Systematic Review of Document Clustering Algorithms and Future Research Directions

Anjali Vashist
Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India
Email: anjali.vashist29@gmail.com

RajenderNath
Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana, India
Email: rnath2k3@gmail.com

----------ABSTRACT----------
Clustering is an efficient technique that organizes a large quantity of unordered text documents into a small number of significant and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms. It is studied by the researchers at broad level because of its broad application in several areas such as web mining, search engines, and information extraction. It clusters the documents based on various similarity measures. So far many clustering algorithms are proposed in the literature but they are not reviewed systematically. This paper intends to review the literature on document clustering algorithms systematically and suggest various research directions in this field.

Keywords-Document clustering, research methodology, similarity measures, research direction, conclusion

1. Introduction

Numbers of users on the Internet are increasing day by day and information on the Web is growing exponentially. Extracting useful information from the Web is becoming increasingly difficult. To address this issue, many data mining techniques have been proposed in the literature. Data mining is the process of extracting the hidden, previously unknown and useful information from data. Document clustering, the sub-problem of data mining is the process of organizing documents into clusters and the documents in each cluster share some common properties according to similarity measure used. The document clustering algorithms play an important role in helping users to quickly navigate, sum up and organize the information. Document clustering can be used for classification of different documents, detecting content duplicity, recommending Web pages to users, optimizing searches etc.

This paper intends to systematically analyze the existing document clustering techniques and draws research directions in this field. The rest of the paper is organized as follows: Section 2 describes general steps of document clustering. Section 3 presents research methodology. Section 4 gives detailed review of document clustering. Section 5 draws research direction in the field of document clustering. Section 6 gives the conclusion and future scope.

2. Document Clustering

There are mainly five main phases in the document clustering - (a) preprocessing (b) feature extraction (c) document representation (d) similarity measures (e) clustering. A brief description of each of these phases is presented in the ensuing paragraphs.

(a) Preprocessing: This phase further consists of four sub phases (i) Filtering (ii) Tokenization (iii) Stemming (iv) Pruning. In filtering punctuation marks and special characters are removed from the plain text documents. Sentences are split into individual words or tokens in tokenization. In stemming stop words are removed from the document as these words do not convey any meaningful information and they are reduced to their base form. In pruning the words which have very low frequency are removed from the resulting dataset.

(b) Feature extraction: A set of keywords is extracted from the dataset and then feature vector is created by extracting those keywords whose frequency is greater than threshold.

(c) Document representation: A model is needed to represent the text documents. In literature many such models have been proposed such as tf-idf, tf probe etc.
(d) Similarity measures: In the literature many similarity measure such as Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson

Correlation Coefficient, and AveragedKullback-Leibler Divergence etc are proposed in the literature

(e) Clustering: It classifies the documents into clusters using some similarity measurement mentioned above.

3. Research Methodology

Firstly, research papers relating to document clustering will be collected from various resources such as ACM, Springer, and IEEE etc. Then these research papers will be classified into different categories based on the clustering technique used in the paper. After that representative research papers along each class will be critically analyzed and research directions will be drawn.

4. Document Clustering Techniques

Total 19 research papers were collected and were classified into three categories viz. partitioning algorithms, hierarchical algorithms and soft-computing algorithms. Eight research papers were found under the category of partitioning algorithms; nine research papers under hierarchical algorithms and four under soft computing algorithms. The research papers under each category are discussed and analyzed below.

4.1 Partitioning Algorithms

It relocates instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters need be pre-defined by the user. To achieve global optimality in partitioned-based clustering, an extensive enumeration process of all possible partitions is required. Now the representative papers under this category are analyzed and discussed below.

Fox [1] proposed document compression method to reduce the run time memory requirement by represent each document in vector format and apply discrete cosine transform on them. The proposed method reduced the memory requirement to 60%.

Forsati et.al [2] proposed k-means algorithm for clustering. The k-means algorithm could generate a local optimal solution. They presented novel harmony search clustering algorithms that deal with documents clustering based on harmony search optimization method. By modeling clustering as an optimization problem, first, they proposed a pure harmony search based clustering algorithm that finds near global optimal clusters within a reasonable time. Contrary to the localized searching of the K-means algorithm, the harmony search clustering algorithm performs a globalized search in the entire solution space. Then harmony clustering is integrated with the k-means algorithm to achieve better clustering. The proposed algorithms improved the k-means algorithm by making it less dependent on the initial parameters such as randomly chosen initial cluster centers, hence more stable.. Experimental results reveal that the proposed algorithms could find better clusters when compared to K-means and the quality of clusters was comparable and converged to the best known optimum faster.

Shameem et.al [3] they identified another problem of K-means algorithm. It was failed to classify the documents in disjoint datasets. To solve this problem, they proposed a technique to measure the initial guess of the centroid points for K clusters. They represented the documents in the vector space model and some dissimilarity measurement techniques could be applied over the document set to find out the most dissimilar K documents. Then K points should be used as K centroid which ensures to classify the document in K disjoint datasets.

Daling et.al [4] the most popular document clustering method K-Means had the shortcoming of its cluster intra dissimilarity. They proposed an optimized K-Means algorithm. It introduced the scalar factor from SOM into means during K-Means assignment stage for controlling the influence to the means from new objects. Experiments showed that the optimized K-Means algorithm had more F-Measure and less Entropy of clustering than standard K-Means algorithm, thereby reduces the intra-dissimilarity of clusters effectively.

Mabel Rani et,al [5] they identified one more problem of k-means. The problem was that K-means algorithm was sensitive to the selection of the initial partition and may converge to local optima. To solve this problem they used Improved Particle Swarm Optimization (IPSO). In IPSO, training data was presented to the algorithm one by one, and the training proceed is a one-shot incremental algorithm. The proposed solution was generated more accurate and better clustering results when compared with existing K-means.

Guran et.al [6] they proposed a method which reduced the size of processed data and execution time of K-means clustering algorithm. They used NMF based dimension reduction methods. It was a text summarization process which applied to each document.

Ranjana Agrawal et.al [7] they proposed a solution to the four major problems of K-means algorithm. The problems were: 1) The number of clusters K had to be given as input and 2) Based on the initializations it converges to different local minima. 3) It was slow and cannot be used for large number of data points.4) It could not handle empty clusters. To resolve all the issues they had modified VSM input to K-means. Then they developed a new algorithm using Cosine Similarity Threshold which resolves all these issues.

Pramod Bid [8] he found the solution of over clustering. To solve this problem they proposed Improved Document Clustering algorithm which generates number of clusters for any text documents and used the cosine similarity measures to place similar documents in proper clusters. They used 20newsgroup dataset for experimental study. The F1 score of proposed algorithm was better than the existing techniques.

4.2 Hierarchical Algorithms

These methods build the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. Now the representative papers under the hierarchical algorithms are analyzed and discussed below.

Naveen et.al [9] Information Retrieval (IR) systems such as search engines retrieve a large set of documents, images and videos in response to a user query. Computational methods such as Automatic Text Summarization (ATS) reduce this information load enabling users to find information quickly without reading the original text. The challenges to ATS include both the time complexity and the accuracy of summarization. To solve this issue they proposed Information Retrieval system consists of three different phases: Retrieval phase, Clustering phase and Summarization phase. In the Clustering phase, they extend the Potential-based Hierarchical Agglomerative (PHA) clustering method to a hybrid PHA-Clustering Gain-K-Means clustering approach. The experimental results showed that the proposed work increase the efficiency and accuracy of clusters when compared to both the conventional Hierarchical Agglomerative Clustering (HAC) algorithm and PHA.

Hammouda et.al [10] proposed a clustering method for phrases. Phrase based analysis means that the similarity between documents should be based on matching phrases rather than on single words only. A novel phrase-based document indexing model, the Document Index Graph (DIG)that captures the structure of sentences in the document set, rather than single words only. But this model was only for Web document clustering.

B. F. Momin et.al [11] proposed a document clustering method for phrases with better accuracy than earlier one They proposed a Document Index Graph (DIG) model that explained the effectiveness of phrase based similarity over term based similarity, and then they proposed Document Index Graph based Clustering (DIGBC) algorithm to enhance the DIG model for incremental and soft clustering algorithm to cluster documents efficiently. The quality of DIGBC algorithm was not very good because of the less accuracy of document-cluster similarity calculation and the threshold value determination.

Lee et.al [12] they combined the two approaches to form a new system named CONDOR system with hierarchical structure based on document clustering using K-means algorithm to reduce the time complexity of hierarchical clustering methods. The proposed system was not very efficient in terms of performance and it was applicable only for small Web data.

Hammouda et.al [13] proposed a solution to the problem of modularity, flexibility, and scalability. To solve this problem they proposed a hierarchically distributed Peer-to-Peer (HP2PC) architecture and clustering algorithm. The architecture was based on a multilayer overlay network of peer neighborhoods. The proposed method did not allow centroids to cross neighborhoods through higher levels.

Ambedkar [14] he proposed WDC (Word sets-based Clustering)an efficient clustering algorithm based on closed words sets to handle the very high dimensionality of the document, very large size of the datasets and understandability of the Cluster description. WDC used a hierarchical approach to cluster text documents having common words. WDC found scalable, effective and efficient when compared with existing clustering algorithms like K-means and its variations.

Murugesan et.al [15] they proposed a hybrid clustering algorithm and it was a combination of divisive and agglomerative hierarchical clustering algorithm which generate better clusters than the bisect K-means divisive hierarchical algorithm and taken less time complexity than the agglomerative hierarchical clustering algorithm (UPGMA) .They used a vector space model in which each document was represented by a set of weights of the terms

appeared in a collection. They used tf -idf term weighing scheme to identify the importance of a term in a collection. Their method used both the Euclidean distance and Cosine similarity measure for finding the relationship between the documents/clusters.

Meena et.al [16] they proposed work of cluster summarization of huge text documents. To implement this they proposed Dynamic Peer to Peer (P2P) document clustering and cluster summarization (DP2PCS) architecture which was based upon bonus words and stigma words. The major limitation of this technique was security level was not very high.

Selangor et.al [17] they proposed a new textual document clustering method which was used to achieve high efficiency and high-quality data clustering. Fuzzy Frequent Item-set- Based Hierarchical Clustering method (F2IHC) [16] was used to improve the clustering quality. By using fuzzy association rules mining, it was easy to realize a relation and integrate linguistic terms. The quality and efficiency was not very good.

4.3 Soft Computing Algorithms

Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. The guiding principle of soft computing is: Exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost. Several papers under this category are analyzed and discussed below.

Truh.cao et.al [18] they identified the problem of KBDC (keyword based document clustering). KBDC had limitation due to simple treatment of words and hard separation of clusters. To solve this problem they introduced named entities which were the key elements defining document semantics. But the problem of this technique was that its entropy and f-measure was not so good.

Zang et.al [19] they introduced a method which was based on genetic algorithm known as GeneticCa to improve the cluster aggregation performance. This method was work only for bit string.

Lailil [20] they proposed a method was known as Latent Semantic Index (LSI) approach. It used the concept of Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA)The objective of this method was to reduce the matrix dimension by finding the pattern in document collection with refers to concurrent of the terms. Each method was implemented to weight of term-document in vector space model (VSM) for document clustering using fuzzy c-means algorithm.

Feng et.al [21] the Performance of Medical Literature Analysis and Retrieval System Online (MEDLINE) articles were enhanced by a novel semisupervised spectral clustering method which is known as SSNCut for clustering over the local cost (LC) similarities. It was worked on two types of constraints: must-link (ML) constraints on document pairs with high MeSH semantic (MS) similarities and cannot-link (CL) constraints on those with fewer similarities. The performance of SSNCut was the limitation of this paper.

5. Research Directions

Document clustering has been used in different fields and several authors face several challenges and proposed different algorithms to solve all the issues like complexity issue, runtime memory requirements. After going through many research papers on partitioning algorithms many problems were identified such as no document clustering based on semantics are proposed yet. In future researchers can develop the clustering algorithms based on semantic analysis of text documents. In hierarchical based algorithms the major bottleneck is the height of the tree so in future techniques can be suggested to shorten the height of the tree. In soft computing algorithms proposed so far they show poor entropy and f-measure. Some methods are based on bit string only. Some are based on local cost, MeSH semantic similarities and must link and cannot link constraints lack in performance so therefore the researchers are needed to address these issues.

6. Conclusion

In this paper the authors have critically analyzed document clustering algorithms reported in the literature .By classifying them in the three broad categories, the limitations of the algorithms under these categories have been identified. The future research directions suggested in this field can be very useful for the researchers.

References

[1] Fox.T, *"Document Vector Compression and Its Application in Document Clustering"*, Conference on Electrical and Computer Engineering, pp. 2029 – 2032, May 2005, IEEE.

[2] Forsati.R Meybodi.MR, Neiat.M, ,"*Hybridization of K-means and Harmony Search Methods for Web Page Clustering*" ,Conference on Web Intelligence

and Intelligent Agent Technology , pp. 329 - 335 2008,IEEE

[3] Ferdous.R, Shameem.M, "An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering", Conference on Internet, pp. 1-6, 3-5 Nov. 2009 , IEEE.

[4]Wang.D,"*An Optimized K-Means Algorithm of Reducing Cluster Intra-dissimilarity for Document Clustering*, pp. 785 – 790, 2005, Springer.

[5] Parthipan.L, Rani,.R, "*Clustering Analysis by Improved Particle Swarm Optimization and K-Means Algorithm*," Conference on Sustainable Energy and Intelligent Systems pp. 27-29 Dec. 2012 IEEE.

[6]Güran.A, Kaptıkaçtı.H ,Naiboğlu.M, "*NMF based Dimension Reduction Methods for Turkish Text Clustering*", International Symposium on Innovations in Intelligent System and Applications, pp.1-5 ,19-21 June 2013 IEEE..

[7]Agrawal.R and Phatak.M, "*Document clustering algorithm using modified k-means*", Conference on Advances in Recent Technology pp. 294-296, 19-20 oct 2012, IEEE.

[8] Pramod Bide, "Improved Document Clustering using K-means Algorithm", conference on Electrical, Computer and Communication technology, pp. 1-5, 5-7 March 2015 IEEE

[9] Gopal.N, Nedunga.P, "*Query-based Multi-Document Summarization by Clustering of Documents*", October 10 - 11 2014, ACM.

[10] Hammouda.K, Kamel.M, "*Efficient Phrase-Based Document Indexing for Web Document Clustering"*,vol. 16, no. 10, october 2004 IEEE.

[11] Chaudhari,A, Kulkarni.P, Momin.B, "*Web Document Clustering Using Document Index Graph*", conference on advance Computing and Communication, pp. 32-37,20-23 Dec 2006 ,IEEE.

[12] BokI, Chung.S, Dongun, Lee.S, Lee.W, Ryu.H, "*Selection of Cluster Hierarchy Depth in Hierarchical Clustering using K-Means Algorithm*", conference on Information Technology Convergence,pp. 27-31,23-24 Dec 2007, IEEE.

[13] Hammouda.K, Kamel.M, "*Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization*", vol. 21, no. 5, may 2009 IEEE.

[14] Ambedkar.B, "*A Wordsets based document clustering algorithm for large datasets"*, conference on Methods and Models in Computer Science ,pp. 1-7,14-15 Dec 2009, IEEE.

[15] Murugesan.k, Zhang.J, "*Hybrid hierarchical clustering"*, pp. 1755-1760, 2011 IEEE.

[16] Meena.S, "*Dynamic Peer-to-peer Distributed Document Clustering and Cluster Summarization*", conference on sustainable energy and Intelligent Systems , pp. 815-819,July 20-22, 2011, IEEE.

[17] Chen.C.L, Liang.T ,Tseng.F, "*Hierarchical Document Clustering Using Fuzzy Association Rule Mining*",conference on Innovative Computing Information and Control, 18-20 June 2008 ,IEEE.

[18] Cao.T, Do.H, Hong.D, "*Quan, Fuzzy Named Entity Based Document Clustering*", 2008 IEEE.

[19] Cheng.H Chen.W, Fang.Q, Zhang.Z ,"*Clustering Aggregation Based on Genetic Algorithm for Documents Clustering*, conference on Evolutionary Computation, pp. 3156-3161, 2008, IEEE.

[20] Muflikhah.L, "*Document Clustering using Concept Space and Cosine Similarity Measurement*", Conference on Computer Technology and Development, pp. 58-62, vol-1, 2009, IEEE.

[21] Feng.W, Gu.J, "*Efficient Semisupervised Medline DocumentClustering with MeSH-Semantic and Global-Content Constraints*", pp. 1265-1276, vol-3, 2013, IEEE.