ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 4, Issue 4, April 2017

Web Document Clustering Algorithm and Similarity Measure

[1]Ms.S.M.Durge, [2]Mr.Y.M.Kurwade, [3]Dr.V.M.Thakare
[1] [2] [3]SGBAU, Amravati, India
[1]sukshmandurge7@gmail.com, [2]yogeshwarkurwade@gmail.com, [3]vilthakare@yahoo.co.in

***Abstract :-* — The Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Most of the approaches perform web documents clustering, i.e., they assign each object to precisely one of a set of clusters. Objects in one cluster are similar to each other. The similarity between objects is based on a measure of the distance between them.This works well when clustering the compact and well-separated groups of data, but in many situations, clusters are different at rerun. This proposed method usek-means++ algorithm,is capable of identifying problem by spreading the initial centers evenly and improves performance.**

***Keywords:* - High-dimensional data, Clustering, Similarity measure.**

I. INTRODUCTION

The high dimensional data are used in many areas like, image processing, pattern recognition, bio-informatics. This data not only increases the computational time and memory requirements of algorithms, but also adversely affects their performance due to the noise effect and insufficient number of samples with respect to the ambient space dimension.

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. The web document Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity[1][3][5].The Sparse Optimization Clustering program is in general NP-hard to consider a convex relaxation and show that under appropriate conditions on the arrangement of the subspaces and the distribution of the datacan be very useful in many different scenarios [2]. The main challenge for most of clustering algorithms is their necessity to know the number of clusters for which to look. Some researchers have tried to estimate or determine it automatically [4].

In this paper, numbers of different optimization methods for clustering process are reviewed. All these methods are compared with their quality and computational time and also accuracy. Most of these approaches perform web document clustering, is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters. Clustering is a technique for extracting information from unlabelled data. The similarity between objects is based on a measure of the distance between them. The proposed clustering methodology is choosing the initial values (or "seeds") from the k-means clustering algorithm. The avoiding, sometimes poor clustering found by the standard algorithm. We can readily apply k-means(fast) algorithms. Since pair-wise similarities or dissimilarities between data points can readily be calculated from the attribute data using similarity measures such as cosine similarity. This paper tries to solve this problem by spreading the initial center evenly using clustering.The cluster similarity measure and find out distance between cluster in different metrics.

II. BACKGROUND

In this paper, analysis Real-time clustering method also called Adaptive Spherical K-Means (ASKM)method [1], sparse subspace clustering to cluster data points that lie in a union of low-dimensional subspaces. [2],TW-k-means, an automated two-level variable weighting clustering algorithm [3],Tripartite clustering method [4] and constrained clustering method that is based on a graph-cut problem.[5].

The real-time processing ASKM [1] used different process,To solve it, an Adaptive dimension reduction method to extend the vector space with new dimensions or remove ones which are not relevant anymore automatically. This technique the dimension of vectors can be kept on a relatively low level.Adaptive cluster creation, fixed the number of cluster is a key problem in the original sOSKM method. The proposed method extends the idea of ASKM clustering method.

A sparse subspace clustering to cluster data points that lie in a union of low-dimensional subspaces.[2],sparse optimization program is in general NP-hard then consider a convex relaxation and show that under appropriate conditions on the arrangement of the subspaces and the distribution of the data.State-of-the-art is that it can deal directly with data nuisances, such as noise, sparse outlying entries, and missing entries, by incorporating the model of the data into the sparse optimization program.

A TW-k-means clustering algorithm [3] ismulti view data which can simultaneously compute weights for views and individual variables.The subprocess of the Optimize Weights must always return a performance vector.The view weights will be only determined in the view level and the variable weights will be only determined in a view.

A Tripartite Clustering [4], algorithm that extends the k-means algorithm, which clusters the three types of nodes (resources, users, and tags) simultaneously by only utilizing the links in the social tagging network and also investigates two other approaches to exploit social tagging for clustering with K-means and Link K-means. The Tripartite Clustering fully relies on the relationships among the Web pages, users, and tags for clustering.

A constrained clustering method that is based on a graph-cut problem.[5]formalized by SDP(Semi-Definite Programming). The Semi-definite programming is a kind of convex optimization that is used to relax several optimization problems such as combinatorial optimization 0-1 integer programming and non-convex quadratic programming.
The rest of this paper is organized as follows. In Section III, we discuss previous work, Section IV,we discuss existing methodologies,Section V, analysis and discussion. Section VI we present the proposed methodology, and we present possible outcome and results in Section VII. We conclude this paper in Section VIII.

III. PREVIOUS WORK DONE

Over the past decades, many clustering algorithms have been proposed, including k-means clustering, spectral clustering, sparse subspace clustering,Most of these approaches perform k-means clustering, i.e.,each object is assigned to precisely one of a set of clusters. Objects in one cluster are similar to each other.

Adrian Pusztat et al. (2013) [1] Real-time textual content clustering of different sources called documents over the Internet. The Adaptive Spherical K-Means (ASKM)used to keep dimension of the document space on a reasonable level and ability to open new and remove old clusters.

EhsanElhamifar et al. (2013) [2], A sparse representation technique called SSC to cluster a collection of data points lying in a union of low-dimensional subspaces. A collection of multi subspace data using sparse representation techniques and motivate and formulate the algorithm for data points that perfectly lie in a union of linear subspaces.

Xiaojun Chen et al. (2013) [3]TW-k-means an automated two-level variable weighting clustering algorithm for multi view data which can simultaneously compute weights for views and individual variables.

Caimei Lu et al. (2011) [4]The Tripartite network consists of a three types of nodes: User, Resource and tag.A user’s interest can be featured by the resources that he has previously annotated and the tags that he has used, a resource’s topic can be represented by the users who have annotated the resource and the tags that have been assigned to the resource. This approach should be able to cluster resources not only based on their direct links to the user nodes and tag nodes but also based on the cluster structures of the user nodes and the tag nodes. The Tripartite networkinvestigates two other approaches to exploit social tagging for clustering with K-means and Link K-means to produced useful information.

Masayuki Okabe et al. (2011) [5], the minimum cut problem of graph partitioning find a method to solve constrained maximum cut problems. The semi-definite programming (SDP) is practically easier to use because it can handle constraints intrinsically. Semi-definite programming is a kind of convex optimization that is used to relax several optimization problems such as combinatorial optimization 0-1 integer programming and non-convex quadratic programming. Thus our proposed methods overcome and used high dimensional clustering method to improved performance.

IV. EXISTING METHODOLOGIES

A. Adaptive Spherical K-Means (ASKM) Clustering

The adaptive methods extend the vector space with new dimensions or remove ones which are not relevant anymore automatically. This technique the dimension of vectors can be kept on a relatively low level. This method cluster creation, Fixed the number of cluster is a key problem in the original sOSKM method. This technique the number of cluster K, start clustering process determines the nearest centroid and its distance to the document vector. The distance is greater than the predefined threshold limit (T) then new cluster created whose centroid is set to the given document vector. Choosing the proper threshold is essential, because it significantly affects the clustering result.The method of cluster deletion, handle out-lier conditions assign a lifetime property to each cluster. If no document gets into a cluster after processing a batch, the clusters lifetime decreases, otherwise it resets to its initial value. Once the lifetime reaches zero, the cluster will be removed.

B. Sparse Subspace Clustering

TheSparse Subspace Optimization Programself-expressiveness property of the data,More precisely, each data point $y_i \in U^n_{\ell=1} S_\ell$ can be written as,

$$\boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i, \qquad c_{ii} = 0,$$

where $\boldsymbol{c}_i \triangleq [\, c_{i1} \;\; c_{i2} \;\; \cdots \;\; c_{iN} \,]^\top$ and the constraint $C_{ii}=0$ eliminates the trivial solution of writing a pointas a linear combination of itself.The efficiently finding a nontrivial sparse representation of y_i in the dictionary Y, consider minimizing the tightest convex relaxation of the ℓ_0-norm, i.e.,

$$\min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i, \; c_{ii} = 0$$

The sparse optimization program for all data points i = 1, . . .,N in matrix form as

$$\min \|\boldsymbol{C}\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{C}, \operatorname{diag}(\boldsymbol{C}) = \boldsymbol{0},$$

Where $\boldsymbol{C} \triangleq [\boldsymbol{c}_1 \;\; \boldsymbol{c}_2 \;\; \cdots \;\; \boldsymbol{c}_N] \in \mathbb{R}^{N \times N}$ is the matrix whose ith column corresponds to the sparse representation of $\boldsymbol{y}_i$, $\boldsymbol{c}_i$, and $\operatorname{diag}(\boldsymbol{C}) \in \mathbb{R}^N$ vector of the diagonal elements of C.

The Clustering Using Sparse Coefficients, A weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$, where V denotes the set of N nodes of thegraph corresponding to N data points and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges between nodes.

The Noise and Sparse Outlying Entriesdo not lie perfectly in a union of subspaces. In some real-world problems the data lie in a union of affine rather than linear subspaces.The success of the SSC algorithm is proposed optimization program recovers a subspace-sparse representation of each data point i.e. a representation nonzero elements correspond to the subspace of the given point.

The data points that lie in a union of independent subspaces which is the underlying model of many subspace clustering algorithms. The independent span a 3D space and the sum of their dimensions is also 3. And the more general class of disjoint subspaces and investigate conditions under the optimization program recovers a subspace-sparse representation of each data point. The disjoint each pair of subspaces intersects at the origin.

C. TW-k-means clustering algorithm

The multiview clustering methods take both multiple views and individual variables into consideration.Compute variable weights automatically and the view weights are given by users. The optimization clustering model process to partition X into k clusters with weights for both views and individual variables is modeled as minimization.

D. Tripartite Clustering model on a social tagging system

The tripartitenetwork denotedby

$$TN = \left(U, R, T, E^{(UR)}, E^{(UT)}, E^{(RT)}\right)$$

U, R, and T are finite sets of users, resources, and tags, respectively and E(UR), E(UT), and E(RT) three types of undirected links in the network.The k-means calculates the cluster centroids and reassigns each document to the closest cluster until no document can be reassigned. Link K-means, users and tags act as bridges which bring topically related Web pages together.

E. Constrained Clustering Method

The minimum cut problem of graph partitioning find a method to solve constrained maximum cut problems.A graph G = (V,E), V is a set of vertices and E is a set of edges. The w_{ij} is weight of an edge between data i ∈ V_1 and j ∈ V2. The Semi-definite Programming Relaxation (SDP)

It cannot-link it is very difficult to select applicable cannot-links during multi-class clustering because the partitioning order determined in the graph cut process is usually unpredictable in advance.

Maximum Cut Problem with SDP Relaxation

$$\begin{array}{ll} \text{maximize} & L \bullet X \\ \text{subject to} & E_{ii} \bullet X = 1, \quad (i = 1 \sim n) \\ & E_{ij} \bullet X = 1, \quad (i,j) \in M \\ & X \succeq O \end{array}$$

The Swapping Rows and Columns in a Label Matrix, First binarize ˜X with 0-1 values then swap rows and columns to maximize the evaluation measure and finally determine the partitioning border.

V. ANALYSIS AND DISCUSSION

To improve accuracy and incomplete web documents clustering result, it use precision as a qualifying measure and F-score combine both precision and recall to calculate result in the Clustering Complete Documents.The captured incomplete textual documents to recognize the character encoding and separate the control elements like HTML tags from the textual content. These preprocessing steps deal with the effects of sampling on the clustering accuracy as well. The downloading document has been thrown out it is called shortfall value an incomplete document stream.[1].

The SSC optimization algorithm using an alternating direction method of multipliers (ADMM) framework of derivation is provided in the online supplementary material.The Motion Segmentation used random projections for dimensionality reduction and use PCA or original 2F-dimensional data. In additions used a CVX solver to compute a subspace-sparse representation. In the Face Clustering to validate the fact that corruption of faces is due to sparse outlying errors so to apply the robust principal component analysis (RPCA) algorithm to remove the sparse outlying entries of the face data in each subject.[2].

The view weights will be only determined in the view level and the variable weights will be only determined in a view. Therefore, the two levels of variable weights will eliminate the unbalanced phenomenon and compute more objective weights. The result is a TW-k-means clustering algorithm is efficient and outperform under the previous k-means clustering process with two additional steps to compute view weights and variable weights in each iteration.[3].

Parameter selection of the clustering results of web pages are quantitatively evaluated against the 14 ODP categories, so kR is also set to 14 for evaluation purposes then decide kU and kT based on the clustering quality of a random sample of Web pages.[4].

The semi-definite programming to adopted the Euclid distance for the weight wij of a graph edge in the maximum graph-cut problem.[5].

A brief comparison of ASKM Clustering, SSC,TW-k-means algorithm, Tripartite Clustering model and Constrained Clustering Method are as shown in Table_1.

Clustering methods	Advantages	Disadvantages
Adaptive Spherical K-Means (ASKM) Clustering	1) Good accuracy. 2) Performs well with incomplete 3) Contents. 4) Good efficiency.	1. The clustering quality decrease.
Sparse Subspace Clustering (SSC)	1) Deal directly with data nuisances, noise, sparse outlying entries, and missing entries. 2) Reduced computational 3) cost and memory requirement.	1. The used of small dataset.
TW-k-means clustering algorithm	1. Reduced noise.	1. More execution time. 2. Scales well only high-dimensional data.
Tripartite Clustering	1.Achieves equivalent or better performance 2. Produces more useful information.	1. The quality of clusters quantitatively.
Constrained Clustering Method	1. Improve normal cluster accuracy. 2. Avoid a trivial solution. 3. Make constrained clustering more efficient.	1. Small number of constraints.

Table 1: Comparison between ASKM, SSC, TW-K-means, Tripartite and Constrained Clustering

VI. PROPOSED METHODOLOGY

By applying k-means++ clustering algorithm attemptsto address existingproblem by spreading the initial centers evenly. K-means++clustering methods, for choosing the initial values (or "seeds") for the k-means clustering algorithm.

ALGORITHM:-

1. Choose one center uniformly at random from data point.
2. For each data point x, compute D(x), distance between x and nearest center already chosen.
3. Choose one new data point at random as a new center, using weighted probability distribution.
   Where point x chosen with proportional to D(x)$_{2.}$
4. Repeat steps 2 and 3 until k centers chosen.

5. Now that initial centers chosen to proceed k-means clustering.

Flow Diagram:

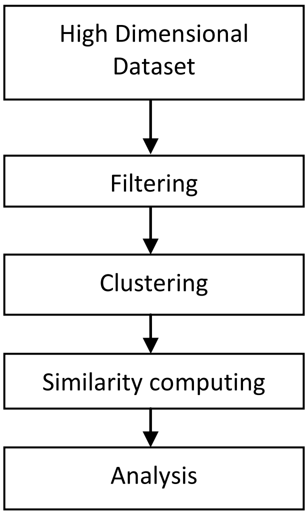

Fig.1. Flowchart of the proposed method

VII. POSSIBLE OUTCOMES AND RESULTS

Performance Measures

The centroid based clustering like the K-Means++ produce a centroid cluster model and a clustered set. It tells which examples are parts of which cluster. It also has information regarding centroids of each cluster. The Cluster Distance Performance takes this centroid cluster model and clustered set as input and evaluates the performance of the model based on the cluster centroids. The performance of clustering result is improved and fast.

VIII. CONCLUSION

This paper exhausted different clustering algorithm for optimizing procedure and provide low computational complexity. The results of these clustering algorithm show that the algorithm is able to achieve superior performance.The similarity between objects is based on a measure of the distance between them. The clusters are semantically related sentences.

REFERENCES

[1] Adrian Pusztat, Janos Sziilet and SandorLaki, “Near Real-Time Thematic Clustering of Web Documents and Other Internet contents”, 4th IEEE International Conference on Cognitive Info-communications, pp. 307-312, DEC. 2-5 2013.

[2] EhsanElhamifar, and Rene Vidal, “Sparse Subspace Clustering: Algorithm, Theory, and Applications”, IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 35, NO. 11, pp. 2765-2781, NOV. 2013.

[3] X. Chen, X. Xu, J. Zhexue Huang and Yunming Ye, “TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multi view Data”, IEEE Transactions on Knowledge and Data Engineering, VOL. 25, NO. 4, pp. 932-944, APRIL 2013.

[4] Caimei Lu, Xiaohua Hu and Jung-ran Park, “Exploiting the Social Tagging Network for Web Clustering”, IEEETransaction on Systems, Man and Cybernetics, VOL.41, NO. 5, pp. 840-852, SEP. 2011.

[5] Masayuki Okabe and Seiji Yamada, “Graph-cut based Iterative Constrained Clustering”, 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 126-129, NOV. 2011.