FUTURE RESEARCH PROSPECTS AND A SYSTEMATIC ANALYSIS OF EXISTING DOCUMENT CLUSTERING ALGORITHMS

Mayank Sharma
Research Scholar
Department of Computer Science
OPJS University

Dr. Kalpana Midda
Research Guide
Department of Computer Science
OPJS University

ABSTRACT

By dividing a vast number of unorganized text documents into a select few meaningful and coherent clusters, clustering is an effective method for creating the foundation for user-friendly navigation and browsing tools. Given that it has a wide range of applications in things like web mining, search engines, and information extraction, it is investigated at a broad level by researchers. The papers are grouped together based on several criteria of similarity. Numerous clustering techniques have so far been put out in the literature, however they are not thoroughly examined. In order to identify potential research trajectories in this area, this study will conduct a comprehensive review of the literature on document clustering techniques.

Keywords*-Document clustering, similarity measures, research direction*

1. Introduction

Internet users are rising daily, and the amount of information on the Web is expanding dramatically. It's becoming harder and harder to get relevant information online. Numerous data mining strategies have been put out in the literature to solve this problem. The practice of removing hidden, previously undiscovered, and usable information from data is known as data mining. Document clustering is a sub-problem of data mining that involves grouping documents into groups that, depending on the similarity metric being used, have certain characteristics in common. The user's ability to explore, summarize, and arrange the information rapidly is greatly aided by the document clustering methods. Document clustering may be used to categorize various documents, find duplicate content, suggest Web sites to users, improve search results, and more. This work seeks to develop research paths in this area by doing a comprehensive analysis of the currently used document clustering approaches..

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 the process of filtering, punctuation and special characters are eliminated from plain text documents. In tokenization, sentences are broken down into individual words or tokens. Stop words are eliminated from a text and reduced to their root form during stemming. In pruning, very infrequent words are eliminated from the resultant dataset.

(b)Feature extraction: From the dataset, a list of keywords is retrieved, and a feature vector is then produced by removing the keywords whose frequency exceeds a threshold.

(c) Document representation: To represent the text documents, a model is required. Numerous similar models, including tf-idf and tf probe, have been put out in the literature.

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

(d) Clustering: It classifies the documents into clusters using some similarity measurementmentioned above.

3. Research Methodology

First, research articles on document clustering will be gathered from a variety of sources, including ACM, Springer, IEEE, and others. On the basis of the clustering approach used in the article, these research papers will then be divided into several groups. Following that, sample research articles from each class will be evaluated critically, and research paths will be determined.

4. Document Clustering Techniques

The collection of 19 research articles was divided into three groups: partitioning algorithms, hierarchical algorithms, and soft- computing techniques. In the category of partitioning algorithms, which conducts a worldwide search throughout the whole solution space, eight research publications were discovered. Then harmony clustering is merged with the k-means technique to obtain better clustering. The suggested techniques made the k-means algorithm more stable by reducing its reliance on the starting parameters, such as the randomly selected beginning cluster centers. According to experimental findings, the suggested algorithms could locate better clusters than K-means, and the clusters' quality was equivalent and they reached the best known optimum more quickly.

Shameem et.al [3] They discovered another issue with the K-means algorithm. The classification of the documents in separate datasets was unsuccessful. They presented a method to measure the first estimation of the centroid locations for K clusters in order to address this issue. The documents were represented by the vector space model, and the most different K documents could be identified by applying certain dissimilarity measuring methods to the document collection. The document should then be classified in K distinct datasets by using K points as the K centroid. Nine research papers under hierarchical algorithms and four under soft computing algorithms. The research papers under each category are discussed and analyzed below.

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] By representing each document in vector format and applying the discrete cosine transform on it, the suggested document compression approach lowers the run-time memory need. The suggested technique lowered the RAM need to 60%.

Forsati et.al [2] suggested k-means clustering method A local optimum solution might be produced via the k-means method. In order to cluster texts using the harmony search optimization approach, they introduced unique harmony search clustering methods. They originally suggested a pure harmony search-based clustering method that quickly locates close to global ideal clusters by characterizing clustering as an optimization problem. The harmony search clustering method, in contrast to the localized searching of the K-means algorithm

Daling et.al [4] K-Means, the most widely used document clustering technique, has the flaw of its cluster intra dissimilarity. They suggested a K-Means method that was improved. In order to limit the effect of additional items on the means during the K-Means assignment step, the SOM scalar factor was included into the means. The enhanced K-Means method significantly minimizes the intra-dissimilarity of clusters by having greater F- Measure and less Entropy of clustering than the regular K-Means algorithm, according to experiments.

Mabel Rani et,al [5] They discovered one additional k-means difficulty. The issue was that the K-means method may converge to local optimum and was sensitive to the initial partition choice. They employed Improved Particle Swarm Optimization to tackle this issue (IPSO). In IPSO, the algorithm received training data one at a time and was trained using an incremental one-shot technique. When compared to current K-means, the suggested approach produced more precise and superior clustering results.

Guran et.al [6] They suggested a solution that lowered the amount of processed data and the K-means These approaches construct clusters by recursively splitting instances either from the top down or the bottom up. The exemplary publications for hierarchical algorithms are evaluated and discussed in the next section.

Naveen et.al [9] In response to a user query, Information Retrieval (IR) systems such as search engines extract a massive collection of documents, photos, and videos. This information burden is reduced by computational approaches such as Automatic Text Summarization (ATS), allowing users to discover information fast without reading the original text. Both the temporal complexity and the accuracy of summarization are obstacles for ATS. To address this problem, they devised an Information Retrieval system with three

clustering algorithm's execution time. They used dimension reduction techniques based on NMF. Every document was subject to the text-summarization procedure.

Ranjana Agrawal et.al [7] They suggested a solution that lowered the amount of processed data and the K-means clustering algorithm's execution time. They used dimension reduction techniques based on NMF. Every document was subject to the text-summarization procedure.

Pramod Bid [8] He discovered the remedy for overclustering. They introduced the Improved Document Clustering technique to overcome this issue, which produces a number of clusters for all text documents and uses cosine similarity measurements to arrange comparable documents in the appropriate clusters. They used the 20newsgroup dataset for experimental research. The suggested algorithm has a higher F1 score than current approaches.

Hierarchical Algorithms

distinct phases: retrieval, clustering, and summarization. During the Clustering phase, they adapt the Potential-based Hierarchical Agglomerative (PHA) clustering technique into a hybrid PHA-Clustering Gain-K-Means clustering strategy. The experimental findings shown that the suggested method improves the effectiveness and precision of clusters in comparison to both the traditional Hierarchical Agglomerative Clustering (HAC) algorithm and PHA.

Hammouda et.al [10] offered an approach for phrase clustering Phrase-based analysis implies that the similarity of texts should be determined by matching phrases, as opposed to individual words.

B. F. Momin et.al [11] suggested an approach for phrase clustering in documents that is more accurate than previous methods. They proposed a Document Index Graph (DIG) model that explained the efficacy of phrase-based similarity over term-based similarity, and then they proposed a Document Index Graph based Clustering (DIGBC) algorithm to enhance the DIG model for incremental and soft clustering algorithm to efficiently cluster documents. Due to the inaccuracy of the document-cluster similarity computation and the selection of the threshold value, the DIGBC method had a subpar level of quality.

Lee et.al [12] In order to minimize the temporal complexity of hierarchical clustering techniques, they merged the two methodologies to create the CONDOR system, which has a hierarchical structure based on document clustering using the K-means algorithm. Performance-wise, the suggested solution was inefficient, and it was only suitable to little Web data.

Hammouda et.al [13] The answer to the modularity, flexibility, and scalability challenge was presented. They presented a hierarchically distributed Peer-to-Peer (HP2PC) architecture and clustering technique to overcome this issue. The design was built on a peer-to-peer multilayer overlay network. The suggested strategy prevented centroids from traversing neighborhoods through higher levels.

Ambedkar [14] He introduced WDC (Word sets- based Clustering), an efficient clustering technique based on closed word sets to deal with the extremely high dimensionality of the text, the very big quantity of the datasets, and the comprehension of the Cluster description. WDC used a hierarchical strategy to group text documents with similar phrases. WDC was proven to be more scalable, effective, and efficient than current clustering methods such as K-means and its modifications.

Murugesan et.al [15] In order to improve upon the existing K-means divisive hierarchical method and the agglomerative hierarchical clustering algorithm (UPGMA), a hybrid clustering algorithm was devised. This algorithm combines the best features of both approaches. Each document was given its own weights based on the phrases used, and these weights were then employed in a vector space model.

Meena et.al [16] They put out the idea of cluster summarizing very large text texts. The Dynamic Peer to Peer (P2P) Document Clustering and Cluster Summarization (DP2PCS) architecture, based on bonus words and stigma words, was suggested to do this. The main drawback of this method was the low degree of security.

Selangor et.al [17] They put out a fresh textual document clustering approach, which was employed to cluster data with excellent effectiveness and quality. To enhance the clustering quality, the Fuzzy Frequent Item-set Based Hierarchical Clustering approach (F2IHC) [16] was used. It was simple to create a linkage and incorporate linguistic phrases employing fuzzy association rules mining. The effectiveness and quality weren't very excellent.

4.3 Soft Computing Algorithms

Soft computing tolerates imprecision, uncertainty, partial truth, and approximation, unlike hard computing. Exploit imprecision, uncertainty, partial truth, and approximation to provide tractability, resilience, and cheap solution cost. This section analyzes and discusses many related works.

Truh.cao et.al [18] They detected the issue with KBDC (keyword based document clustering). KBDC has limitations because to its rudimentary word processing and difficult cluster separation. In order to resolve this issue, they implemented named entities as the primary components determining document semantics. The issue with this method was that its entropy and f-measure were subpar.

Zang et.al [19] GeneticCa, a technique based on a genetic algorithm, was developed to increase cluster aggregation performance. This approach only worked with bit strings.

Lailil [20] The mechanism they proposed was known as the Latent Semantic Index (LSI) approach. It used the Singular Vector Decomposition (SVD) or Principal Component Analysis concepts (PCA) The purpose of this strategy was to lower the matrix dimension by discovering a pattern in a collection of documents that relates to concurrent phrases. Each approach was applied to weight term-document in the vector space model (VSM) for document clustering using the fuzzy c-means algorithm.

Feng et.al [21] A new semi supervised spectral clustering technique called as SSNCut for clustering over the local cost (LC) similarities improved the performance of Medical Literature Analysis and Retrieval System Online (MEDLINE) articles. Two different kinds of constraints were worked on: cannot-link (CL) constraints on document pairings with less similarities in MeSH semantics and must-link (ML) constraints on document pairs with more similarities. The performance of SSNCut was this paper's flaw.

5. Research Directions

Document clustering is utilized in a variety of domains, and several authors have developed various techniques to address constraints including complexity and runtime memory needs. Many issues were found after reading numerous research articles on partitioning methods, such as the absence of any proposals for document clustering based on semantics. In the future, researchers may create clustering algorithms based on text document semantic analysis. Future strategies may be offered to reduce the height of the tree since it is a key bottleneck in hierarchical based algorithms. The entropy and f-measure of the soft computing algorithms that have been suggested so far are inadequate. Some techniques solely use bit strings. Researchers are required to solve these problems since some are based on local cost, MeSH semantic similarities, and must link and cannot link restrictions have poor performance.

6. Conclusion

The authors of this study have conducted a thorough analysis of document clustering techniques that have been disclosed in the literature. The constraints of the algorithms within these categories have been determined by categorizing them into the three major groups. The researchers may benefit greatly from the proposed routes for further study in this area.

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 - 3352008,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

[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 2009IEEE.

[14] Ambedkar.B, "A Wordsets based document clustering algorithm for large datasets", conference on Methods and Models in Computer Energy and Intelligent Systems pp. 27-29 Dec. 2012 IEEE.

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

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 NamedEntity 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.

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

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