IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

An Effective Model for User Clustering To Track Similarities

1 M Teja, 2 Dr. M Arathi
1 Mtech Scholar 2 Associate Professor
1 Computer Science,2 Department of Computer Science
1,2 School of Information Technology - JNTUH, Hyderabad, India

ABSTRACT: Clustering of information with high dimensions and variable densities makes an uncommon challenge to the conventional density-based clustering strategies. As of late, entropy, a numerical measure of the instability of data, can utilized to compute the border degree of samples in information and select significant highlights within the highlight set. The proposed model focuses on handling the issue of client grouping within the setting of their distributed content streams. It was utilized in this system depend on sparsity-density entropy. to cluster the information with high measurements and variable densities. To begin with, sparsity density entropy (SDE) directs a high-quality inspecting for multidimensional information and chooses the significant highlights utilizing sparsity score entropy. Then DE comes into picture the DE successively decides the border set depend on the global minimum of border degrees. The performance and exactness of proposed model are validated on data sets and comparing with different clustering algorithms. The outcomes demonstrated that proposed method can detected the noises and prepared the data with high dimension and various densities.

1. INTRODUCTION

Data clustering is most extensively used methods in data mining; data clustering has useful applications in pattern recognition, picture processing, and data compression. Clustering algorithms are grouped into five types they are hierarchical, grid-based, density-based algorithms and model-based algorithms. Partition clustering methods are divided into k medoids,k means and also Fuzzy c-means, if you take any of this methods in this methods assign the incoming data points to K disjoint subsets, such that points within a same cluster are similar than those in other clusters. Even though, the number of clusters are pre given, and the results are very sensitive to initial points, shapes and sizes of clusters.

Hierarchical clustering contains both agglomerative methods and also divisive methods: the agglomerative methods begin with single-point clusters. Grid based clustering models classify the data into a fixed no of cells that constitute a grid structure, However, the relation between effectiveness and correctness remains challenge in the grid clustering. Clustering algorithms are categorized on probability models utilize parametric models to improve the wellness among data and models, namely the expectation maximization models.

Moreover, The clustering algorithms are grouped on graphs and artificial networks additionally been proposed, e.g., spectral clustering and self-organizing maps The basic intention of spectral clustering is to develop a weighted graph, where the vertex represents data points and each weighted edge indicates the closeness between each corresponds to pair of vertex Normalized cut and NJW are classic spectral clustering methods. The Artificial network based self-organizing maps generates a low-dimensional data of information utilizing unsupervised competitive learning and has been applied to pattern recognition, process monitoring, image analysis and fault diagnosis.

In addition to the above methods, density-based methods has formed an important research focus of clustering algorithms.Density-based clustering, e.g., density-based clustering applications with noise such ordering points and DBSCAN to recognize the clustering structure, separate clusters by the density of points in regions. In DBSCAN, a cluster is determined as a dense component with high connectivity it can expand in any direction where the density leads.

However, DBSCAN is parameter-sensitive, which can severely impact the exactness. The parameters are hard to find without a reference range. In addition, it is not capable to find the clusters with variable densities or acquire the entire boundary of each cluster, arbitrary shapes, there are some drawbacks. The foremost one is the low to inform these problems; researchers developed a series of improved algorithms. VDBSCAN finds clusters with variable densities and automatically determines the values of input specifications depending on the features of the datasets. However, the selection of parameter K on various datasets is still a challenge.

In reality, each cluster has its individual density. Utilizing consistent density measures might lead to group the boundary data points into noises or noises into certain cluster. The second one is the high model complexity, especially in multi dimensional information. Moreover, redundant features would hurt the performance of clustering and decrease accuracy. to control these disadvantages we proposed the use of sparsity score entropy to get more informative features and density entropy for density-based clustering. Our method lays two assumptions. The first one is that the cluster border has a trivial impact on the whole density distribution in the data space. The cluster boundary is located on the transitional space among high- and low-density regions of a given dataset. Core points in the high- density region have a great proportion in data space and can impact the entire density distribution.

2. LITERATURE SURVEY

This paper address the issue of user clustering with regards to their context of uploaded text streams. The density clustering strategies are capable of finding clusters with different shapes however there are still few downsides. The foremost one is the low robustness against density variances among the clusters and the subsequent one is high model complexity the existing model is used k means clustering algorithm increasing sample size generally leads to higher accuracy but efficiency reduced.Utilizing statistical Optimal Sample Size which can decide the sample size on enormous data sets by measuring the sample quality dependent on the data distribution

To wipe out these drawbacks introduce a novel framework, called the SDE system, which can process both low- and high-dimensional data. For high-dimensional information we use sparsity score entropy method to automatically find significant features and remove the redundant features without manual threshold limit by computing the entropy ratio of each feature to gain representative features and weighing them, we can not only explore the advances in dimension reduction, but also ensure the data quality.

SDE clustering outperforms different techniques for clusters with variable densities through a two step grouping for each cluster. The initial clustering outcomes and a labelled border set are acquired based on the absolute minimum of all the border degrees.

3. SYSTEM ANALYSIS

System analysis consists of the various existing systems, drawbacks in them, and proposed system with advantages

3.1 Existing System

The Existing System based on k means algorithm. Increasing sample size usually leads to a high accuracy however efficiency reduced. Using a strategy known as Statistical Optimal Sample Size (SOSS) which can decide the sample size on large information sets by computing the sample quality based on the information. In general, Existing Algorithms are able to of discover clusters with random shapes, bur there are still some disadvantages .The first is the low strength against density variances between the clusters, i.e., density-based methods find clusters by utilizing global parameters, the parameters are the radius of the clusters (Eps) and the MinPts, which may not manage with individual groups of various densities.

3.2 Disadvantages

1. More Complexity in multi-dimensional space
2. Redundant characteristics would hurt clustering performance and reduce exactness
3. It depends on the center of the cluster

3.3 Proposed system

The proposed system uses a novel system, called the Sparsity- Density Entropy framework, which can successfully process both low- and high-dimensional information. At the first step sparisity density entropy directs a high-quality sampling for multidim. After the initial step then DE naturally decides the border set based on the global minimum of border degrees It finds similar things in user documentsensional data.

3.4 Advantages

1. One of the advantage of this is Automatically decides the values of input values dependent on the characteristics of the datasets

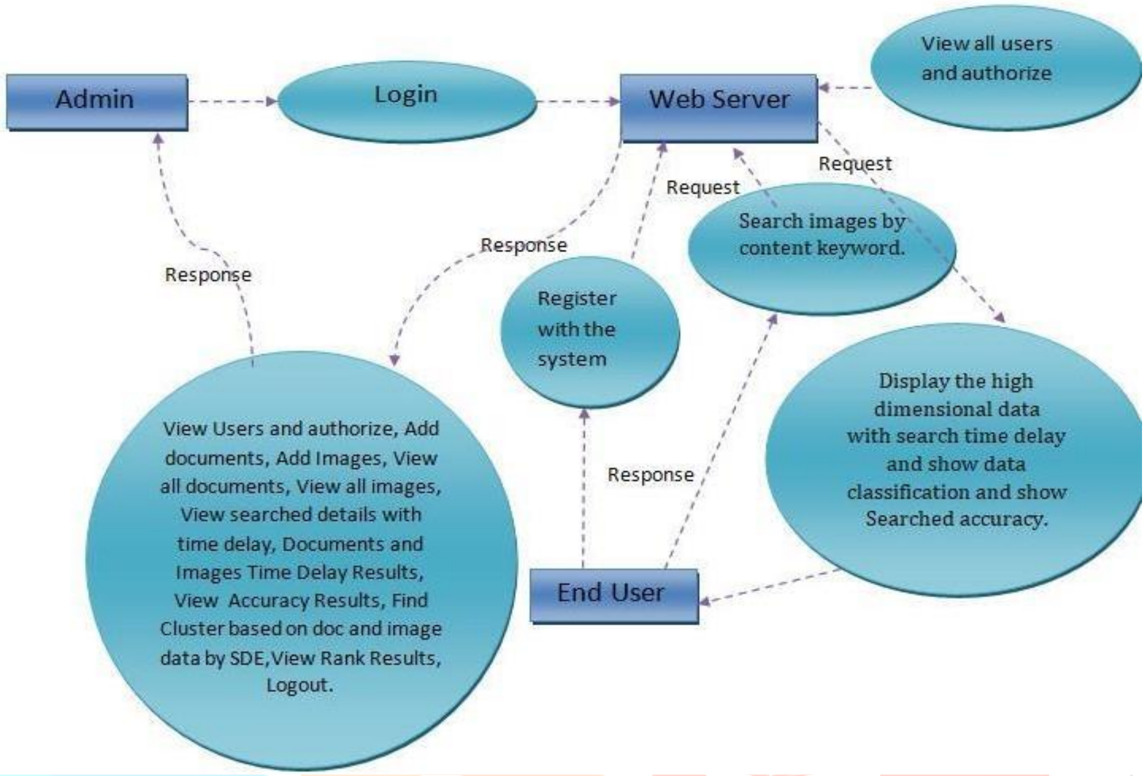
2. High Accurate and Efficient.

4. IMPLEMENTATION

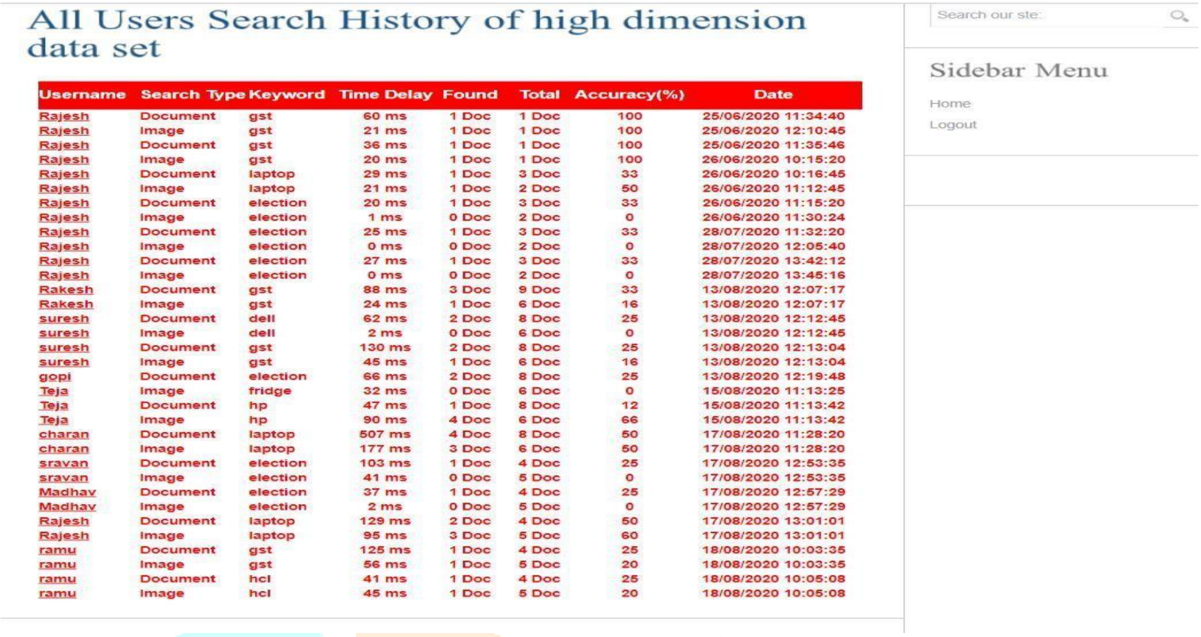
In order to handle multidimensional data and eliminate redundancies to use two methods they are sparsity score entropy and DE method. The sparsity score entropy method is utilized to select important features Once the first phase is completed then DE comes into picture it is useful for performing density based clustering .All features are sorted with respect to sparsity score from smallest to largest .the smaller the sparsity value more the sparse preserving ability. According to this algorithms space complexity is O(n) the approach based on the total dataset will lead to higher algorithm complexity

In order to solve this complexity here first utilized the SOSS algorithm this algorithm used to determine the optimal sample size and based on this size compute the sparsity score of each feature and this methods are very useful to reduce the algorithm complexity and keep the same accuracy.In general use value of threshold to select the important features based on the sparsity scores different thresholds to might lead to different outcomes and it is tough to find an accurate threshold to solve this use the information entropy ration method instead of manual threshold.

5. DATA FLOW DIAGRAM


Admin
Login
Web Server
View all users and authorize
Request
Request
Search images by content keyword.
Response
Response
Register with the system
Display the high dimensional data with search time delay and show data classification and show Searched accuracy.
View Users and authorize, Add documents, Add Images, View all documents, View all images, View searched details with time delay, Documents and Images Time Delay Results, View Accuracy Results, Find Cluster based on doc and image data by SDE,View Rank Results, Logout.
Response
End User

6. RESULTS



All Users Search History of high dimension data set

Search our site:

Sidebar Menu

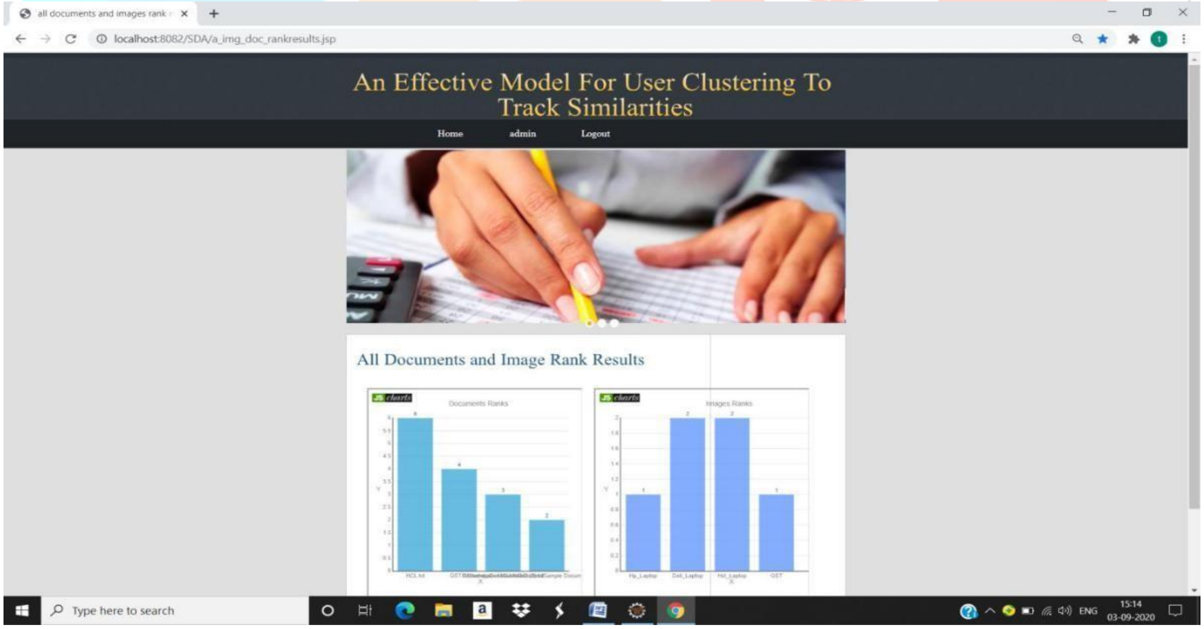
Home

Logout

Username	Search Type	Keyword	Time Delay	Found	Total	Accuracy(%)	Date
Rajesh	Document	gst	60 ms	1 Doc	1 Doc	100	25/06/2020 11:34:40
Rajesh	Image	gst	21 ms	1 Doc	1 Doc	100	25/06/2020 12:10:45
Rajesh	Document	gst	36 ms	1 Doc	1 Doc	100	25/06/2020 11:35:46
Rajesh	Image	gst	20 ms	1 Doc	1 Doc	100	26/06/2020 10:15:20
Rajesh	Document	laptop	29 ms	1 Doc	3 Doc	33	26/06/2020 10:16:45
Rajesh	Image	laptop	21 ms	1 Doc	2 Doc	50	26/06/2020 11:12:45
Rajesh	Document	election	20 ms	1 Doc	3 Doc	33	26/06/2020 11:15:20
Rajesh	Image	election	1 ms	0 Doc	2 Doc	0	26/06/2020 11:30:24
Rajesh	Document	election	25 ms	1 Doc	3 Doc	33	28/07/2020 11:32:20
Rajesh	Image	election	0 ms	0 Doc	2 Doc	0	28/07/2020 12:05:40
Rajesh	Document	election	27 ms	1 Doc	3 Doc	33	28/07/2020 13:42:12
Rajesh	Image	election	0 ms	0 Doc	2 Doc	0	28/07/2020 13:45:16
Rakesh	Document	gst	88 ms	3 Doc	9 Doc	33	13/08/2020 12:07:17
Rakesh	Image	gst	24 ms	1 Doc	6 Doc	16	13/08/2020 12:07:17
suresh	Document	dell	62 ms	2 Doc	8 Doc	25	13/08/2020 12:12:45
suresh	Image	dell	2 ms	0 Doc	6 Doc	0	13/08/2020 12:12:45
suresh	Document	gst	130 ms	2 Doc	8 Doc	25	13/08/2020 12:13:04
suresh	Image	gst	45 ms	1 Doc	6 Doc	16	13/08/2020 12:13:04
gopi	Document	election	66 ms	2 Doc	8 Doc	25	13/08/2020 12:19:48
Teja	Image	fridge	32 ms	0 Doc	6 Doc	0	15/08/2020 11:13:25
Teja	Document	hp	47 ms	1 Doc	8 Doc	12	15/08/2020 11:13:42
Teja	Image	hp	90 ms	4 Doc	6 Doc	66	15/08/2020 11:13:42
charan	Document	laptop	507 ms	4 Doc	8 Doc	50	17/08/2020 11:28:20
charan	Image	laptop	177 ms	3 Doc	6 Doc	50	17/08/2020 11:28:20
sravan	Document	election	103 ms	1 Doc	4 Doc	25	17/08/2020 12:53:35
sravan	Image	election	41 ms	0 Doc	5 Doc	0	17/08/2020 12:53:35
Madhav	Document	election	37 ms	1 Doc	4 Doc	25	17/08/2020 12:57:29
Madhav	Image	election	2 ms	0 Doc	5 Doc	0	17/08/2020 12:57:29
Rajesh	Document	laptop	129 ms	2 Doc	4 Doc	50	17/08/2020 13:01:01
Rajesh	Image	laptop	95 ms	3 Doc	5 Doc	60	17/08/2020 13:01:01
ramu	Document	gst	125 ms	1 Doc	4 Doc	25	18/08/2020 10:03:35
ramu	Image	gst	56 ms	1 Doc	5 Doc	20	18/08/2020 10:03:35
ramu	Document	hcl	41 ms	1 Doc	4 Doc	25	18/08/2020 10:05:08
ramu	Image	hcl	45 ms	1 Doc	5 Doc	20	18/08/2020 10:05:08

Fig All Users Search History

Above fig displays all users search history with accuracy and search type and time delay and also shows which type of content he search such as document or image



An Effective Model For User Clustering To Track Similarities

All Documents and Image Rank Results

All Documents Image Rank Results

The above picture displays all Document Image Rank Results of users

7. CONCLUSION

This paper is about the density-based automatic clustering framework called SDE. It has the advantages, including that it could concurrently recognize the noises, detect data clusters with varied densities and arbitrary shapes after two phases of clustering, and automatically select important and enlightening features according to the inborn properties of datasets. The existing algorithms are efficient for detecting the global noises but cannot deal with the local noises well, as the noises have extraordinary impact on the data distribution of the local cluster. In SDE, the global outermost border threshold to conduct initial clustering using the DE method. Initially determining the clustering scope, we carried out clustering once again on each local scope with its local outer-most border threshold. Through two step clustering, we eliminated both the global and local noises, and also clustered the datasets according to similar density metrics. Moreover most of the clustering algorithms need the clustering number as a prior, just need to set the number of closest neighbours to assess the density metrics of data points. The power of the proposed SDE algorithm has been evaluated on some synthetic datasets and real datasets.

Future Enhancements:

The main aim of this paper clustering of users to track their similarities it can be done by using two step clustering and using a algorithm called SDE to improve accuracy .We also reduce dimensions by selecting significant features with large sparsity score entropy values. It based on the information quality and does not need any threshold.In the future it will be useful to track similarities of users and find their interests in the form of their search content and send same information(adds,images,information) to group of people.

8. REFERENCES

[1] A. K. Jain, M. N. Murthy, and P. J. Flynn, “Data clustering: a review,” Acm Computing Surveys, vol. 31, no. 2, pp. 264–323, 2015.J. A. Hartigan and M. A. Wong, “A k-means clustering algorithm.” Applied Statistics, vol. 28, no. 1, pp. 100–108, 2013.

[2] X. Y. Wang and J. Bu, “A fast and robust image segmentation using fcm with spatial information,” Digital Signal Processing, vol. 20, no. 4, pp. 1173–1182, 2010.

[3] B. S. Guha and R. Rastogi, “Shim k: Rock: a robust clustering algorithm for categorical attributes,” Information Systems, vol. 25, no. 5, pp. 345–366, 2010

[4] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, 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, 2013.

[5] H. S. Park and C. H. Jun, "A simple and fast algorithm for k-melodies clustering," Expert Systems with Applications, vol. 36, no. 2, pp. 3336–3341, 2009.