



International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Document Clustering Using Enhanced Tw-K-Means

M.Kumaresan, G.Ashwitha, S.Bhuvaneshwari, N.Priya Dharshini, M.Udhaya

Assistant Professor, Dept. of CSE, Angel College of Engineering and Technology, Tamilnadu, India

ABSTRACT: This paper proposes TW-k-means, an automated two-level variable weighting clustering algorithm for multiview data, which can simultaneously compute weights for views and individual variables. In this algorithm, a view weight is assigned to each view to identify the compactness of the view and a variable weight is also assigned to each variable in the view to identify the importance of the variable. Both view weights and variable weights are used in the distance function to determine the clusters of objects. In the new algorithm, two additional steps are added to the iterative k-means clustering process to automatically compute the view weights and the variable weights.

I. INTRODUCTION

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others. Document clustering is a process that involves a computational burden in measuring the similarity between document pairs. Similarity measure is the function which assigns a real number between 0 and 1 to the documents. A zero value means that the documents are dissimilar completely whereas one indicates that the documents are identical practically.

SYNCLUS is the first variable weighting multiview clustering algorithm which uses weights for both views and individual variables in the clustering process. But it only computes variable weights automatically and the view weights are given by users. Recently, Tzortzis and Likas proposed a weighted combination of exemplar-based mixture models (WCMM) that assigns different weights to the views and learns those weights automatically, but their method does not consider variable weights. The two algorithms have a big weakness that they are not scalable to large data sets. In this paper, we propose TW-k-means, a novel twolevel variable weighting k-means clustering algorithm for multiview data. In the TW-k-means algorithm, to distinguish the impacts of different views and different variables in clustering, the weights of views and individual variables are introduced to the distance function. The view weights are computed from the entire variables, whereas the weights of variables in a view are computed from the subset of the data that only includes the variables in the view. Therefore, the view weights reflect the importance of the views in the entire data, while the variable weights in a view only reflect the importance of variables in the view. We present an optimization model for the TW-kmeans algorithm and introduce the formulae, derived from the model, for computing both view weights and variable weights. We define the TW-k-means algorithm as an extension to the standard k-means clustering process with two additional steps to compute view weights and variable weights in each iteration. Since the two steps do not require intensive computation, the new clustering algorithm remains efficient in clustering large high dimensional multiview data. Compared with SYNCLUS and WCMM, TW-k-means can automatically compute both view weights and individual variable weights. Moreover, it is a fast clustering algorithm which has the same computation complexity as k-means.

II. RELATED WORK

Text mining spans through various areas and has its applications including recommendation systems, tutoring, web mining, healthcare and medical information systems, marketing, predicting, and telecommunications to specify a few among many applications. The authors (Hussein Hashimi, Alaaeldin Hafez, & Hassan Mathkour, 2015), study and propose various criteria for text mining. These criteria may be used to evaluate the effectiveness of text mining



International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

techniques used. This makes the user to choose one among the several available text mining techniques. In (Yannis Haralambous & Philippe Lenca, 2014), the authors use the concept of text item pruning and text enhancing and compare the rank of words with the tf-idf method. Their work also includes studying the importance and extending the use of association rules in the text classification. Association rule mining is playing an important role in text mining and is also widely studied, used and applied by the researchers in text mining community. In (Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, & David Chek Ling Ngo, 2014) authors, discuss the importance of text mining in the predicting and analyzing the market statistics. In short, they perform a systematic survey on the applicability of text mining in market research. In (Sajid Mahmood, Muhammad Shahbaz, & Aziz Guergachi, 2014), the authors work towards finding the negative association rules. Earlier in the past decade, the data mining researchers and market analysts were only interested in finding the dominant positive association rules. In the recent years, significant research is carried out towards finding the set of all possible negative association rules. The major problem with finding negative association rules is the large number of rules which are generated as a result of mining. The negative association rules have important applications in medical data mining, health informatics and predicting the negative behavior of market statistics. In (Wen Zhang, Taketoshi Yoshida, Xijin Tang, & Qing Wang, 2010) the authors use the approach of first finding the frequent items and then using these computed frequent items to perform text clustering. They use the method called “maximum capturing”. With the vast amount of information generating in the recent years, many researchers started coming out with the extensive study and defining various data mining algorithms for finding association rules, IADIS International Journal on Computer Science and Information Systems obtaining frequent items or item sets, retrieving closed frequent patterns, finding sequential patterns of user interest (Ning Zhong, Yuefeng Li, & Sheng-Tang Wu, 2010). All these algorithms are not suitable for their use in the field of text mining because of their computational and space complexities. The suitability of these techniques in text mining must be studied in detail and then applied accordingly. One of the important challenges in text mining is handling the problems of misinterpretation and less frequency. An extensive survey on dimensionality reduction techniques is carried in (Fodor, I.K., 2002). The authors discuss the method of principal factor analysis, maximum likelihood factor analysis and PCA (principal component analysis). A fuzzy approach for clustering features and text classification which involves soft and hard clustering approaches is discussed in (Jung-Yi Jiang, Ren-Jia Liou, & Shie-Jue Lee, 2010). An improved similarity measure overcoming the dis-advantages of conventional similarity measures is discussed in (Yung-Shen Lin, Jung-Yi Jiang, & Shie-Jue Lee, 2014; Shie-Jue Lee, & Jung-Yi Jiang, 2014) their work also involves clustering and classification of text documents. In (Sunghae Jun, Sang-Sung Park, & Dong-Sik Jang, 2014), the concept of support vector machines, SVM is used for document clustering. The other significant findings and research works in the area of text mining include work by the researchers (Sofus A. Macskassy, & Haym Hirsh, 2003; Libiao Zhang, Yuefeng Li, Chao Sun, & Wanvimol Nadee, 2013; Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, & Dino Isa, 2012; Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, & Dino Isa, 2012; Dell Zhang & Wee Sun Lee, 2006; Wen Zhang, Taketoshi Yoshida, & Xijin Tang, 2008). In the present work, our idea is to design a similarity measure overcoming dis-advantages in Euclidean, Cosine, Jaccard distance measures (Yung-Shen Lin et al., 2014). The proposed measure considers distribution of features of the global feature set.

III. PROPOSED MODULE

- 1.Document/Dataset Classification
 - 1.1. Stemming Rule Implementation
 - 1.2. Cluster Preserving Indexing (CPI) Implementation
- 2.Frequency Calculation
- 3.Cluster Based Modeling

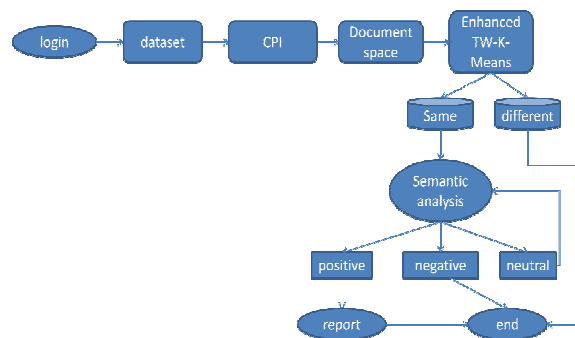
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

ARCHITECTURE DIAGRAM



IV. DESCRIPTION

Uploading forensic data set .perform Cluster preserving indexing(CPI) to Construct the local neighbour patch, and compute the matrices. Create document space .perform Enhanced TW-K-means to detect homogenous cluster and heterogenous cluster separately. By considering homogenous cluster semantic analysis process is performed to analysis positive ,negative and neutral data where separated out. Repeat sematic process for neutral data to purity of the document and generate the report .

V. ALGORITHM USED

For subsequent experiments, the standard Enhanced TW-K-means algorithm is chosen as the clustering algorithm. This is an iterative Partitional clustering process that aims to minimize the least squares error criterion . As mentioned previously, Partitional clustering algorithms have been recognized to be better suited for handling large document datasets than Hierarchical ones, due to their relatively low computational requirements .The standard Enhanced TW-K-means algorithm works as follows. Given a set of data objects D and a pre-specified number of clusters k , k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroids are recomputed for each cluster and in turn all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids. The generated clustering solutions are locally optimal for the given data set and the initial seeds. Different choices of initial seed sets can result in very different final partitions. Methods for finding good starting points have been proposed.However, we will use the basic K-means algorithm because optimizing the clustering is not the main focus of this paper. The K-means algorithm works with distance measures which basically aims to minimize the within-cluster distances. Therefore, similarity measures do not directly fit into the algorithm, because smaller values indicate dissimilarity.

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
- 3.compute view and variable weights automatically.
4. Recompute the centroid of each cluster.
5. Repeat steps 2 and 3 until the cluster form.



International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

VI. MODULE DESCRIPTION

DOCUMENT/DATASET CLASSIFICATION

- Document classification or document categorization is a problem in library science, information science and computer science.
- The task is to assign a document to one or more classes or categories and this may be done "manually" (or "intellectually") or algorithmically.
- The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is used mainly in information science and computer science.
- The problems are overlapping, however, and there is therefore also interdisciplinary research on document classification.

CLUSTER PRESERVING INDEXING (CPI) IMPLEMENTATION

- A set of documents $x_1; x_2; \dots; x_n$. Let X denote the document matrix.
- The algorithm for document clustering based on CPI can be summarized.
- Construct the local neighbor patch, and compute the matrices. Project the document vectors into the Singular Value Document (SVD) subspace by throwing away the zero singular values.
- With the help of CPI implementation all zero singular values in X have been removed. Accordingly, the vectors in U and V that correspond to these zero singular values have been removed as well.

STEMMING RULE IMPLEMENTATION

- The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and in flexional endings from words in English.
- Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.
- Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form generally a written word form.

REQUENCY CALCULATION

- The term frequency vector can be computed as follows:
- Transform the documents to a list of terms after words stemming operations.
- Remove stop words. Stop words are common words that contain no semantic content.
- Compute the term frequency vector using the TF/IDF weighting scheme.

CLUSTER BASED MODELING

- The experimental results of LPI and CPI on data set are obtained when the number of nearest neighbors is set to seven or eight. .
- For data sets, the number of nearest neighbors used for CPI varies from 3 to 24.
- CPI algorithm has two essential parameters: the dimension of optimal semantic subspace and the number of nearest neighbors.
- Unfortunately, how to determine the optimal dimension of the semantic subspace is still an open problem. In typical spectral clustering, the dimension of semantic subspace is set to the number of clusters.

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

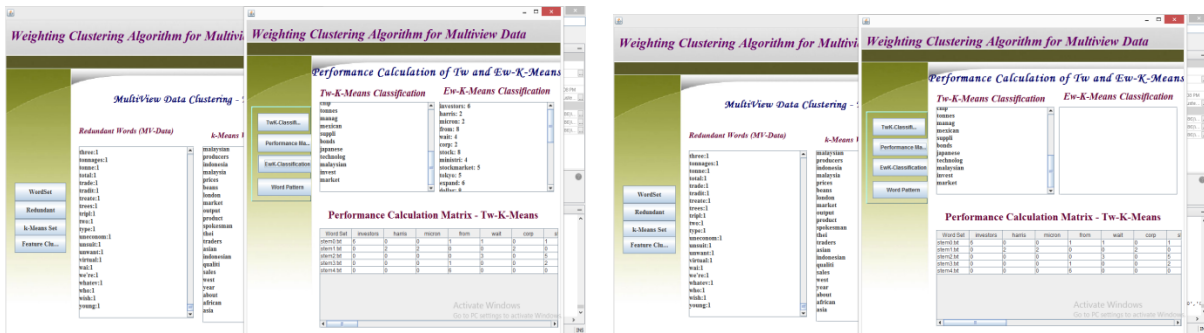
Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

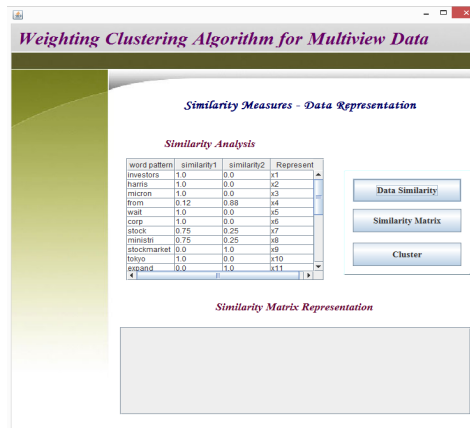
VII. RESULT ANALYSIS AND DESCRIPTION

Clustering technique using Enhanced TW-K-Means algorithm and using semantic analysis process to increase purity of the document

RESULT SHAPSHOTS

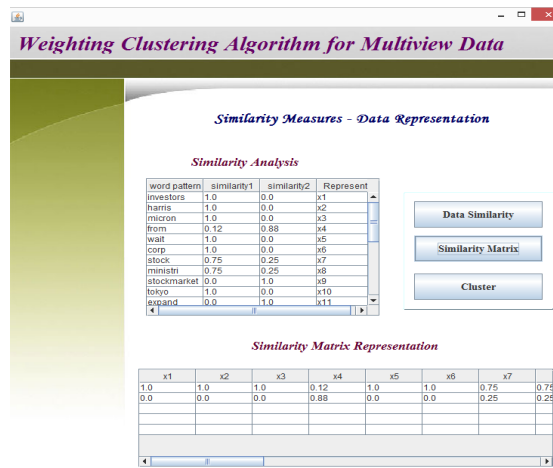


The screenshots show the 'Performance Calculation of Tw and Ew-K-Means' window. The left pane displays 'Redundant Words (MV-Data)' with a list of words like 'investors', 'harris', 'micron', etc. The right pane shows 'Performance Calculation Matrix - Tw-K-Means' with a table of similarity values for various words.



This screenshot shows the 'Similarity Measures - Data Representation' window. It includes a 'Similarity Analysis' table and a 'Similarity Matrix Representation' area.

word pattern	similarity1	similarity2	Represent
investors	1.0	0.0	x1
harris	1.0	0.0	x2
micron	1.0	0.0	x3
from	0.12	0.88	x4
wait	1.0	0.0	x5
corp	1.0	0.0	x6
stock	0.75	0.25	x7
ministri	0.75	0.25	x8
stockmarket	0.0	1.0	x9
tokyo	1.0	0.0	x10
expand	0.0	1.0	x11



This screenshot shows the 'Similarity Matrix Representation' area with a detailed matrix of similarity values between clusters.

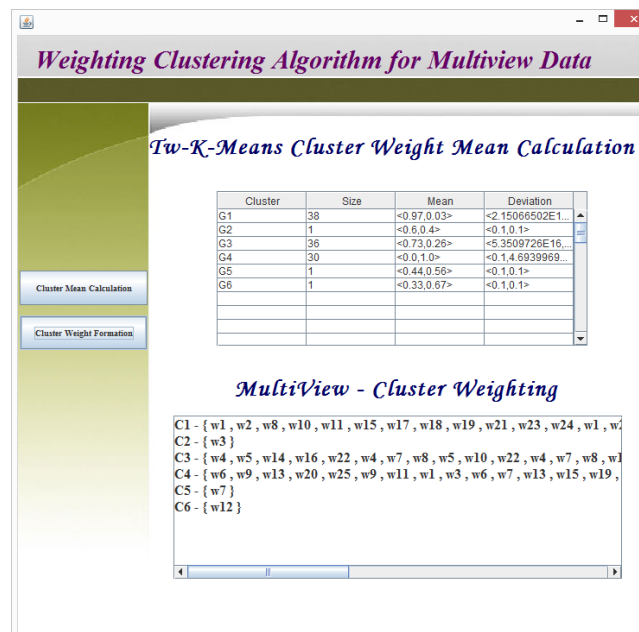
	x1	x2	x3	x4	x5	x6	x7
1.0	1.0	1.0	1.0	0.12	1.0	1.0	0.75
0.0	0.0	0.0	0.0	0.88	0.0	0.0	0.25

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019



VIII. CONCLUSION AND FUTURE WORK

In this experimental studies we concludes that the proposed topic overall performance is similar. So we introduced multi-viewpoint based similarity measure and two related clustering methods. Using multiple viewpoints, more informative assessment of similarity could be achieved and performance is much better than above similarity measures. Enhanced models for document representation looks promising aspect to probe. For example: the incorporation of semantic information and taking account of the semantic relatedness using WordNet between documents in the bag of words model could lead to better accuracy. This is due to the fact that similar meaning words weights will add up the weight for that particular term in the respective document.

REFERENCES

- [1] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, "The expanding digital universe: A forecast of worldwide information growth through 2010," *Inf. Data*, vol. 1, pp. 1–21, 2007.
- [2] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London, U.K.: Arnold, 2001.
- [3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Engle-wood Cliffs, NJ: Prentice-Hall, 1988.
- [4] L. Kaufman and P. Rousseeuw, *Finding Groups in Gata: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley-Interscience, 1990.
- [5] R. Xu and D. C. Wunsch, II, *Clustering*. Hoboken, NJ: Wiley/IEEE Press, 2009.
- [6] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse frame-work for combining multiple partitions," *J. Mach. Learning Res.*, vol. 3, pp. 583–617, 2002.
- [7] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, "Evolving clusters in gene-expression data," *Inf. Sci.*, vol. 176, pp. 1898–1927, 2006.
- [8] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps," in *Proc. IFIP Int. Conf. Digital Forensics*, 2005, pp. 113–123.
- [9] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Im-proving information retrieval effectiveness by thematically clustering search results," *Digital Investigation, Elsevier*, vol. 4, no. 1, pp. 49–54, 2007.
- [10] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis frame-work," *Digital Investigation, Elsevier*, vol. 5, no. 3–4, pp. 124–137, 2009.