DATA OBJECT REDUNDANCY & COMPLEXITY REDUCTION USING CONTINGUITY BASED GLOBULAR DATA CLUSTERING

Dr. S. Adaekalavan
Assistant Professor, Department of Computer Applications
J.J. College of Arts and Science (Autonomous), Pudukkottai, Tamilnadu, India

ABSTRACT:
The technology is discovered for extracting knowledge out of collected information in the data mining process. Clustering analysis is the most usually used data mining tool for grouping the set of the similar objects. However the data objects used for the clustering process may contain the sensitive information, so that the detection operation is a complicated task. Moreover the redundancy in the clustering group is not removed in the real applications. The higher level of the redundancy rate in the data mining clustering process leads to the time and space complexity. The complexity on the clustering produces the lesser quality rate and arise the problem on computing the similarity measure. To reduce the redundancy level on the clustering of the data objects, Contiguity based Globular Data Clustering (CGDC) Model is proposed in this paper. Initially, CGDC clusters the set of the nearest neighbor similar data objects with the bound of the globular density level. Globular provides the relative clustering of the different densities of the data objects by reducing the redundancy level in CGDC model. The reduced redundant data objects in the cluster, then steps into the Enhanced Sparsification technique. Enhanced Sparsification technique is the second step followed in CGDC model to keeps connection link with the most similar nearest neighbor data objects. The CGDC model thereby breaks the connections with lesser similar data objects using graph, by improving the clustering quality rate. The amount of time required to cluster the data object in CGDC model is drastically reduced along with minimal complexity on the state space. Experiment is conducted on factor such as clustering quality rate, redundancy reduction rate, and time complexity measure.

Keywords: Enhanced Sparsification, Globular Clustering, Nearest Neighbor, Redundancy Level, Data Objects Density

1. INTRODUCTION
    Clustering of data objects is used to ensure the high similarity into same information. The irrelevant data objects which are unlabeled are removed from the set. The development of the grouping performance based on the compactness typically reflects the investigator's
theory of clusters. To tackle the problem of compactness, a variety of clustering algorithm is developed in the literature. Since the number of data objects size increases, a more attention is paid on clustering the categorical data.

Clustering is one of the fastest emergent research areas since the accessible of large amount of data. Data modeling puts clustering in a chronological perspective order which is embedded in statistics, mathematics, and geometric analysis. All the data objects which are placed in a cluster have the density nearest to that data object. After processing the data objects, density value are recalculated, and the whole process is repetitive. The redundant clustered data is illustrated in Fig 1.

![Data Objects Redundant Subset Object against Comparison Redundant Cluster Formation](image)

**Fig 1 Redundant Cluster Formation**

Data clustering is used regularly in many applications such as data mining, vector quantization, and in many real world applications. The most well-known, extensively used and fast methods for clustering is K-means clustering, however the redundancy level is not reduced. K-means clustering is a partitioning clustering method that split information into k equal groups. Through such the iterative partitioning, K-means clustering minimizes the sum of distance from each data to its clusters, thereby increases the space complexity on storing the data information. K-Means unsupervised learning methods in [12] partition the data objects into clusters. The entire data objects are placed with the centroid nearest to the data object.

Document clustering plays an important role in indexing and retrieval of the large and high dimensional data. Document clustering algorithm named Semantic Smoothing Vector Space Kernel (S-VSM) in [2] is used to smooth the similarity and the demonstration of text documents. Top-k S-VSM, which judge only the top-k semantic interconnected terms which fails on examining the space and time complexity parameters. An original semi-supervised text clustering algorithm, called Seeds Affinity Propagation (SAP) that confine the similarity measurement of the texts [4]. The complete SAP algorithm get hold of higher F-measure and lower entropy to improve considerable clustering execution time. Semi-supervised learning approach exploits the information from a few labeled data objects versus a large number of unlabeled data objects.

Dark Block Extraction (DBE) in [7] mechanically estimate the large number of unlabeled data objects. The unlabeled data objects are worked with Visual Assessment of Cluster Tendency (VAT) on data set, using numerous ordinary sign processing technique.
The center samples for every significant cluster from any well structured uses the Reordered Dissimilarity Image (RDI), thus reducing the amount of the data clustering latency. A clustering-based pre-fetching system as demonstrated in [13] contains the graph-based clustering algorithm with ranking score. The algorithm recognizes clusters of connected web pages based on the users’ prototype. The scheme incorporated easily into a web proxy server for improving its performance and the added similarity metrics are not used for directed test generation.

Data mining with privacy preservation concept in [11] has been a significant research area for vast spectrum of application. PPC is used to look after the sensitive attributes value data object that is subjected to clustering analysis. Non-parametric data driven approach as illustrated in [9] approximate the number of clusters in a dataset. Non-parametric data driven approach uses the clustering algorithm and make out the highest number of groups with that dataset information. The slope method combined with the silhouette statistic increase the efficiency and correctly identifies the number of clusters. The different probability distributions and in the graph partition problem are solved.

Redundancy is one of the important issues in data clustering operation. Optimization-based model for generic document summarization as illustrated in [1] generates the rule to extract the sentences from the documentation. The optimization model selects the sentences to identify the similarity level, and deliver the majority level of the similarity. However, the redundancy occurred on the larger percentage on the subtopics extracted from documents.

In this work, Contiguity based Globular Data Clustering (CGDC) Model main goal is to cluster the nearest neighbor similar data objects. The data objects with similar densities in the neighbor are clearly clustered and enhanced Sparsification technique is employed. The main work of the enhanced Sparsification is to perform the linear graph cut. The linear graph cut improves the connectivity between the more similar data objects. The limited similarity between the data objects are discarded, thereby reducing the space complexity. Sparsification improves the speed of the data object clustering, in that way time complexity is reduced.

The structure of this paper is as follows. In Section 1, describe data objects different type of clustering model and limitations in the existing works. A detailed review of different types of the existing clustering algorithm is carried out. Section 2 presents the Contiguity based Globular Data Clustering (CGDC) Model to reduce the redundancy level. Section 3 performs the experimental evaluation on CGDC model and section 4 briefly describes about the results part. A result on a benchmark data set as well as the comparison with the selected existing algorithms is carried out to identify the result percentage. Section 5 describes about related existing clustering algorithms. Section 6 lastly concludes the work.

2. CONTIGUITY BASED GLOBULAR DATA OBJECTS CLUSTERING MODEL

The proposed work integrates the globular clustering technique with the concept of the nearest neighbor to reduce the redundancy level. In addition to reduce the time and space complexity on the clustering of data objects, the graph based enhanced sparsification approach is applicable in CGDC model. The globular clustering is operated on the large
galaxies (i.e., larger gene data objects from which the similar data objects are grouped as the star forming subsystems. Clearly, these star subsystems is capable of forming clusters with much more massive in CGDC model than the less significant open clusters that have formed.

The globular cluster uses the fossils (i.e., density of the data objects) at the initial stage to group the similar density nearest neighbor clusters, where the key density value used for the easier way of clustering process. The Globular cluster is exclusive class of the data object with the recurrent observation of data object properties for reducing the redundancy level. Normal Clustering with Globular Clustering structure is clearly described in below Fig 2.

Fig 2 (a) Normal Open Clustering  (b) Globular Clustering

Fig 2 explains the globular clustering usage in proposed CGDC model. The normal clustering groups the similar data objects with higher redundancy level, whereas CGDC model uses the massive complexes to group based on the nearest neighbor data objects density properties. The density properties are examined clearly and form the star subsystems thereby reducing the redundancy level on grouping of data objects. Architecture Diagram of Contiguity based Globular Data Clustering (CGDC) Model is described in Fig 3.
As illustrated in Fig 3, larger set of the data objects are measured to perform the clustering using the Globular clustering. Nearest neighboring data objects are monitored about the densities using the density value and perform the clustering operation. The similar density level data objects are grouped using the Contiguity based Globular Clustering model. The contiguity refers the nearest neighbor data objects in CGDC model. The nearest neighbor
data objects similarities are observed through the density properties for the easier clustering. The grouping of the similar density objects and dissimilar density are also clustered according to the particular value.

The reduced redundant level clustering integrates the enhanced Sparsification technique in CGDC model based on the graph form. The enhanced sparsification based graph cut is the simpler and faster system which easily estimates the edge connective link of the data objects. The easier estimation of the connectivity’s in CGDC model helps to reduce the time complexity rate. The linear graph cut procedure is used to produce the sparsifiers in CGDC model with minimal time factor. The disconnection of the unwanted connected links reduces the search space complexity in CGDC model automatically.

2.1 Contiguity based Globular Density Model

Assume data objects \( D = \{ \) \( d_1, d_2, d_3 \) be the vector space in the CGDC model to group the similar densities. Each cluster contains the ‘n’ similar data objects after the contiguity based Globular density operation. If the and are similar density data objects, then the objects comes under the same cluster. Contiguity based Globular density is suitable while following the density point similarity measure. The density point similarity uses the specified density value to perform the grouping process. The contiguity based density form is represented as,

\[
\text{Density} = \text{No. of objects within speci} \quad \text{...... Eqn (1)}
\]

Where ‘r’ is the radius covered. Density point similarity measure built an regular scaling factor to reduce the redundancy level. The nearest neighbor sharing data objects used on the grouping operation. The radius is specified to identify the nearest neighbor sharing with similar density function. A data objects with the specified radius ‘r’ is considered as the neighborhood sharing function. The neighbor sharing data objects is used in the Globular clustering. In particular the neighbors which are similar to the ‘r’ radius are matched through the similarity measure in CGDC model. The clustering based on the Globular is measured as,

\[
\text{Globular Clustering} (C) = n \ (SD \ (d_1, d_2, d_3) \quad \text{...... Eqn (2)}
\]

The globular clustering takes the ‘n’ similar density data objects’d’ along the radius ‘r’. The similar density ‘SD’ level is used for the easy star subsystem clustering operation. Similar density level value is used on plotting the different clusters in CGDC model. The Average Globular Cluster Similarity (AGCS) measure is employed in CGDC model to test the data objects similarity level. The average similarity is computed as,

\[
\text{AGCS} = \frac{n_{c_1}(D(c_1))+n_{c_2}(D(c_2))+\ldots+n_{c_m}}{n_{c_1}+n_{c_2}+\ldots+n_{c_m}} \quad \text{...... Eqn (3)}
\]

The nearest neighbor data objects with similar density ‘D’ are plotted and Globular clustering operation is carried out. The different type of the density function value is taken as the input factor to group the data objects. The group of different densities is constructed, which produce the reduced redundancy level. The algorithmic step is explained through the following manner in CGDC model.
Contiguity based Globular Clustering

Begin
Step 1: Construct the initial nearest neighbor data object list
Step 2: Place the density specified value as input factor
Step 3: Perform the clustering through contiguity based Density value
Step 3.1: Density point similarity measure is computed
Step 3.2: Computed point value is used to identify the similarity between the two nearest data objects
Step 4: Globular Clustering through Average Globular Cluster Similarity Measure
Step 4.1: Similarity between the two data cluster groups
Step 4.2: Compute
Step 5: Update the nearest neighbor list, reduces redundancy level

End

The contiguity based globular clustering model follows the similarity measure with the density points. Such that the contiguity needs the uses to provide the precise result about the nearest neighbor data objects in CGDC model. The larger the data objects are easily handled by specifying the density value at the input stage. Observations of globular clusters in CGDC model show that these star subsystem formations improves the clustering of different density data objects with reduced redundancy level.

Enhanced Sparsification Technique

Enhanced Sparsification technique establishes and maintains the connection to the most of the similar nearest neighbor data points. The nearest neighbor data objects belong to the similar cluster reduces the impact of the space and time complexity in SGDC model. Enhanced sparsification uses the linear graph cut procedure to easily reduce the time complexity rate. The SGDC model recursively partition the different density clusters, thereby representing the cluster through the graph structure.

The graph structure guarantees the link between the similar density data objects, thereby eliminating the unwanted links. The unwanted link eliminated in SGDC model, reduces the O (log (m/n)) space complexity, where ‘m’ cluster link with ‘n’ data objects. The connective link through the enhanced sparsification based on graph attains the O (log (m/n)) minimal time complexity. Overall work of the proposed system is to eliminate the overhead of the time and space by partitioning the vertices and edges of the data objects clustered through the graph cut procedure.

2.1.1 Linear Graph Cut Procedure

The linear graph cut procedure in the Enhanced sparsification compute the global minimum cuts of the dissimilar clusters to split up the edges. The edges are splitted using the splitting off operation in SGDC model. The main purpose of the graph cut is to improve the cluster quality rate. The linear graph cut procedure is described as,

\[ \text{Linear Graph (G)} = \ldots \ldots \text{Eqn (4)} \]

The graph contains the vertex with edges ‘ \', where the edges reduced as the redundancy level decreased in the proposed work. The main density properties are analyzed.
recurrently to perform the graph cut operation. The linear graph cut from the Globular Clustering is represented in Fig 3.

![Contiguity based Globular Clustering](image1)

![Sparsification Graph Cut](image2)

edges with low probabilities leads to a sparse sample. Each node in this sparsification graph corresponds to the subsystems formed by the edges in the Globular clustering group. The least probabilities are cut through the graph based enhanced sparsification. A graph property is referred to sparse, and enhanced sparsification denotes the partitioning of the edges. The partitioning which exactly cuts the less similar data objects, so that the space complexity is reduced. Linear cut speeds up the process and identify the high similar data objects with minimal time complexity.

3. EXPERIMENTAL EVALUATION

Contiguity based Globular Data Clustering (CGDC) Model is implemented in JAVA platform on avoiding the redundancy level on clustering the data objects. The experiment uses E. Coli Genes Data Set from UCI repository where the characteristics of the each gene are described to perform the clustering operation without any duplication of data objects inside the clusters. Sequence, homology (similarity to other genes), structural information, and functions are provided for easy evaluation of the experimental parameters with lesser time and space complexity. The data composed of several sources such as GenProtEC, and SWISSPROT. E. coli genes are connected to each other by forecasting ecoli_to_ecoli. All the data objects for a single gene are together with these delimiters. The gene functional classes are taken in the globular data clustering operation.

Homology search offer the gene information with more relationships for performing the hierarchical agglomerative clustering. The CGDC Model is compared against the existing improved differential evolution algorithm and Semantic smoothing Vector Space kernel (S-VSM) for text documents clustering. A biological data object is experimented on the factors such as object clustering quality rate, redundancy reduction rate, state space and time complexity measure, and nearest neighbor data objects connectivity linkage rate.
The data objects with the similar density level within the nearest neighbor achieve the higher quality rate. The higher quality rate is measured in terms of percentage (\(\%\)).

\[
\text{Quality Rate} = \frac{Q - r}{Q}
\]

‘Q’ denotes the data objects taken from the E. Coli Genes Data Set taken for the experimental evaluation. The experimental data object is assumed to be ‘100’. The ‘r’ denotes the radius of the nearest neighbor visited to perform the clustering operation. The clustering is carried out with the minimal repetition of the same data objects in the different clusters. The concept of minimizing the repetition is taken as the redundancy reduction rate. The reduction rate denotes the result with the reduced percentage count.

\[
\text{Redundancy Reduction} = \frac{(C1 \cap C2)}{\text{Total no. of data objects in } C1, C2}
\]

Here the computation carried out on the cluster ‘1’ and ‘2’, where common repetitive objects are analyzed. The redundancy level reduces on proposed work through computation with the overall data objects. The reduction in the space occupation by avoiding the lesser similar data objects leads to the minimal space complexity. The higher connectivity with higher speed of the proposed operation reduces the time complexity.

\[
\text{Minimal Time Complexity} = \text{Start time of process} - \text{End Time}
\]

The higher amount of linkage on the nearest neighbor data objects depends based on the higher ratio of the similarity level. The higher similarity level improves the data object connective ratio result.

4. RESULTS AND DISCUSSION

In section 4, CGDC model results are analyzed with the Improved Differential Evolution (IDE) algorithm and Semantic smoothing Vector Space kernel (S-VSM) model. The biological data object taken from E. Coli Genes Data Set for the experimental evaluation. The approach is based on exploiting the link structure based on the similarities of the data objects. The objects belonging to the same density level are more likely to group together in the proposed CGDC work.

<table>
<thead>
<tr>
<th>Relative Radius ‘r’</th>
<th>Object Clustering Quality Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDE Algorithm</td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
</tr>
<tr>
<td>10</td>
<td>0.85</td>
</tr>
<tr>
<td>15</td>
<td>0.81</td>
</tr>
<tr>
<td>20</td>
<td>0.72</td>
</tr>
<tr>
<td>25</td>
<td>0.69</td>
</tr>
<tr>
<td>30</td>
<td>0.62</td>
</tr>
<tr>
<td>35</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 1 Tabulation of Object Clustering Quality Rate

Dr. S. Adaekalavan
Table 1 and Fig 4 illustrate the object clustering quality rate based on the radius ‘r’ count. The clustering quality is maintained with higher rate in CGDC model by building the regular scaling factor. The scaling factor for the experimental work is assumed to be ‘100’. The radius is specified to identify in CGDC model with 5 – 14 % improved quality rate when compared with the IDE Algorithm [1]. For instance, the relative radius ‘r’ is about taken as the ‘5’, then clustering quality is achieved about 0.95 % in CGDC, whereas 0.90 % in IDE Algorithm and 0.92 % in S-VSM model. A data objects with the specified radius ‘r’ is considered as the neighborhood sharing function. The higher the relative radius ‘r’ range, the clustering of the data objects based on the density is also improved. In particular the neighbors which are similar to the ‘r’ radius are matched through the similarity measure in CGDC model, which improves the clustering quality by 3 – 9 % when compared with S-VSM model [2].

<table>
<thead>
<tr>
<th>No. of Data Objects</th>
<th>Redundancy Reduction Rate (reduced % count)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDE Algorithm</td>
</tr>
<tr>
<td>50</td>
<td>0.30</td>
</tr>
<tr>
<td>100</td>
<td>0.31</td>
</tr>
<tr>
<td>150</td>
<td>0.35</td>
</tr>
<tr>
<td>200</td>
<td>0.38</td>
</tr>
<tr>
<td>250</td>
<td>0.40</td>
</tr>
<tr>
<td>300</td>
<td>0.43</td>
</tr>
<tr>
<td>350</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 2 Tabulation of Redundancy Reduction Rate
The redundancy reduction rate is measured on the IDE Algorithm, S-VSM model and CGDC Model. As the data objects increases, the redundancy level is decreased through the contiguity based clustering. The globular cluster uses the fossils at the first stage to group the similar density nearest neighbor clusters. The key density value used for the easier way of clustering process and reduce the redundancy level.

CGDC model uses the massive complexes to group based on the nearest neighbor data objects density properties. Data objects density properties reduces the redundancy rate by 11 – 18 % when compared with the IDE Algorithm [1]. The density properties are examined clearly and form the star subsystems thereby reducing the redundancy level on grouping of data objects as shown in Fig 5. The nearest neighbor data objects with similar density ‘D’ are plotted to reduce the redundancy rate by 5 – 15 % when compared with the existing S-VSM model [2]. The group of different densities is constructed, which produce the reduced redundancy level in our proposed work.

| Object Size (KB) | Complexity Measure |  |
|-----------------|---------------------|-----------------|-----------------|-----------------|-----------------|
|                 | IDE Algorithm       | S-VSM model     | CGDC Model      |                 |                 |
|                 | State Space (bytes) | Time (sec)       | State Space (bytes) | Time (sec)       | State Space (bytes) | Time (sec)       |
| 2               | 1192                | 89              | 1150            | 80              | 1101            | 70              |
| 4               | 1320                | 104             | 1270            | 92              | 1220            | 80              |
| 6               | 1650                | 121             | 1596            | 111             | 1550            | 100             |
| 8               | 1950                | 135             | 1915            | 120             | 1885            | 106             |

**Fig 5 Measure of Redundancy Reduction Rate**
DATA OBJECT REDUNDANCY & COMPLEXITY REDUCTION USING CONTINGUITY BASED GLOBULAR DATA CLUSTERING

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2052</td>
<td>149</td>
<td>2002</td>
<td>134</td>
<td>1965</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>2179</td>
<td>165</td>
<td>2110</td>
<td>152</td>
<td>2066</td>
<td>139</td>
</tr>
<tr>
<td>3</td>
<td>2521</td>
<td>201</td>
<td>2485</td>
<td>189</td>
<td>2412</td>
<td>169</td>
</tr>
</tbody>
</table>

Table 3 Tabulation of State Space and Time Complexity Measure

The above table (table 3) describes the time and space complexity based on the object size. Data object size measures the space and time complexity in IDE Algorithm, S-VSM model and CGDC Model. Time complexity is measured in terms of seconds (sec) and space complexity is measured in terms of bytes.

![Graph](image)

Fig 6 (a) State Space Complexity Measure

Fig 6 (a) clearly describes the object size based on the space complexity measure. Enhanced Sparsification with graph cut reduces the space complexity by 3 – 7% when compared with IDE Algorithm [1]. The disconnection of the unwanted connected links reduces the search space complexity in CGDC model automatically. For instance the object of 2 KB it taken and performs the measure of the space through the graph cut. The CGDC model is 7.63% lesser when compared with IDE Algorithm [1] and also reduced to 4.26% lesser when compared S-VSM model. The reduced redundant level clustering integrates the enhanced Sparsification technique in CGDC model based on the graph form to measure the time and space complexity. CGDC model reduced to 1 – 4% when compared with existing S-VSM model [2].
Fig 6 (a) clearly describes the time complexity measure based on the object size. The enhanced sparsification based graph cut is the simpler and faster system which easily estimates the edge connective link of the data objects. The easier processing and discarding of the dissimilar data objects reduce the time complexity by 15 – 23 % when compared with IDE Algorithm [1]. The linear graph cut procedure is used to produce the sparsifiers in CGDC model with 8 – 13 % minimal time factor, when compared with existing S-VSM model [2]. Enhanced sparsification uses the linear graph cut procedure to easily reduce the time complexity rate in proposed work.

<table>
<thead>
<tr>
<th>No. of Iterations</th>
<th>Nearest Neighbor Data Object Connection Linkage Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDE Algorithm</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>82</td>
</tr>
<tr>
<td>7</td>
<td>83</td>
</tr>
<tr>
<td>8</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 4 Data Object Connection Linkage Rate Tabulation
Fig 7 illustrates the nearest neighbor data object connection based on the iterations. The nearest neighbor data objects similarities are observed through the density properties for the effective linkage. On the different iterations the linkage percentage of the IDE Algorithm, S-VSM model and CGDC Model are analyzed. The Most edges in a density based Globular graph are well connected, and hence 12 – 21 % improved linkage rate when compared with the IDE Algorithm [1]. The nearest neighbor data objects belong to the similar cluster improve the more connectivity. The SGDC model recursively partition the different density clusters, thereby representing the connection linkage through the graph structure. The connection linkage is improved by 2 – 5 % when compared with existing S-VSM model [2].

Finally, Contiguity based Globular Data Clustering (CGDC) Model main goal is to integrate the nearest neighbor data objects cluster with enhanced Sparsification technique. The data objects with similar densities carried out the linear graph cut procedure on reducing the time and space complexity.

5. RELATED WORK

Clustering ensure the data objects to achieve the higher similarity rate with extensive applications. Optimization Objective function is used in [15] to update the formula and convergence between the cluster information. The weaken cluster information is equalized to zero value. Spatio Temporal data objects issues and difficulties associated with data depiction, analysis, mining and visualization of information are presented in [5]. A variety of data mining tasks are association rules; classification clustering for determining information from spatiotemporal datasets is examined and reviewed. The cluster quality measure is not recognized to all of the spatiotemporal data mining tasks.

Cluster-based Temporal Mobile Sequential Pattern Mine (CTMSP-Mine) find out the cluster based mobile patterns [8]. CTMSP-Mine forecast the succeeding mobile behaviors. In
CTMSP-Mine, user clusters are constructed with novel algorithm. Cluster-Object-based Smart Cluster Affinity Search Technique (CO-Smart-CAST) measures the similarity among users. The clustering scheme is not used further to enhance the precision by predicting the user performance. Web People Search approach via connection analysis in [14] clusters the web pages based on the relationship of the diverse peoples. Web People search the information on the different diversity of semantic information. The semantic data is taken out from web pages, such as name unit, a hyperlink which is enclosed by namesakes on the web pages. The information stored in the top-k web pages are not upon the utmost quality mark for cluster processing.

Automatic annotation approach in [16] first aligns the data units on a web result page into different groups. Clustering-based shifting technique employed in this existing work arranges data objects into different group set so that the data objects within the same group have the similar semantics. Clustering-based shifting technique fails to incorporate the dynamic clustering to improve the performance rate. An automatic and dynamic database clustering technique as illustrated in [10] actively re-cluster a database. Database with little intervention of a database administrator (DBA) sustain an adequate query response at all the time of query processing.

Logical Analysis of Data (LAD) and Shadow Clustering (SC) as demonstrated in [3] adopt the effective heuristic procedure to retrieve the logical data. The retrieval of the logical data uses the AND-OR expressions. LAD consists of the breadth first search with the whole prime implicates to reduce the time complexity rate. The computational cost required for LAD is of higher rate, which works on the larger set of the data object domain. An encrypted index-based technique as demonstrated in [6] transforms the analyzed logical data to the service provider with minimal cost factor. The query object is an outlier whose Nearest Neighbor (NN) resides in a dense cluster location. At query time, the query user identifies the query object and then presents the altered query object to the server for similarity search.

6. CONCLUSION

The rapid growth of data clustering operation motivated us to develop the model for reducing the redundancy level on the clustering of the data objects. Contiguity based Globular Data Clustering (CGDC) Model is employed largely to improve the cluster quality rate. More focus is shown on clustering the data objects using the Enhanced Sparsification technique. Enhanced Sparsification technique achieves the goal of reducing the time and space complexity measure using graph cut procedure. CGDC model keeps connection link with the most similar nearest neighbor data objects of the varying domain knowledge. The less similar data object connection are discarded in the Enhanced Sparsification based on the graph, by this means reduce the space complexity measure. Theoretical analysis and experimental result shows that the CGDC model attains the maximal object clustering quality rate, and nearest neighbor data objects connectivity linkage rate. The time complexity is reduced to 10.95 % lesser when compared with present work and also the space complexity is reduced about 5.47 % percentage.

REFERENCES
DATA OBJECT REDUNDANCY & COMPLEXITY REDUCTION USING CONTINGUITY-BASED GLOBULAR DATA CLUSTERING


