International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12232-12236

Maximal Frequent Term Based Document Clustering

[1]Harsha Patil and [2]Dr. Ramjeevan Singh Thakur

[1]Research Scholar, Department of Computer Applications,
[2]Associate Professor, Department of Computer Applications,
[1,2]Maulana Azad National Institute of Technology(MANIT),Bhopal, Madhya Pradesh, India.
[1]Orcid Id: 0000-0002-1801-4086

Abstract

The significance behind to world wide acceptance of Internet is the information. The internet crafts indefinite openings to access information and knowledge, to interact and to support us in our business and daily lives. Downloading and uploading information on Internet now becomes part of the daily work-routine. Vigorous generation of digitalized documents on Internet challenges technology for storing, retrieving and processing them. Out of which unstructured text documents are big task for researcher due to their high dimensionality. Document Clustering increases the quality of searching query. Clustering using maximal frequent item sets provide control over high dimensionality of text document during clustering. Here in our approach we try to tradeoff between high dimensionality with high accuracy of clustering and we got good results. We evaluate our method on the bases of F-Score on standard datasets and the results shows that our method performs comparatively better.

Keywords: Document Clustering, Text Mining, F-score, Item Sets, Maximally frequent, Score Function

INTRODUCTION

Document clustering is the unsupervised process of grouping unlabeled documents into sets called clusters. It is one of the main techniques for organizing large volume of documents into a small number of clusters. Document clustering can be classified in three categories: Partitioning method , Agglomerative and divisive clustering and item set based clustering Over the past decades, many document clustering algorithms have been proposed by researchers like K-Means Bisecting K-Means , Hierarchical Agglomerative clustering (HAC) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) to improve the quality of clusters. For handling high dimensions of text documents, frequent item sets based algorithms have high contribution. Item set which has frequency in document more than a user specified threshold value is become the candidate for making cluster. This threshold value is actually defined that a particular item set or term is frequent or not. This user specified threshold value is also called as global support. Approach presented in this paper use frequent item sets for dealing with high dimensional text documents. Proposed algorithm provides the tradeoff between cluster quality and global support to improve the result of clustering.

In this paper, we present improved version of algorithms which uses frequent item sets for text clustering. Frequent term sets are sets of terms co-occurring in more than a threshold percentage of all documents of a database. These frequent sets can be efficiently generated by using algorithms such as FP , Apriori etc. Application of Frequent item set for reduce drastically the dimensionality of the data, is efficient way for huge databases.

The Maximal Frequent term based Document Clustering (MTDC) method is evaluated on three standard datasets: Classic4, WAP and Reuters.

DOCUMENT CLUSTERING

Velocity of digital documents generation breaks all the records on Internet. Handling, analysis and management of this huge data is big challenge for researchers and IT professionals. Text mining methods provides a way to handle above problems. Nowadays, search engines are become boon for people to getting information for any aspects of the Life. Optimum solution of any query required precise retrieval of documents. Retrieval quality of documents can be improvised by using document clustering. Huge volume of documents can be classified using different methods of clustering. Document clustering algorithms classify documents in to different groups. Documents from one group are more relevant to each other as compare to documents in another group. Due to the diversified nature of documents, it becomes awkward to finalized general technique of document clustering which is suitable for all kinds of text data. The main objective of document clustering technique is to maximize cohesiveness of the clusters. Document clustering is being studied by many researchers from long time but still there is a scope for finding precise solution to challenges for text mining. The challenges are:

1. Appropriate Features Extraction.
2. Appropriate way for finding similarity between text documents.

3. Proper Preprocessing of text documents..

5. Appropriate methodology for evaluating performance of clustering process.

Document Clustering Process

Any document clustering process consists following steps:

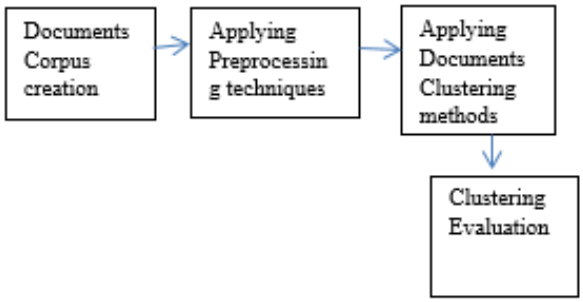

Figure 1: Steps for Document Clustering Process

The first step accumulates interesting documents and create corpus of its. Document collection process may include crawling, indexing, filtering etc. After creating corpus the next step is document preprocessing, which include removing stop words, stemming of terms and finally creating bag of words. After that suitable clustering technique is used for create clusters. Clustering technique and cluster quality evaluation is the last steps of document clustering process.

Item set-based Document Clustering Problem

Frequent item set is core of association rule mining. It provides significant way for finding text Patterns from documents. Frequent term sets are set of terms which come together frequently. Many algorithms like Apriori [16], FP [30] etc. are proposed by researchers for finding frequent item sets. A frequent item set based approach is widely acceptable because it efficiently reduce dimensionality of huge corpus. Global frequent item sets are basically set of items which seem together in more than user defined fraction of the document set.

In our algorithm we used Maximal frequent item sets.(MFI). MFIs are the compressed form of FIs. An frequent item set is called as Maximal Frequent item sets if there exists no super Item-set of that particular FI [34]. In our algorithm we eliminate small size MFIs by setting minimum size of MFI as 2.

RELATED WORK

As we discussed earlier high dimensionality of text documents is big challenge for researchers. Most of the Hierarchical and Partitioning methods works efficiently on low dimensional data, but results very poorly on high dimensional space. Text documents corpus usually contains thousands of words. Frequent-term-based algorithms support in sinking dimensionality of the text documents. Many researcher works remarkable in the area of document clustering using frequent item sets.

Frequent Term based Clustering (HFTC) [4] proposed by Beil F, Ester M, Xu X (2002) works for reducing high dimensionality of the documents and accuracy of clusters using frequent item sets. But it is not scalable.

After that Fung, et al came up with Hierarchical Document Clustering using frequent item sets (FIHC) [5] which leave behind HFTC. It provides meaningful labels to the clusters. Hasan H Malik,et al. [6] use closed frequent item sets but it results in loss of information. Puctada et al.(2006)[27] use hierarchical clustering method and generate cluster labels automatically. Hotho A, Staab S, Stumme G (2003) [8] uses WordNet as an ontology for improve quality of clustering. C. Su et al. (2009) proposed Maximal Frequent Term Set Clustering (MFTSC) [26] algorithm. MFTSC extracts MFIs using FP-Growth [30] algorithm, which helps in direction of more reducing documents dimensionality.

Our approach Maximal Frequent Term Based Document Clustering (MTDC) falls into the category of hard clustering, frequent item set based document clustering, which outputs Quality clusters in clustering process.

Framework: MTDC Approach

Maximal Frequent Term Based Document Clustering (MTDC) approach consist four modules, namely Preprocessing Module, Dimension reduction Module, Initial Cluster Construction Module and Final Cluster Construction Module. In MTDC framework, first module is responsible for corpus analysis which results in document vector. Second Module is responsible for reduce the dimensionality of document vector through generating maximal frequent itemsets. Third Module constructs initial clusters. These Initial clusters are soft clusters; it means they have overlapping of documents. After constructing initial clusters fourth module constructs Final clusters. Final clusters are constructed by using score function.

Preprocessing Module

First Module is responsible for receiving documents and creating corpus of text documents. This set of text documents is preprocessing using following steps:

Each document of corpus is splits into sentences. Then from each sentence, terms are extracted as features.

Removing common words also known as stop words that have no analytic value.

Stemming is used to convert remaining terms to their base forms.

After that the next step is to construct a document vector model which represent each document using a vector item frequencies.

Dimension reduction Module

The high dimensional document vector converted into feature vector by using only maximal frequent itemsets.We first apply Apriori algorithm on all the documents to mine the maximally frequent item sets (MFIs) .We limit our MFIs min-size to 2. The use of MFIs advances efficiency, accuracy and the removal of small size MFIs add further improvement.

Constructing Initial Clusters

A good clustering process always results in cohesive clusters i.e. Documents belongs to same cluster have more similarity as compare to documents belongs to different cluster. A global support value of any item sets specified that how many documents of corpus support that item set.

Any global frequent item sets are called cluster frequent for any cluster Ci if they present in some minimum fraction of documents in Ci. A minimum cluster support is used to find cluster frequent items.

This step constructs a cluster for each global 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 global frequent item sets from document vectors. Then we construct an initial cluster for each global frequent item sets i.e. we have five initial clusters.

Because any document may have more than one global frequent item sets so it may got membership of more than one cluster. So this step results in soft clustering i.e. one document may belongs to more than one cluster.

Finding Final Cluster

This step finds, final cluster for each document. Final clustering outputs hard clustering. Final cluster is basically most suitable cluster for any document. For that, We compare the global frequent items present in a document with the cluster frequent items of each of its initial cluster. For calculating the similarity between the document and the cluster frequent items score function is used.

Score Function

In our method, the score function consist three parts: Rewarding part, Penalty part and Bonus part. Suppose that item x appears in dj. For calculating score of cluster Ci for any dj: we reward Ci if x appears in cluster Ci otherwise we penalize Ci.

The Bonus part is global_support(hidden term) of the hidden term which has less threshold support than global threshold value, but it is present in document and also has a particular minimum support to the cluster label. This minimum support is less than global threshold support but greater than a predefined support. We called it hidden support (HS) and this hidden support is weighted to find accurate cluster of the documents. Hidden term bonus can be taken by only that cluster which has all terms of cluster label supported by Hidden term. So if the hidden term is fully supported with the cluster label than only hidden weight will be given otherwise it will be ignored.

$$\text{Score}(C_i \leftarrow d_j) = [\sum_{x} n(x) * \text{cluster\_support}(x)] - [\sum_{x'} n(x') * \text{global\_support}(x')] + HS$$

x represents a global FI in dj and also cluster frequent in Ci
x' represents a global FI in dj but not cluster frequent in Ci
$n(x)$ is frequency of x in the feature vector of dj
$n(x')$ is the frequency of x' in the feature vector of dj

Input: A document set D; explicit stop word list

Output: Target Cluster set C

1. Extract the termset $T_D = \{ t_1, t_2, t_3 \ldots\ldots\ldots t_n \}$
2. Remove all stop words from T_D.
3. Apply stemming for T_D.
   //create document term matrix which represent document in form of $d_i = \{ (t_1,fi_1), (t_2,fi_2), \ldots\ldots\ldots\ldots\ldots (t_n,fi_n)\}$
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 $C_i \in C$
   Cft = (C_i , min_csup)
8. For each $d_j \in D$ do //finding final cluster
   For each cluster $C_i \in$ Cft do
   Calculate Score($C_i \leftarrow d_j$) using formula given in Score Function
9. Assign d_j in final cluster Cft which has maximum Score for membership

Figure 2 : Algorithm MTDC

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 and WAP datasets were used for experiment purpose.

Reuters: This dataset of 21578 news articles that appeared on Reuters newswire in 1987. Out of 21578 articles, 8654 articles are uniquely assigned to one of these classes.

Wap: This dataset has 1560 web pages collected during WebACE project[28] from the yahoo! Subject hierarchy.This dataset consist of 20 different classes. Datasets for clustering were obtained from Cluto clustering tolkit [15].

Classic4:Classic collection consists 7095 documents from 4 different categories: CACM, CISI, CRAN, and MED.

The experiment results of some algorithms like TDC [32], FIHC [5], etc were taken from the results reported in HCCI [6]. HCCI [6] presents many approaches for clustering; we have taken the best scores out of all of them.

Table 1: Comparison of F-Score using our approach

Datasets	TDC	FIHC	MTDC
Reuters	0.46	0.506	0.571
Wap	0.47	0.391	0.51
Classic4	0.61	0.623	0.594

Our Maximal Frequent Term based document clustering Method reduces the cluster overlapping and produces more qualitative final clusters. In MTDC method, we work with improvements to existing score function used in FIHC and use Maximal Frequent item sets. We can see that results in Table 4 that our approach outperforms its companion algorithms.

CONCLUSION AND FUTURE WORK

In this paper, we presented document clustering using Maximal frequent item sets. Proposed algorithm constructs more precise clusters by using hidden term support value. We evaluate performance of our method on three standard datasets and found that our algorithm results comparatively good. In future we would like to work with external knowledge base for support out hidden term semantically. We also like to work with more datasets to find applicability and scope of our algorithm.

REFERENCES

[1] Pudi, V., Haritsa, J.R.: Generalized Closed Item sets for Association Rule Mining,In Proc. of IEEE Conf. on Data Engineering. (2003)

[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)

[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] Yu, H., Searsmith, D., Li, X., Han, J.: Scalable Construction of Topic Directory with Nonparametric Closed Termset Mining, In Proc. of Fourth IEEE Intl. Conf. on Data Mining. (2004)

[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] Hotho, A., Maedche, A., Staab, S.: Text Clustering Based on Good Aggregations, In Proc. of IEEE Intl. Conf. on Data Mining. (2001)

[10] 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)

[11] Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, In Proc. of The 21st National Conf. on Artificial Intelligence. (2006)

[12] Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, In Proc. of The 20th Intl. Joint Conf. on Artificial Intelligence. (2007)

[13] Hu, X., Zhang, X., Lu, C., et al.: Exploiting Wikipedia as External Knowledge for Document Clustering, In Proc. of Knowledge Discovery and Data Mining. (2009)

[14] Hu, J., Fang, L., Cao, Y., et al: Enhancing Text Clustering by Leveraging Wikipedia Semantics, In Proc. of 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval.

(2008)

[15] Cluto.:http://glaros.dtc.umn.edu/gkhome/views/cluto

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

[17] K. Chakrabarti, S. Mehrotra. Local Dimension reduction: A new Approach to Indexing High Dimensional Spaces, VLDB Conference, 2000.

[18] C. C. Aggarwal, P. S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces, ACM SIGMOD Conference, 2000.

[19] Harsha P., Ramjeevan Singh T., Document clustering : a summarized survey " Book entitled:Pattern and data analysis in healthcare settings ,ISBN: 9781522505365 (hardcover),9781522505372 (ebook- www.igi-global.com)

[20] G. V. R. Kiran, K. Ravi Shankar, Vikram Pudi, "Frequent Item set based Hierarchical Document Clustering using Wikipedia as External Knowledge", In: Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems , 2010

[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. 1988. "Recent Trends in Hierarchic Document Clustering: A Critical Review." Information Processing & Management, 24(5), 577-97

[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] 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, Page 1424 (2014)

[25] 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, pp. 142-151, 2013.

[26] Chen CL, Tseng FSC, Liang T (2010) Mining fuzzy frequent item sets for hierarchical document clustering. Inf Process Manag 46(2): 193–211

[27] 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, pp.1551-1556, (2009)

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

[29] Eui-Hong (Sam) Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin.kumar, B. Mobasher, and Jerry Moore, WebAce: A Web Agent for Document Categorization and Exploration. Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)

[30] Borgelt, C.: An Implementation of the FP-growth Algorithm. In: Workshop ―Open Source Data Mining Software (OSDM'05, Chicago, IL)‖, pp. 1-5, ACM Press, New York, NY, USA (2005)

[31] C.L. Chen, F.S.C. Tseng, T. Liang, An integration of fuzzy association rules and Word Net for document clustering, Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09), 2009, pp. 147–159.

[32] Yu, H., Searsmith, D., Li, X., Han, J.: Scalable Construction of Topic Directory with Nonparametric Closed Termset Mining, In Proc. of Fourth IEEE Intl. Conf. onData Mining. (2004)

[33] Hartigan, J. A., Wong, M. A.: Algorithm AS 136: A K-Meanss Clustering Algorithm. In: Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, pp. 100-108, Royal Statistical Society (1979)

[34] 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‖, pages 443-452, Heidelberg, Germany (2001)