Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Document Clustering

[1]Monika Gupta, [2]Kanwal Garg
[1] M.Tech Scholar, [2] Asstt. Professor
[1, 2] Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana, India

Abstract- ***Text mining has becoming an emerging research area in now-a-days that helps to extract useful information from large amount of natural language text documents. The need of grouping similar documents together for different applications has gaining the attention of researchers in this area. Document clustering organizes the documents into different groups called as clusters. The documents in one cluster have higher degree of similarity than the documents in other cluster. The paper provides an overview of the document clustering reviewed from different papers and the challenges in document clustering.***

Keywords- Text Mining, Document Clustering, Similarity Measures, Challenges in Document Clustering

I. INTRODUCTION

Clustering is the unsupervised technique that group data objects into classes or groups or clusters such that objects within the same cluster have higher degree of similarity than the objects in different cluster. The objects are highly dissimilar to objects in other classes or groups. Thus, clustering helps the efficient visualization of documents in a collection by grouping similar and relevant documents together in one cluster[20]. Cluster is an ordered list of data which have familiar characteristics. Cluster analysis can be done by finding similarities between data. The similarities can be found by comparing the characteristics. A good clustering process able to produce high superiority clusters. The intra class similarities high and interclass similarity is low[2]. Clustering can be performed without the knowledge of the category structure of class or presumptions. Clustering uses the technique to define similarity or dissimilarity among objects[10]. By organizing the similar documents together, large collection of documents can be easily navigate, browsed and organized. Document clustering plays an important role in various fields like business applications, knowledge discovery etc.[8]. Common approach for all clustering techniques is to find cluster centre that will represent each cluster[1]. Clustering method and clustering algorithm are different things. A clustering method is a general strategy applied to solve a clustering problem where as a clustering algorithm is simply an instance of a method[12].

II. LITERATURE SURVEY

After reviewing the research papers, it has been observed that the term data clustering was first appeared in the title of a 1954 article dealing with anthropological data[12]. Clustering is the efficient technique that helps to make clusters without the knowledge of category structure of class or pre-assumptions[10]. To carry on the present work the researcher explains the document clustering papers that carried the latest research from 2004 to 2015.

Anuj Sharma et al. [3] explains the document clustering based on wordset. The paper presents that it is a efficient way for clustering the documents. WDC uses a hierarchical approach for the clustering of text documents that have common words. H. N. Ganganave et al. [4] presents the methodology for the identification of criminal. The paper provides the methodology to evaluate the document clustering of the criminal database by using k-means clustering and with the help of outlier detection and cluster the criminals data based on the type of crime. Hung Chim et al. [5] proposes the phrase based document similarity that determines the pair wise similarity of document based on Suffix Tree Document model. The each node of suffix tree is mapped into a unique feature term in VSD(Vector Space Document) model. The VSD and STD model helps in text based information retrieval. J. Jayabharathy et al. [6] presents the modified semantic based model in which the related terms can be extracted and treated as concepts for the concept based document clustering and concept based topic discovery. The comparison shows that document clustering by terms and related terms is better than document clustering by single term only. Khaled M. Hammouda et al. [7] explains the two key parts of document clustering. The first one is phrase based document index model, the Document Index Graph that helps for the incremental construction of phrase based index of document set. The second one is incremental document clustering algorithm that forms its basis on maximizing the tightness of clusters by determining the pair wise documents similarity distribution inside cluster. Livin Sebestian Matei et al. [9] presents the documents using time series. The time series model can be used as a another way to the vector space representation. Dynamic time wrapping is used to find out the distance between time series. The similarity between two documents can be defined with the help of time series representation and the dynamic time wrapping. Pramod Bide et al. [14] proposes an algorithm that takes inputs as keywords found after extraction. It helps to solve the problem of over clustering by dividing the documents into small groups using divide and conquer strategy. The improved document clustering algorithm is provided that generates the number of clusters for any text document and uses the cosine similarity measure to place similar documents in proper

cluster. Ruizhang Huang et al. [16] proposes the Dirichlet Process Mixture Model Feature Partition(DPMFP) approach that find out the latent cluster structure. The cluster structure has its basis on DPM model that does not require the knowledge of numbers of clusters uses as input. The document clustering and feature selection are handled simultaneously. Rupesh Kumar Mishra et al. [17] explains a new inter-passage based clustering technique which will cluster the segment of documents on the basis of similarities. The approach helps for the consideration of word count as well as the SentiWordNet score of that word in the segment which can be useful for the systematic organization of documents with the availability of large amount of documents. T. W. Fox et al. [21] focuses on the document vector compression because the memory requirement is high when you have vast number of documents. By using document vector compression, the runtime memory requirement can be reduced by 60%. Wael M. S. Yafoz et al. [23] focuses on the fact to group the textual documents based on the similarities. It also shows the importance of dynamic clustering for mining the frequent terms with included named entity. Zhenya Zhang et al. [25] presents an approach based on genetic algorithm for the aggregation problem of clustering. The paper explains that GeneticCA is found better than the clustering performance of original clustering divisons with clustering precision.

The review shows that the researches carried out at different times helps to improve the cluster performance by not degrading the cluster quality. All the experiments tried to increase the accuracy in terms of F-measure and time complexity. Better clustering results are provided in terms of precision and recall in a short time.

III. DOCUMENT CLUSTERING

Document clustering is the process of collecting similar documents into groups, where similarity is some function on a document[26]. It is the automatic organization of documents into clusters[11]. It has applications in automatic organization of documents, topic extraction, fast information retrieval or filtering[26].

A. Procedure of Document Clustering

There is not a single operation from the collection of documents to the clustering of document collection. It includes the number of stages that consist generally four main stages[11]-

1) *Preprocessing:* Before document representation, we require some preprocessing. Firstly, we need to remove stop words such as ‘a’, ‘any’, ‘the’, since they are frequent and irrelevant. Secondly, we need to stem the words. For eg. ‘flying’ and ‘flew’ are stemmed to ‘fly’[24].
2) *Feature Extraction:* It employs to produce the set of features by parsing each document. It helps to remove the noise and reduce the dimensionality of feature space. The most commonly used feature selection metric are term frequency and inverse document frequency[11].
3) *Document Representation:* Most of the clustering approaches use the vector space model for document representation. In VSM, the document is represented as the vector of keywords. A collection of n documents with m unique words is represented as an m*n matrix, where each document is a vector of m dimension[20].
4) *Document Clustering:* At this stage, the target documents are grouped into different clusters on the basis of selected features[11].

B. Document Clustering Similarity Measures

The cluster analysis method form their basis on measurement of similarity between a pair of objects. The determination of similarity includes three steps between a pair of objects:

- Variables selection to characterize the objects.
- Weighting scheme selection for the variables.
- Similarity coefficient selection that describes the degree of resemblance between two attribute vectors[11].

The different similarity measures are:

1) *Euclidean Distance:* For two documents d_a *and* d_b having term vectors t_a and t_b respectively, the Euclidean distance is given as:

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2},$$

Where the termset is T = t_1, ..., t_m.

2) *Cosine Similarity:* When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors

Given two documents $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$, their cosine similarity is

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|},$$

Where t_a and t_b are the m-dimensional vectors over the term set T = {t_1,….,t_m}.

3) *Jaccard Coefficient:* It measures the similarity as the intersection divide by the union of objects. It is given by:

$$SIM_J(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}|^2 + |\overrightarrow{t_b}|^2 - \overrightarrow{t_a} \cdot \overrightarrow{t_b}}.$$

It ranges between 0 and 1, where 1 means the two objects are same and 0 means they are completely different.

4) Pearson Correlation Coefficient: For the term set T = $\{t_1,\ldots,t_m\}$, the formula is given as:

$$SIM_P(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{m\sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{[m\sum_{t=1}^{m} w_{t,a}^2 - TF_a^2][m\sum_{t=1}^{m} w_{t,b}^2 - TF_b^2]}}$$

$$\text{where } TF_a = \sum_{t=1}^{m} w_{t,a} \text{ and } TF_b = \sum_{t=1}^{m} w_{t,b}.$$

It ranges from +1 to -1, it is 1 when $t_a=t_b$[15].

5) Manhattan Distance: It is popularly known as block distance. It computes the distance that would be travelled to get from one data point to the other if a grid like path is followed[19]. The manhattan distance between two items is the sum of the differences of their corresponding components.

The distance between a point ***X***=(***X1***, ***X2***, etc.) and a point ***Y***=(***Y1***, ***Y2***, etc.) is expressed as:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Where n is the number of variables, ***Xi*** and ***Yi*** are the values of *i*th variable at points **X** and **Y** respectively[27].

C. Challenges In Document Clustering

Document clustering technique is well known from many decades but still it is far from a solved problem. The challenges are:

- *Appropriate feature selection:* The feature selection is a major problem in document clustering because the clustering is an unsupervised technique, it does not know the structure of class, so it is harder to select the features due to absence of class labels that would guide the search for relevant information[22].
- *Cluster labeling:* The problem of labeling clusters is preferring descriptive, relevant, human-readable labels for the clusters that are produced by a document clustering algorithm[18]. A good clustering labeling helps in browsing by giving the brief but meaningful description about cluster. So, the clustering method should provide proper labels to cluster that are easily understandable[20].
- *Selection of appropriate similarity measure:* The similarity measures like euclidean, manhattan can be used for numerical attributes, but for the categorical attribute, the selection of similarity measure is difficult[13].
- *Knowledge about input parameters:* The clustering algorithm requires the certain information before clustering like number of clusters. The accuracy of clustering result may depend on such input parameters but identifying the exact value of such parameters before execution is difficult[20].

IV. CONCLUSION

The overall goal of text mining is to extract information from the large dataset and provide to its users into an understandable form. Document clustering is a fundamental and crucial operation in various application such as document organization, summarization, information retrieval etc. The paper provides an overview on document clustering. Further work in this area is cluster improvement by not affecting the cluster quality.

REFERENCES

[1] Aastha Joshi, Rajneet Kaur, " A Review: Comparative Study of Various Clustering Techniques in Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X, vol. 3, Issue 3, 2013

[2] Amandeep Kaur Mann, Navneet Kaur, "Review Paper on Clustering Techniques", *Global Journal of Computer Science and Technology Software & Data Engineering*, vol. 13, Issue 5, Version 1.0, 2013

[3] Anuj Sharma, Renu Dhir, "A Wordset Based Document Clustering Algorithm For Large Dataset", *International Conference on Methods and Models in Computer Science*, 2009

[4] H. N. Gangnave, M. C. Nikose, P. C. Chavan, "A Novel Approach for Document Clustering to Criminal Identification by Using ABK-Means Algorithm", *IEEE International Conference on Computer, Communication and Control (IC4-2015)*

[5] Hung Chim, Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering", *IEEE Transactions On Knowledge And Data Engineering*, vol. 20, No. 9, 2008

[6] J. Jayabharathy, S. Kanmani, A. Ayeshaa Parveen, "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature", *Department of Computer Science & Engineering, IEEE*, 2011

[7] Khaled M. Hammouda, Mohamed S. Kamel, "Efficient Phrase-Based Document Indexing for Web Document Clustering", *IEEE Transactions On Knowledge And Data Engineering*, vol. 16, No. 10, 2004

[8] Latika, "An Effective and Efficient Algorithm for Document Clustering" *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X vol. 5, Issue 5, 2015

[9] Livin Sebestian Matei, Stefan Trausan-Matu, "Document clustering based on time series", *19th International Conference on System Theory, IEEE,* 2015

[10] Mamta Mahilame, Mr. K. L. Sinha, " A Survey Paper On Different Techniques Of Document Clustering" *Department of Comp. Sci. & Engg.* vol. 2, Issue-1,2015

[11] Neepa Shah, Sunita Mahajan, "Document Clustering: A Detailed Review*", International Journal of Applied Information System*, ISSN : 2249-0868, vol. 4, No.5, 2012

[12] Neha Soni, Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review*", International Journal of Advanced Research in Computer Science and Software Engineering,* ISSN: 2277 128X, vol. 2, Issue 8, 2012

[13] Parul Agarwal, M. Afshar Alam, Ranjit Biswas, "Issues, Challenges and Tools of Clustering Algorithms", *International Journal of Computer Science Issues*, vol. 8, Issue 3, No. 2, 2011

[14] Pramod Bide, Rajashree Shedge, ''Improved Document Clustering using K-means Algorithm", Dept. Computer Engg, IEEE, 2015

[15] Pranjal Singh, Mohit Sharma, "Text Document Clustering and Similarity Measures", Dept. of Computer Science & Engg., 2013

[16] Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi, "Dirichlet Process Mixture Model for Document Clustering with Feature Partition*", IEEE Transactions On Knowledge And Data Engineering*, vol. 25, No. 8, 2013

[17] Rupesh Kumar Mishra, Kanika Saini, Sakshi Bagri, "Text Document Clustering on the basis of Inter-passage approach by using K-means*", International Conference on Computing, Communication and Automation, IEEE*, 2015

[18] Sujata Kolhe, Dr. Sudhir Sawarkar, " Review of Document Clustering Techniques: Issues, Challenges and Feasible Avenue", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, Issue 4, ISSN: 2277 128X, 2015

[19] S. Mahalakshmi ," Challenging Issues and Similarity Measures for Web Document Clustering", IOSR Journal of Computer Engineering, vol. 17, Issue 1, 2015

[20] Sunita Bisht, Amit Paul, "Document Clustering: A Review", *International Journal of Computer Applications*, vol. 73, No.11, 2013

[21] T. W. Fox, "Document Vector Compression and Its Application in Document Clustering*",Intelligent Engines, IEEE*, 2005

[22] Volker Roth, Tilman Lange, "Feature Selection in Clustering Problems", Institute of Computational Science

[23] Wael M. S. Yafooz, Siti Z.Z. Abidin, Nasiroh, Omar, Rosenah A. Halim, "Dynamic Semantic Textual Document Clustering Using Frequent Terms and Named Entity*", IEEE 3rd International Conference on System Engineering and Technology*, 2013

[24] Yogesh Jain, Amit Kumar Nandanwar, "A Theoretical Study of Text Document Clustering", *International Journal of Computer Science and Information Technologies*, vol. 5, ISSN: 0975-9646, 2014

[25] Zhenya Zhang, Hongmei Chang, Shugang Zhang, Wanli Chen, Qiansheng Fang, "Clustering Aggregation based on Genetic Algorithm for Document Clustering", *IEEE*, 2008

[26] https://en.wikipedia.org/wiki/Document_Clustering

[27] http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Manhattan_Distance_Metric.htm

