ISSN NO: 2347-4890
International Journal of software & Hardware Research in Engineering
Volume 1 Issue 2, October 2013

Study on Different Clustering Approaches for Document Clustering

J.Sankari

Research Scholar,
Department of Computer Applications,
K.S.R angasamy College of Arts and Science ,
Tiruchengode-637215, India
jgksankari@gmail.com

Dr. R Manavalan

Department of Computer Applications,
K.S.R angasamy College of Arts and Science ,
Tiruchengode-637215, India
Manavaln_r@yahoo.com

ABSTRACT

*Clustering is an unsupervised learning data mining technique. The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad demand and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. The principle of clustering depends on the concept of Distance Metric or Similarity Metric. Because of the variety of feature types and scales, the distance measure (or measures) must be chosen carefully. T*his paper represents a survey of clustering techniques in Text mining. The clustering techniques are categorized based upon different approaches. *The paper ends by addressing some important issues and open questions that can be subject of future research.*

Key Terms: Clustering, data mining, Distance Metric, Similarity Measures, Text Mining

1. INTRODUCTION

A literature review is a description of the literature relevant to a particular field or topic. It gives an overview of what has been said, who the key writers are, what are the prevailing theories and hypotheses, what questions are being asked, and what methods and methodologies are appropriate and useful. As such, it is not in itself primary research, but rather it reports on other findings.

2. DOCUMENT CLUSTERING: A REVIEW

Document clustering is a technique for identifying clusters or groups of documents which share some common features or have overlapping content. These groupings of documents can be useful in document retrieval. Document clustering has been used in experimental Information Retrieval (IR) systems for decades. It was initially proposed as a means for improving efficiency and also a way to categorize or classify documents [1] [2]. Clustering documents are part of a search process. The deep study on document clustering is provided in section 3.

3. STUDY ON DOCUMENT CLUSTERING

In 2000, Alexander Strehl, Joydeep Ghosh, and Raymond Mooney introduced Impact of Similarity Measures on Web-page Clustering [3]. Four popular similarities measures (Euclidean, cosine, Pearson correlation and extended Jaccard) in conjunction with several clustering techniques (random, self-organizing feature map, hyper-graph partitioning, generalized k- means, weighted graph partitioning), were compared on high dimensional sparse data representing web documents. Performance measured against a human-imposed classification into news categories and industry categories. Number of experiments was conducted and t-tests used to assure statistical significance of results. Metric distances such as Euclidean are not

appropriate for high dimensional, sparse domains. Cosine, correlation and extended Jaccard measures are successful in capturing the similarities. The clustering results indicated by accuracy.

Document Clustering Based on Nonnegative Matrix Factorization developed by Wei Xu, Xin Liu and Yihong Gong in 2001 [4]. A document Clustering method based on the Nonnegative Factorization of the term document matrix of the given document corpus was proposed and achieved 72% of accuracy. The proposed document clustering method surpasses the latent semantic indexing and the spectral clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies.

Spectral Relaxation for K-means Clustering was introduced by Hongyuan Zha and Xiaofeng He in 2001 [5]. To tackle the problem from a different angle: an equivalent formulation of the sum-of-squares minimization as a trace maximization problem with special constraints. The Relaxing the constraints, leads to a maximization problem that possesses optimal global solution was used to achieve 67% of accuracy. It was actually pretty tricky to compute the accuracy using the confusion matrix because they do not know which cluster matches which newsgroup category.

Inderjit S. Dhillon presented Co-clustering documents and words using Bipartite Spectral Graph Partitioning in 2001 [6]. The document collection as a bipartite graph between documents and words, using which the simultaneous clustering problem can be proposed as a bipartite graph partitioning problem. To solve the partitioning problem, spectral co-clustering algorithm used with the second left and right singular vectors of an appropriately scaled word-document matrix to yield good by partitioning. This is in stark contrast to the spherical k-means algorithm that gave poor results on small document collections.

In 2003, Inderjit S. Dhillon et al. proposed an Information-Theoretic Co-clustering [7]. An innovative co-clustering algorithm that monotonically increased the preserved mutual information by intertwining both the row and column clustering's at all stages was presented by an accuracy of 96%.

An Efficient Online Spherical K-means Clustering method was proposed by Shi Zhong in 2004. The spherical k-means algorithm, i.e., the k-means algorithm with the cosine similarity method used for clustering high-dimensional text data [8]. Each document as well as each cluster means represented as a high-dimensional unit-length vector. However, it has been mainly used in batch mode. Each cluster means vector updated, only after all document vectors being assigned, as the (normalized) average of all the document vectors assigned to that cluster. An efficient online spherical k-means algorithm was investigated, which employs the "Winner Take-All" competitive learning technique, together with an annealing-type learning rate schedule.

Clustering objects on Subsets of Attributes in 2004 proposed by Jerome H. Friedman and Jacqueline J. Meulmany [9] for clustering attribute value data for increasing sensitivity for detecting especially low cardinality groups clustering on a small subset of variables. The validity and usefulness of the output of different clustering methods are evaluated by the user in the context of each particular application.

A Survey of Correlation Clustering was discussed by Hila Becker in 2005. The problem of partitioning a set of data points into clusters found in many applications [10]. Correlation clustering is a clustering technique motivated by the problem of document clustering, in which given a large corpus of documents such as web pages, This was used to find their optimal partition into clusters. While most commonly used clustering algorithms such as k-means, k-clustering some and k-center require prior knowledge of the number of clusters that used to divide the data into web documents, finding the number of clusters not a trivial task. Correlation Clustering, introduced by Bansal, Blum and Chawla, provides a method for clustering a set of objects into the optimal number of clusters, without specifying that number in advance.

Similarity Measures for Text Document Clustering was introduced by Anna Huang in 2008.

Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms [11]. Partitional clustering algorithms have been recognized to be more suitable as opposed to the hierarchical clustering schemes for processing large data sets. A wide variety of distance functions and similarity measures have been used for clustering, such as squared Euclidean distance, cosine similarity, and relative entropy. The effectiveness of these measures in Partitional clustering for text document data sets were compared and analyzed. Experiments use the standard K-means algorithm and reported the results on seven text document data sets and 5 distance/similarity measures commonly used in text clustering.

Hans-Peter Kriegel et al. proposed a General Framework for Increasing the Robustness of PCA-based Correlation Clustering Algorithms in 2008. Most correlation clustering algorithms rely on principal component analysis (PCA) as a correlation analysis tool [12]. The correlation of each cluster is learned by applying PCA to a set of sample points for outlier analysis.

Dino Ienco, Ruggero G. Pensa, and Rosa Meo in 2009, introduced Context-based Distance Learning for Categorical Data Clustering [13] to learn a context-based distance for categorical attributes. This method was competitive categorical data clustering approaches in the state of the art. It does not investigate the application of our distance learning approach to different distance based tasks such as: outlier detection and nearest neighbor classification. The results were presented in the Normalized Mutual Information and accuracy like 89% and 93% respectively.

Analyzing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes was presented by Parul Agarwal, M. Afshar Alam and Ranjit Biswas in 2010. Hierarchical Clustering is an important technique of data mining, groups similar objects together and identifies the cluster to which each object of the domain being studied belongs to [14]. It provided in-depth explanation of implementation adopted for k-pragna, an agglomerative hierarchical clustering technique for categorical attributes.

Ontology based Similarity Measure in Document Ranking was developed by Sridevi.U.K in 2011 for ontology based semantic annotation of web pages with annotation weighting scheme that takes advantage of the different relevance of structured document fields [15]. The retrieval model was based on the importance factors of the structural elements, which used to re-rank the document retrieval by the ontology based distance measure. The relevance concept similarity combined with the annotation-weighting scheme to improve the relevance measures. It has been evaluated on USGS Science directory collection. Preliminary experiment results shown, it generated relevant document in the top rank.

Clustering with Multiviewpoint-Based Similarity Measure was proposed by Duc Thang Nguyen, Lihui Chen and Chee Keong Chan in 2012 [16]. The similarity between a pair of objects defined either explicitly or implicitly. Multi Viewpoint-based similarity measure and two related clustering methods were introduced. Using multiple viewpoints, more informed assessment of similarity could be achieved. Theoretical analysis and empirical study conducted to support this claim. Two criterion functions for document clustering proposed based on this measure. Comparison was made with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the reward of the proposed approach.

Table 1: Survey on document clustering

Method	Description	Author	Result	Year
Similarity Measures [3].	Cosine, correlation and extended Jaccard measures are successful in capturing the similarities.	Alexander Strehl, et al	89%	2000

Non-negative Matrix Factorization [4].	Surpasses the latent semantic indexing and the spectral clustering methods	Wei Xu, Xin Liu and Yihong Gong	72%	2001
Spectral Relaxation [5].	An equivalent formulation of the sum-of-squares minimization as a trace maximization problem with special constraints.	Hongyuan Zha & Xiaofeng	67%	2001
Bipartite Spectral Graph Partitioning [6].	Modeling the document collection as a bipartite graph between documents	Inderjit S. Dhillon	64%	2001
co-clustering algorithm [7].	Monotonically increases the preserved mutual information	Inderjit S. Dhillon et al	96%	2003
Spherical K-means Clustering method [8]	An efficient online spherical k-means algorithm, which employs the "Winner Take-All" competitive learning technique.	Shi Zhong	78%	2004
Different clustering methods [9].	Clustering attribute - value data for increasing sensitivity	Jerome H. Friedman and Jacqueline J. Meulmany	56%	2004
Correlation Clustering [10].	Two different approximation algorithms for the Correlation Clustering problem.	Hila Becker	89%	2005
Partitional Clustering Algorithm [11].	compare and analyze the effectiveness of these measures in Partitional clustering for text document datasets	Anna Huang	91%	2008
Correlation Clustering [12].	Most correlation clustering algorithms rely on principal component analysis (PCA) as a correlation analysis tool.	Hans-Peter Kriegel et al.	57%	2008
DILCA (DIstance Learning in Categorical Attributes) Method [13].	This method is competitive categorical data clustering approaches in the state of the art.	Dino Ienco et al.	89% and 93%	2009
Agglomerative hierarchical Clustering Algorithm [14].	Provide in depth explanation of implementation adopted for k-pragna, an agglomerative hierarchical clustering technique for categorical attributes.	Parul Agarwal, M. Afshar Alam and Ranjit Biswas	92%	2010
Ontology based Similarity Measure [15].	The ontology based semantic annotation of web pages with	Sridevi. U.K	84%	2010

	annotation weighting scheme.			
Multiviewpoint-Based Similarity Measure, Spherical K-Means clustering algorithm [16].	more informative assessment of similarity could be achieved. Information from multiple documents from different categories is clustered	Duc Thang Nguyen et al.	67% and 73%	2012

3. CONCLUSION

This survey projected various clustering approaches and algorithms in document clustering. The areas of document clustering have many issues, which need to be solved. We hope, the paper gives interested readers a broad overview of the existing techniques. As a future work, improvement over the existing systems with better results which offer new information representation capabilities with different techniques like search result clustering, collection clustering and co clustering can be attempted.

4. REFERENCES

1. Estivill-Castro, V. (2002). "Why so many clustering algorithms". ACM SIGKDD Explorations Newsletter 4: 65. doi:10.1145/568574.568575.
2. Nicholas O. Andrews and Edward A. Fox, Recent Developments in Document Clustering, October 16, 2007 .
3. Alexander Strehl, Joydeep Ghosh, Raymond Mooney, "Impact of Similarity Measures on Web-page Clustering", 2000.
4. Wei Xu, Xin Liu, Yihong Gong, "Document Clustering Based On Non-negative MatrixFactorization", 2001
5. Hongyuan Zha & Xiaofeng He, "Spectral Relaxation for K-means Clustering",2001.
6. Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning in", 2001, KDD 2001 San Francisco, California, USA Copyright 2001 ACM.
7. Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha, "InformationTheoretic Coclustering", August 2427, 2003, Washington, DC, USA. Copyright 2003
8. Shi Zhong, "Efficient Online Spherical K-means Clustering", 2004
9. Jerome H. Friedman_ Jacqueline J. Meulmany, "Clustering Objects on Subsets of Attributes", 2004, http://www-stat.stanford.edu/~jhf/ftp/cosa.pdf.
10. Hila Becker, "A Survey of Correlation Clustering", May 5, 2005, COMS E6998: Advanced Topics in Computational Learning Theory.
11. Anna Huang, "Similarity Measures for Text Document Clustering", NZCSRSC 2008, April 2008, Christchurch, New Zealand.
12. Hans-Peter Kriegel , Peer Kröger , Erich Schubert , Arthur Zimek, "A general framework for increasing the robustness of PCA-based correlation clustering algorithms (2008) , www.dbs.informatik.uni-muenchen.de
13. Dino Ienco, Ruggero G. Pensa, and Rosa Meo, "Context-based Distance Learning for Categorical Data Clustering", 2009.
14. Parul Agarwal, M. Afshar Alam, Ranjit Biswas, "Analysing the agglomerative hierarchicalClustering Algorithm for Categorical Attributes", International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010 ISSN: 2010-0248.
15. Sridevi.U.K, Nagaveni .N, "Ontology based Similarity Measure in Document Ranking", ©2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 26.
16. Duc Thang Nguyen, Lihui Chen and Chee Keong Chan , "Clustering with Multiviewpoint-Based Similarity Measure" 2012*ieeexplore.ieee.org*

ISSN NO: 2347-4890
International Journal of software & Hardware Research in Engineering
Volume 1 Issue 2, October 2013

