
[2] *Associate Professor, Department of Computer Applications*
[3] *Maulana Azad National Institute of Technology (MANIT), Bhopal, Madhya Pradesh, India*
**Corresponding author E-mail: Harsha.kun.patil@gmail.com*

Abstract

Document Clustering is an unsupervised method for classified documents in clusters on the basis of their similarity. Any document get it place in any specific cluster, on the basis of membership score, which calculated through membership function. But many of the traditional clustering algorithms are generally based on only BOW (Bag of Words), which ignores the semantic similarity between document and Cluster. In this research we consider the semantic association between cluster and text document during the calculation of membership score of any document for any specific cluster. Several researchers are working on semantic aspects of document clustering to develop clustering performance. Many external knowledge bases like WordNet, Wikipedia, Lucene etc. are utilized for this purpose. The proposed approach exploits WordNet to improve cluster member ship function. The experimental result shows that clustering quality improved significantly by using proposed framework of semantic approach.

Keywords: *Document Clustering; Frequent Item Sets; Semantic; Similarity Measures; WordNet.*

1. Introduction

Now a days to solve any query, search engine is very useful and instant tool. Internet is fastest method to learn, understand and solve any problem or get any information from worldwide knowledge base. But many times when we search for any query we get many irrelevant information with less relevant information with respect to our query. Generally all search engines are using document clustering to display query results in organized and in effective manner. Document clustering is unsupervised methodology which collects relevant document in one group. Grouping of similar document in a group such that documents in a group are more similar then a document belongs to another group. This process is called document clustering. But many of the traditional clustering algorithms are mostly based on only BOW, which ignores the semantic similarity between document and Cluster. Document Clustering handles unstructured text, which has many challenges. Text documents are normally full of abstract concepts, which difficult to represent by using traditional methodology of text mining. Due to lacking of this, traditional document clustering algorithms are not capable to present semantic associations among the words and penalties in less qualitative output.

Use of external knowledge base is being very helpful to develop semantic based approaches for document clustering. WordNet (Miller, 1995) is the most extensively used lexica for English language. In recent research work, WordNet has been broadly used to increase quality of document clustering. WordNet is a lexical knowledge, based on conceptual look up which organize lexical information in terms of word meaning, rather than word form. The use of WordNet in clustering captures the relations between the words and help to identify the precise cluster of the documents. Our experiment results shows that by using proposed framework more pure clusters are generated. The reminder of this paper is organized as per the following: Section 2 present the related works. Section 3 discussed on WordNet. Section 4 shows proposed algorithm's framework, Section 5 presents experiment evaluation and results and finally section 6 provide conclusion and scope for future work

2. Related works

So far, document clustering has been comprehensively explored and many techniques has been suggested to deal with it. Document clustering techniques can be categorised in to three broad classes: Partitioning method [21], Agglomerative and divisive clustering [2] and item set based clustering [5]. Many clustering algorithms, based on partitioning or hierarchical methodology like K-Means [12], [14] and its variants, Hierarchical Agglomerative clustering (HAC) [1], [3], [22], and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [21] performs efficiently for low dimensional data but in case of high dimensional data they results in poor clustering. Frequent item set based algorithms handle the complexity of high dimensionality of text documents by selecting only frequent item sets as features for clustering. Hierarchical Frequent Term based Clustering (HFTC) [4] proposed by Beil F, Ester M, Xu X (2002) did prodigious contribution in this direction. Many researchers [15], [19], [20] had discussed the performance improvisation of algorithms, based on association rule mining. After that Fung, et al proposed Hierarchical Document Clustering using frequent item sets (FIHC) [5] which use association rule mining and provides meaningful labels [11] to the clusters. All the above algorithms are not considered the semantic associations of the words. From last nine-ten years many researchers [9] have been using External knowledge base

for associate meanings with words. WordNet is extensively used by researchers for this purpose. Related works done by researchers in past ten years using WordNet is exhaustively summarized in tabular form:

Table 1: Summarized Details on Review of Document Clustering Algorithms Using WordNet

Paper Title	Clustering Algorithm	Evaluation Parameters
WordNet improves Text Document Clustering[8]	Bisecting K-Means	Purity and Inverse Purity
Document Clustering with Semantic Analysis[25]	Word sense disambiguation method, semantic relatedness measures among senses: senseno method offset method	Entropy F-measure
Exploiting noun phrases and semantic relationships for text document clustering.[27]	detection of noun phrases with the use of WordNet as background knowledge	Purity and entropy
A semantic approach for text clustering using WordNet and lexical chains[7]	Bisecting K means	F-measure, Entropy, Purity
A concept driven document clustering using WordNet [24]	LSI (Latent Semantic Indexing)	F-measure, Entropy, Purity
An Integration of Fuzzy Association Rules and WordNet for Document Clustering[28]	Fuzzy Frequent Itemset-based Document Clustering (F2IDC): fuzzy association rule mining	Overall F-measure
Query based Text Document Clustering using its Hypernymy Relation [29]	K-means	Cluster Accuracy
WordNet-based suffix tree clustering Algorithm[30]	WordNet-based suffix tree clustering algorithm (WNSTC).	F-measures

3. WordNet

WordNet is the extensively acknowledged lexical system developed at Princeton University based on psycholinguistic concepts of human lexica memory. Many of the Natural language processing applications used WordNet for Word Sense disambiguation, find semantic distance between words, Machine Translation, Search engine processing, Plagiarism detection, Sentiment analysis etc. WordNet organized terms into taxonomic hierarchies. WordNet established the lexical or semantical connections between noun, verb, adjective, and adverb form of words, by creating synonyms sets called synsets. One synsets is link with other synsets by using relationships like Hyponym/Hypernym, Meronym/Holonym relationships. Various methods are proposed by researchers for finding semantic similarity between terms. Semantic similarity is confidence score of two words which explains likeness of their meaning. According to Meng et al. [26] Semantic similarity measures [9] can be broadly categorized in four classes: path length based measures, information content based measures, feature based measures, and hybrid measures.

1) Path length based measures: It is based on the length of the path connecting the concepts and the location of the concepts in the taxonomy [18]. It counts edges between concepts. The disadvantage of this method is two pairs with equal length of shortest path will results in same similarity.
2) Information content based measures: It is based on the principle is that if two concepts are sharing more common information that means they are more similar [17].
3) Feature based measures: According to this measure two concepts becomes more similar if they have more common features and less uncommon features [16]. This measure is not work properly if complete feature sets of concepts are not available.
4) Hybrid measure: This measure combines the principles proposed in path length based measures, information content based measures and feature based measure. It also consider the relations like IS-A, Part- of more finding semantic similarity.

4. Framework: semantic maximal frequent term based document clustering (SMFTDC).

Proposed approach consist four steps, namely Pre-processing, Dimension reduction, Initial Cluster Construction and Final Cluster Construction.

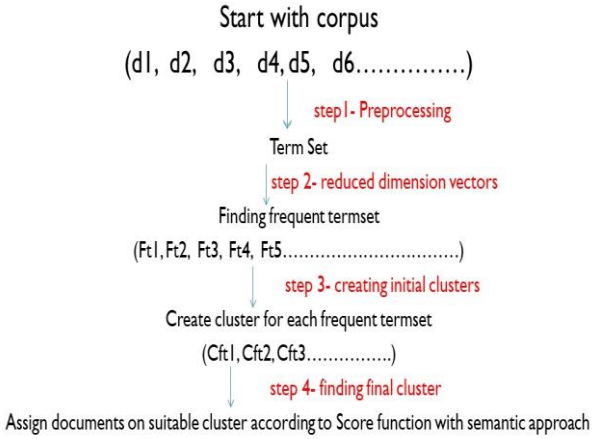

Fig. 1: Flow of Document Clustering.

The whore process of proposed algorithm is explained in following section, before which the projection on important definitions is given below:

Definition 1: *The global frequent itemset is a set of items that seem together in more than a minimum fraction of the corpus.*

Definition 2: *The global support of an itemset is the percentage of documents encompassing the itemset.*

Definition 3: *A global frequent item refers to an item that belongs to some global frequent itemset.*

Definition 4: *A global frequent item is called cluster frequent item for any cluster C_i if the item is present in some minimum fraction of documents in C_i.*

Definition 5: *Maximal Frequent item set is a frequent item set for which none of its immediate supersets are frequent.*

Definition 6: *Hidden Support is a global support of an item, which is not global frequent item, but item is semantically related with global frequent item set of cluster.*

Step 1: Preprocessing
In first step targeted documents are collected, which also knew as text corpus. Each document of corpus is splits into terms are extracted as features. Preprocessing of terms includes, removing of stop words and then stemmed them to their base form. After that weight of each term is computed in reference with its frequency of occurrence in a document. Proposed algorithm used the term frequency – inverse document frequency (tfidf) model for computing weight for the term.

$$d_i = tf_i * idf_i \quad (1)$$

Where tfi is the frequency of the term i in the document and idfi is the inverse frequency of I in the corpus.
Step 2: Dimension reduction Step

In our approach, we first apply Apriori algorithm on all the documents to mine the maximally frequent item sets (MFIs) .Maximal frequent itemsets (MFIs) are the closed frequent itemsets which has no immediate frequent superset. We limit our MFIs min-size to 2. The use of MFIs advances efficiency [10], [13], [23], accuracy and the removal of small size MFIs add further improvement.

Step 3: Constructing Initial Clusters

This step constructs a cluster for each maximal frequent item set. All documents that containing same item set are included in the same cluster. So if frequent item set mining algorithm generates five maximal frequent item sets from document vectors. Then we construct an initial cluster for each maximal frequent item sets i.e. we have five initial clusters. These initial clusters have very much overlapping among them. So this step results in soft clustering.

Step 4: Finding Final Cluster

Final cluster is basically most suitable cluster for any document, so this step outputs hard clustering. In this research we consider the semantic association between cluster and text document during the calculation of membership score of any document for any specific cluster. The score function consist three parts: Rewarding part, Penalty part and Bonus part. Suppose that item x appears in d_j. For calculating score of cluster C_i for any d_j: we reward C_i if x, which is present in d_j is also cluster frequent item for cluster C_i otherwise we penalize C_i. The Bonus part is global support (hidden weight) of the hidden term. So if the hidden term is semantically related with the cluster label than only hidden weight will be given otherwise it will be ignored.

$$Score\ (Ci \rightarrow dj\ ) = \left[\sum_{x} n(x) * cluster\_support\ (x) \right] - \left[\sum_{x'} n(x') * global\_support\ (x') \right] + HS \quad (2)$$

x represents a global frequent item in d_j and also cluster frequent item in Ci

x' represents a global frequent item in d_j but not cluster frequent item in Ci

n (x) is frequency of x in the feature vector of d_j

n (x') is the frequency of x' in the feature vector of d_j

In our experiment we used Wu and Palmer measure for finding the association between two terms. It is a path based method. It calculates the association of two concepts using the lowest common subsumer of two concepts lcs(c1,c2), which is the first shared concept on the paths from the concepts to the root concept of the ontology hierarchy. Using path based method we calculate semantic association between hidden term and cluster labels.

$$sim(c_1,c_2) = \frac{2*depth(lcs)}{l(c_1,lcs) + l(c_2,lcs) + 2*depth(lcs)} \quad (3)$$

Input: A document set D; explicit stop word list
Output: Target Cluster set C

1. Extract the termset T_D ={ t_1 , t_2, t_3…….t_n }
2. Remove all stop words from T_D.
3. Apply stemming for T_D.
   //create document term matrix which represent document in form of d_i={ (t_1,fi_1), (t_2,fi_2), ……..(t_n,fi_n)}
4. Create dtm= DocumentTermMatrix(T_D , method=tfidf)
5. Find Maximal frequent item set
   MFI = apriori(dtm, min_sup, conf)
6. C = assignment (dtm, MFI) //creating initial clusters
7. Find cluster frequent terms for each $C_i \in C$
   C_{ft} = (C_i , min_csup)
8. For each $d_j \in$ D do //finding final cluster
   For each cluster $C_i \in C_{ft}$ do
   Calculate Score($C_i \leftarrow d_j$) using eq(1) given score function
   For each $H_t \in d_j$ do
   Calculate the semantic relation between Ht and MFI of C_i Using eq(2)
   If $\propto$ >0.5 then calculate HS for score function
   Else exit ; End if
9. Assign d_j in final cluster C_{ft} which has maximum Score for membership

Fig. 2: Semantic Maximal Term Based Document Clustering Algorithm.

5. Experimental evaluation

To evaluating the cluster quality, we used F-Score. The F-Score values are in the range [0..1] and largest F-score value indicate higher cluster quality. We compare F-Score value of our algorithm with other algorithm i.e. FIHC and TDC. Classic4, Reuters WAP, Hitech and Re0 datasets were used for experiment purpose. The experiment results of some algorithms like TDC [32], FIHC [5], etc were taken from the results stated in [6].

Table 2: Comparison of F-Score Using Our Approach

	F - Score		
Datasets	TDC	FIHC	SMFTDC
Reuters	0.46	0.506	0.60
Wap	0.47	0.391	0.51
Classic4	0.61	0.623	0.69
Hitech	0.57	0.458	0.56
Re0	0.57	0.53	0.57

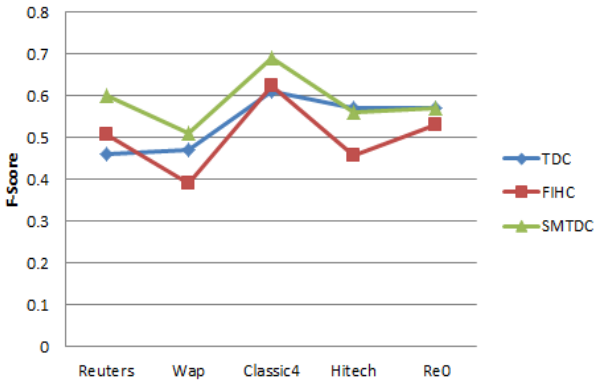

Fig. 3: Graphical Comparison of F-Score Using Our Approach.

The scalability of SMFTDC testing experiment consider 10 random datasets from Reuter dataset with the number of document increase from 1000 to 10000. Two min support at 30% and 50% are experimented. Fig. 4 shows that the execution time of both FIHC and SMFTDC is linear with respect to the number of documents and also SMFTDC outperforms.

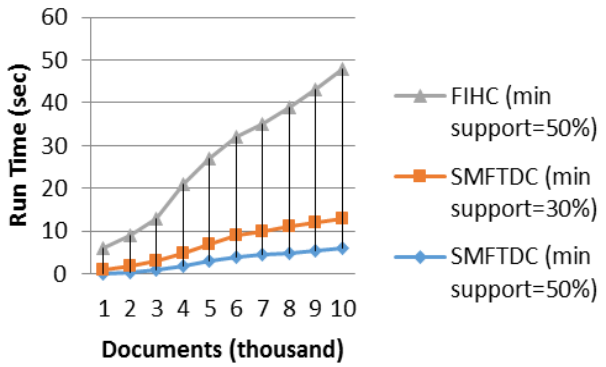

Fig. 4: Running Time of FIHC and SMFTDC with Respect to Document Size.

Our Semantic Maximal Frequent Term based document clustering Method results in less overlapping clusters and improve quality of final clusters. The SMFTDC method contributes in improvement of score function for generate quality clusters and use Maximal Frequent item sets. As per our method results showed in Table 2 that our approach outperforms its companion algorithms.

6. Conclusion and future work

In this paper, we presented document clustering using Maximal frequent item sets with semantic approach. Proposed algorithm use WordNet to take advantage of semantic aspects during clustering. We evaluate performance of our method on five standard datasets and found that our algorithm results comparatively good. In future we would like to work with another external knowledge base like Wikipedia for comparative analysis.

References

[1] Y. Zhao, and G. Karypis, Hierarchical Clustering Algorithms for Document Datasets, Data Mining and Knowledge Discovery, Vol. 10, No. 2, 2005, pp. 141-168. https://doi.org/10.1007/s10618-005-0361-3.

[2] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data, An introduction to Cluster Analysis, John Wiley & Sons, Inc (1990).

[3] Zhao, Y., Karypis, G.: Evaluation of Hierarchical Clustering Algorithms for Document Datasets, In Proc. of Intl. Conf. on Information and Knowledge Management. (2002) https://doi.org/10.21236/ADA439551.

[4] Beil, F., Ester, M., Xu, X.: Frequent Term-based Text Clustering, In Proc. of Intl. Conf. on Knowledge Discovery and Data Mining,. (2002).

[5] Fung, B., Wang, K., Ester, M.: Hierarchical Document Clustering using Frequent Item sets, In Proc. of SIAM Intl. Conf. on Data Mining. (2003).

[6] Malik, H.H., Kender, J.R.: High Quality, Efficient Hierarchical Document Clustering Using Closed Interesting Item sets, In Proc. of IEEE Intl. Conf. on Data Mining. (2006).

[7] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering usingWordNet and lexical chains," Expert Systems with Applications, vol. 42, no. 4, (2015) pp. 2264–2275.

[8] Hotho, A., Staab, S., et al.: Wordnet Improves Text Document Clustering, In Proc. of Semantic Web Workshop, the 26th Annual Intl. ACM SIGIR Conf. (2003).

[9] Zhang, X., Jing, L., Hu, X., et al: A Comparative Study of Ontology Based Term Similarity Measures on Document Clustering,In Proc. of 12th Intl. Conf. on Database Systems for Advanced Applications. (2007) https://doi.org/10.1016/j.eswa.2014.10.023.

[10] Su, C., Chen, Q., Wang, X., Meng, X.: Text Clustering Approach Based On Maximal Frequent Term Sets. In: Proceeding of 2003 IEEE International Conference on ―Systems, Man and Cybernetics", Harbin Institute of Technology, Shenzhen, China, (2009), pp.1551-1556.

[11] P. Treeratpituk and J. Callan. "Automatically labeling hierarchical clusters." Proceedings of the Sixth National Conference on Digital Government Research (2006), pp 167-176.

[12] Hartigan, J. A., Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. In: Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, (1979), pp. 100-108. https://doi.org/10.2307/2346830.

[13] M., Burdick, Calimlim, M., Gehrke, J.: MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. In: Proceedings of the 17th BIBLIOGRAPHY 63 International Conference on ―Data Engineering‖, Heidelberg, Germany (2001), pp 443-452. https://doi.org/10.1109/ICDE.2001.914857.

[14] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In Proc. of the Fifth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, (1999), pp 16–22. https://doi.org/10.1145/312129.312186.

[15] Agrawal, R., Srikant R.: Fast Algorithms for Mining Association Rules in Large Databases, Proc. VLDB 94,Santiago de Chile, Chile, 1994, pp. 487-499.

[16] A. Tversky, "Features of Similarity", Psycological Review, vol. 84, no. 4, (1977).

[17] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, (1995) August 20-25; Montréal Québec, Canada.

[18] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis and E. E. Milios, "Semantic similarity methods in WordNet and their application to information retrieval on the web", Proceedings of the 7th annual ACM international workshop on Web information and data management, (2005) October 31- November 05, Bremen, Germany. https://doi.org/10.1145/1097047.1097051.

[19] N. Negm, P. Elkafrawy, M. Amin, and A. M. Salem. Investigate the Performance of Document Clustering Approach Based on Association Rules Mining, International journal of Advanced Computer Science and Applications, Vol. 4, no. 8, (2013), pp. 142-151.

[20] Daniel, R.M. Shukla, A.K., "Improving Text Search Process using Text Document Clustering Approach", ISSN 2319-7064, International Journal of Science and Research (IJSR), Volume 3 Issue 5, (2014), pp 1424.

[21] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Proc. Of the 6th ACM SIGKDD international conference on TextMining Workshop, KDD 2000,2000

[22] WILLETT, P., Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing & Management, 24(5), 577-97 https://doi.org/10.1016/0306-4573(88)90027-1.

[23] Noor Asmat, Saif Ur Rehman, Jawad Ashraf and Asad Habib, Maximal Frequent Item sets Based Hierarchical Strategy for Document Clustering, in International Conference on Computer Science, Data Mining & Mechanical Engg. (ICCDMME'2015) April 20-21, 2015 Bangkok (Thailand).

[24] Sujata R. Kolhe ; S. D. Sawarkar, A concept driven document clustering using WordNet, In Proc. of the International conference Nascent Technologies in Engineering (ICNTE),2017.

[25] Yong Wang, Julia Hodges, Document Clustering with Semantic Analysis, In Proc. of the 39th Annual Hawaii International Conference on System Sciences, HICSS, Vol. 03, (2006), pp. 543,

[26] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), (2013), pp 1–12.

[27] Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim, "Exploiting noun phrases and semantic relationships for text document clustering," Journal of Information Sciences, Vol. 179, Issue 13, Jun 2009, pp. 2249-2262. https://doi.org/10.1016/j.ins.2009.02.019.

[28] Chun-Ling Chen, Frank S. Tseng, Tyne Liang, "An Integration of Fuzzy Association Rules and WordNet for Document Clustering," In Proc. of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD,2009, pp. 147-159.

[29] S.Vijayalakshmi, Dr.D.Manimegalai, "Query based Text Document Clustering using its Hypernymy Relation," International Journal of Computer Applications 23(1):Jun 2013, pp. 13–16, Jun. 2011

[30] Dang, Q., Zhang, J., Lu, Y., & Zhang, K. WordNet-based suffix tree clustering algorithm. In Paper presented at the 2013 international conference on information science and computer applications (ISCA 2013).