International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.102.26

Document clustering – Application of PCA and K-means on Domain Ratio Tables

Padmaja Ch V R[1], Lakshminarayana S[2], Divakar Ch[3]
[1]*Dept. of Computer Science & Engineering, Raghu Engineering College, Visakhapatnam, AP, India*
[2]*Former Principle Scientist, National Institute of Oceanography, Visakhapatnam, AP, India*
[3]*Professor, Dept. of Information Technology, SRKR Engineering College, Bhimavaram, AP, India*

Abstract— These days, using internet is becoming a usual and almost daily activity of various people. This leads to the availability of huge amounts of electronic textual data in the form of mails, tweets, blogs, research articles, new articles and so on. Grouping this collection into meaningful groups is a challenging task. Document clustering is one research area which facilitates the organization of documents into different groups based on their similarity. It is an unsupervised learning method and having major applications like finding similar documents, organizing large document collections, duplicate content detection, search optimization. In this study, we tried to cluster research articles of various authors using PCA and K-means algorithms and found encouraging results.
Keywords-Document clustering; PCA, K-means.

I. INTRODUCTION

There is a tremendous growth in the volume of the text documents available on the internet, news sources, social media articles like blogs, tweets and mails due to the advancement of technology. Organizing this massive collection of text data is challenging task for the researchers. Data mining is the process of extracting the implicit, previously unknown and potentially useful information from the data [1] whereas, text mining is the process of extracting the useful information from the unstructured text data [2].

Organizing text data manually is very difficult task [3] for obvious reasons. Document clustering is the process of organizing documents into different groups based on similarity measure. These groups are commomly known as clusters [4]. In document classification, class labels are used for classifying the documents which comes under supervised machine learning but, in document clustering no class labels are used to classify the text. Document clustering comes under unsupervised machine learning algorithms.

Finding similar documents, organizing large document collections, recommendation systems, duplicate content detection and search optimization are the main applications of document clustering. The major challenges in document clustering are high dimensionality, scalability, selection of appropriate features of the document and similarity measures between documents, giving meaningful cluster labels and accuracy.

This study focuses on reducing the dimensionality by means of Domain Ratio Based Allocation (DRBA), which is used for document classification in our previous work [5]. Using DRBA approach, the domain ratio tables were obtained. The clustering methods PCA and K-means algorithms were applied on these domain ratio tables and analyzed their performance. In this paper, section 2 addressees literature survey, section 3 document representation, section 4 application of PCA and K-means, results in section 5 and conclusion in section 6.

II. RELATED WORK

Several studies provide information about various algorithms on document clustering [6]. Document clustering is divided into two classes hard and soft clustering methods [7]. In hard clustering, each document is assigned to exactly one cluster, forming disjoint clusters. In soft clustering, each document can appear in multiple clusters, resulting cluster overlapping. The soft clustering is again divided into partitioning, hierarchical, and frequent itemset-based clustering. In the partitioning clustering, the documents are grouped into a fixed number of non-empty clusters. K-means and it variants are the examples for partitioning clusters. Hierarchical clustering follows tree-based structure known as dendrograms. The well-known hierarchical clustering methods are Hierarchical Agglomerative Clustering (HAC) and Unweighted Pair Group Method with Arithmetic mean (UPGMA).

Hierarchical clustering methods are classified into two categories: agglomerative and divisive methods. Agglomerative methods follow bottom up approach in which, each object forms a cluster initially and most similar clusters are combined iteratively until some condition is satisfied to terminate. In contrast, the divisive method follows top-down approach where all objects are considered into a single cluster and then split into smaller clusters recursively until some termination criteria is occurred.

Frequent itemset-based use frequent item sets generated by the association rule mining to cluster the documents. These methods reduce the dimensionality of term features efficiently for large data sets and helpful in labelling the clusters by the obtained frequent item sets. Some of these methods are Hierarchical Frequent Term-based Clustering (HFTC) [8], Hierarchical Document Clustering using Frequent Itemsets (FIHC) [9], and Fuzzy Frequent Itemset-based Document Clustering (F^2IDC) [10].
The high dimensional data can be transformed into lower dimensional data using Principal Component Analysis (PCA) [11] and then k-means can be applied [12]. This approach is used in this study on domain ratio tables to form clusters.

III. METHODOLOGY

3.1 Document Pre-processing

Text document can be represented in many ways. The most common way is to represent in bag-of-words approach [13]. The data set is considered from our previous work [5] as shown in the following Table 1. All the publications are converted into text files for analyzing. The pre-processing steps like tokenizing, converting into lower case, removing numbers and punctuations and removal of stop words are carried out on the corpus. Next, word stemming is applied to bring the words to the root form and sparse words are also eliminated. The corpus is then transformed into Document Term Matrix (DTM) [] which is a word-document frequency table. A DTM is very high dimensional one. To reduce this dimensionality, Domain Ratio Tables are calculated [5].

Table 1Publications Collection

S NO	Institution	Number of Publications collected
1	Institution 1	409
2	Institution 2	465
3	Institution 3	311

3.2. PCA

Principal Component Analysis (PCA) [11] is a statistical method. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. These uncorrelated variables are called principal components. The number of principal components should not exceed the number of original variables. In this approach, the largest possible variance is determined by the first principal component.

3.3 K-means

Among various clustering algorithms, k-means is the most significant and simple one to implement [14]. The goal of K-means is to find the minimum of the following function

$$F = \sum_{i=1}^{K} \sum_{x \in C_i} d(C_i, x)^2, \quad C_i = \frac{1}{|C_i|} \sum_{x \in C_i} x, \tag{1}$$

Where C_i is the i-th cluster, c_i its centroid, and d a distance function.

IV. RESULTS & DISCUSSION

As the principal components are eigenvectors of the covariance matrix which is symmetric, they are orthogonal. However, this noise reduction property alone is inadequate to explain the effectiveness of PCA. Applying PCA to Domain Ratio Table of Author5, the publications can be classified as shown in the Figure 1. The same experiment is carried out for all publications in the data set shown in Table 1 and the result is visualized through the Figure 2.

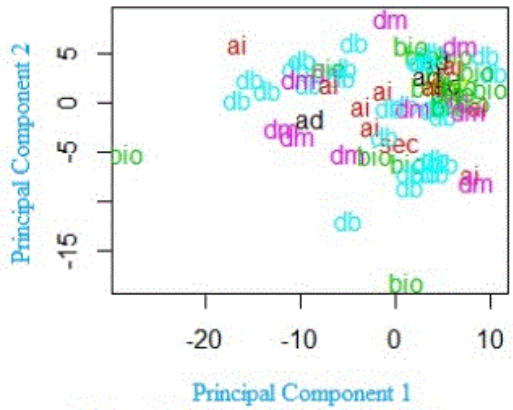

Figure. 1: PCA result for Domain Ratio Table of author 5 publications [5]

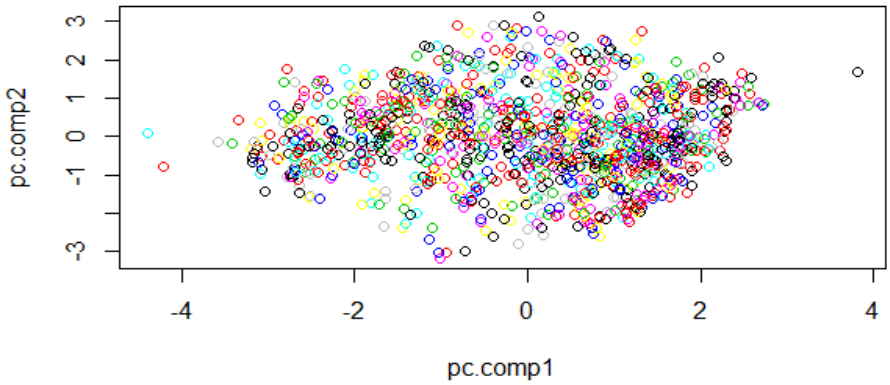

Figure 2: PCA result for the Domain Ratio Table of all publications in the data set

In this work, we explore the connection between these two widely used methods. We tried to prove that principal components are the continuous solution of the cluster membership indicators in the K-means clustering method, i.e., the PCA dimension reduction automatically performs data clustering according to the K-means objective function. K-means uses K samples to express the data in terms of the centroids of clusters. They are calculated by minimizing the sum of squared errors,

$$J_k = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2 \quad (2)$$

where $(x_1, \ldots x_n) = X$ is the data matrix and $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of cluster C_k and n_k is the number of points in C_k.

Standard iterative solution to K-means suffers from a well-known problem: as iteration proceeds, the solutions are trapped in the local minima due to the greedy nature of the update algorithm. Next, we apply K-means clustering in the PCA subspace and shown in the Figure 3.

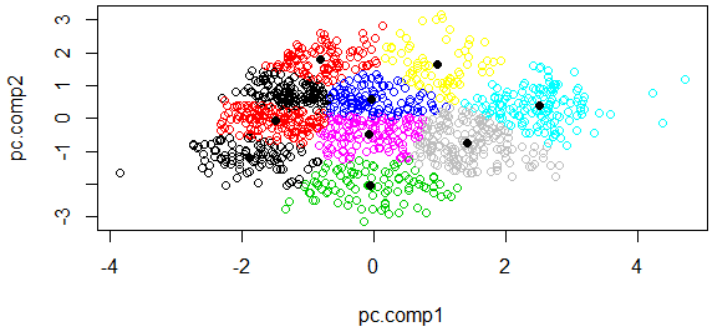

Figure 3: K-means clustering using Principle Components of Domain Ratio Table

V. CONCLUSION

Document clustering is one of the most significant research area in recent years, as it reduces the human effort in organizing them for future reference. It is a trivial fact that every document is very high in dimensions and extracting the theme of it is a complex task. In this study, we transformed the collection of documents into domain ratio table using Domain Ratio Table Allocation method, which uses some external knowledge as citation-based keywords. This domain ratio table contains numerical data which gives us the flexibility to apply any statistical methods or machine learning algorithms for clustering.

In this work, we explore the connection between these two widely used methods. We tried to prove that principal components are the continuous solution of the cluster membership indicators in the K-means clustering method, i.e., the PCA dimension reduction automatically performs data clustering according to the K-means objective function.

VI. REFERENCES

[1] J. Han, M. Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. Pages 335 – 388 and 428 – 435, 2001.

[2] Miner G, Elder J, Hill T, Nisbet R, Delen D, Fast A (2012) Practical text mining and statistical analysis for non-structured text data applications. 1st edn. Academic Press, Boston.

[3] Farial Shahnaz and Michael W. Berry. March 2006. Document Clustering Using Non-Negative Matrix Factorization. Information Processing and Management: An International Journal, Volume 42 Issue 2, Pages 373-386.

[4] Rekha Baghel and Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm," International Journal of Computer Applications, vol. 4, No.5, pp. 0975 – 8887, Jul. 2010

[5] Padmaja Ch V R, Lakshmi Narayana S, Divakar Ch., "Identifying the Research Specialization from the Publications using Text Mining and Linear Discriminant Analysis", International Journal of Applied Engineering Research, Volume 13, pp 6667-6672, 2018.

[6] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," Mining Text Data, Springer US, 2012, pp. 77-128.

[7] Chun-Ling Chen, Frank S.C. Tseng, and Tyne Liang, "An integration of WordNet and fuzzy association rule mining for multi-label document clustering," Data and Knowledge Engineering, vol. 69, issue 11, pp. 1208-1226, Nov. 2010

[8] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," Proc. of Int'l Conf. on knowledge Discovery and Data Mining (KDD'02), pp. 436–442, 2002.

[9] Benjamin C.M. Fung, Ke Wang, and Martin Ester, "Hierarchical Document Clustering Using Frequent Itemsets," In Proc. Siam International Conference On Data Mining 2003, SDM 2003

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

[11] Jolliffe, I. (2002). Principal component analysis. Springer. 2nd edition.

[12] Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for K-means clustering. Advances in Neural Information Processing Systems 14 (NIPS'01), 1057–1064.

[13] Chade-Meng Tan, Yuan-Fang Wang, Chan-Do Lee, The Effectiveness of Bigrams in Automated Text Categorization. ICMLA: 275-281, 2002.

[14] Satchidanandan, D., Chinmay, M., Ashish, G., Rajib, M.: A Comparative Study of Clustering Algorithms. Information Technology Journal, vol.5, p. 551-559, (2006)