ISSN 2277-3061

Enhancement of Sales promotion using Clustering Techniques in Data Mart

Dr.S.Suguna[1], M.Vithya[2]
[1] Assistant Professor, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2, kt.suguna@gmail.com
[2]Lecturer, Sri Meenakshi Govt. Arts College for Women (A), Madurai -2, vithyagopal20@gmail.com

Abstract

Clustering is an important research topic in wide range of unsupervised classification application. Clustering is a technique, which divides a data into meaningful groups. K-means algorithm is one of the popular clustering algorithms. It belongs to partition based grouping techniques, which are based on the iterative relocation of data points between clusters. It does not support global clustering and it has linear time complexity of $O(n^2)$. The existing and conventional data clustering algorithms were not designed to handle the huge amount of data. So, to overcome these issues Golay code clustering algorithm is selected. Golay code based system used to facilitate the identification of the set of codeword incarnate similar object behaviors. The time complexity associated with Golay code-clustering algorithm is O(n). In this work, the collected sales data is pre processed by removing all null and empty attributes, then eliminating redundant, and noise data. To enhance the sales promotion, K-means and Golay code clustering algorithms are used to cluster the sales data in terms of place and item. Performance of these algorithms is analyzed in terms of accuracy and execution time. Our results show that the Golay code algorithm outperforms than K-mean algorithm in all factors.

Keywords

Pre-processing; K-means clustering; Golay code clustering; Sales Promotion.

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 15, No. 2

www.ijctonline.com, editorijctonline@gmail.com

ISSN 2277-3061

1. Introduction

Cluster analysis or clustering is task of grouping a set of objects (item) in such a way that objects in same group (places) are more similar to each other than to those in other groups. Popular notions of clusters include groups with small distance among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Cluster analysis itself is not one specific algorithm, but general task to be solved. Various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them can achieve it [1]. The appropriate clustering algorithm and parameter setting includes values such as the distance function to use, a density threshold or the number of expected clusters depends on the individual data set and intended use of the result. A Vladimir Estivill-Castro model [2] is key to understand the difference between various algorithms. Typical cluster models include connectivity models for hierarchical clustering, builds model based on distance connectivity, centroid models for K-means algorithm represents each cluster by a single mean vector. Today's [3] central issues in the business supply make it imperative develop new concepts to reduce the emerging costs and ensure high quality standards. Applying ICT (Information Communication Technology) and especially online sales- technologies that offer the chance to optimize sales data transfer is regard as the promising strategy, when developing cost saving concepts. The organizations have more branches for selling same product. Sales promotion is any initiative undertaken by an organization to promote an increase in sales, usage or trial of a product or service. Today technologies produce petabytes of information that is big data presents new challenges related to time and memory complexity. Therefore, a new clustering techniques model is required to processing big data faster and efficiently as possible.

The paper is organized as follows: Section 2 deals with survey of related works. Section 3 deals with pre-processing process such as removal of null and redundant values. In Section 4, we analyzed K-means algorithm. Section 4 introduced and analyzed the Golay code-clustering algorithm. Section 5 presents the performance evaluation of K-means and Golay code algorithms for sales data collected. Finally, section 6 gives conclusion.

2. Survey of Related works

In large datasets, finding relationships between one cluster's attributes to another is not a simple task since the data is not organized or most of the time classified with significant biases [4]. There are two types of clustering method 1) Hierarchical clustering proceeds successively by either merging smaller cluster into larger ones, or by splitting larger clusters. 2) Partition clustering: To attempts, directly decompose the data set into a set of disjoint clusters. Each object is a member of the cluster with which it is most similar[5]. The basic purpose of K-mean clustering algorithm is for clustering only scalar data. K-mean depending on initial centroid condition, which causes the algorithm, may coverage to suboptimal solution[3]. Spatial algorithm should be effective in processing data with noise and outliers in the geographical datasets [11]. Goaly code (7,12,23) clustering algorithm is able to classify all 23-bit codeword into a large number clusters. The maximum Hamming distance for each cluster is either 7 or 8 and the total number of bit positions that have common bit values for each cluster is either 16 or 15 [6]. A unique binary label can be applied to the data item, which in turn efficiently allows the system to classify and cluster the data item accurately and added error-correcting bit to each data sets [7].

3. Pre-processing

Data pre-processing is an important step in the data mining process. If there is much irrelevant and redundant information present or noisy and unreliable data then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction etc [8]. We proposed to enhance sales promotion using clustering technique application.

The pre-processing process handled in our work is described as below:

1. Remove all null or empty attributers.
2. Eliminate redundant data.
3. Handling noisy data.

ISSN 2277-3061

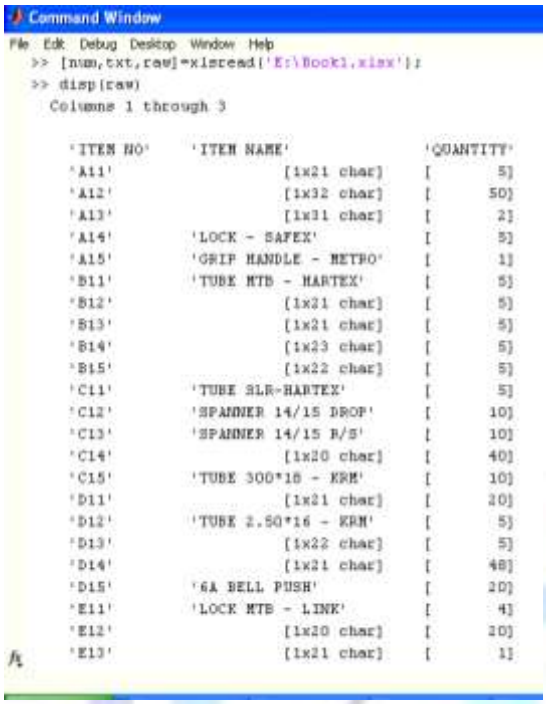


Figure 1: Data before pre-processing (with noisy data)

Figure 2: Data after pre-processing (without noisy data)

Original dataset shown in Fig.1 includes the fields item no, item name, quantity. The quantity field is considered as noisy data since there is no role for this field in our proposed work. So, we removed this field from the dataset and the resultand set is shown in Fig2. Similarly null attributes are also handled and the results are shown in Fig.3 and Fig.4.

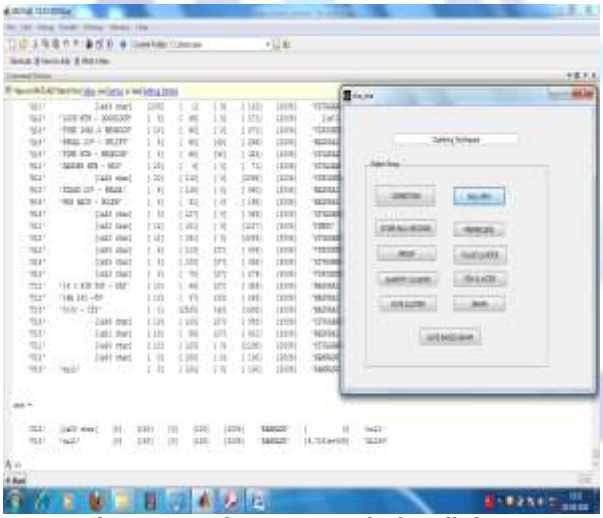


Figure 3: Before Removal of null data

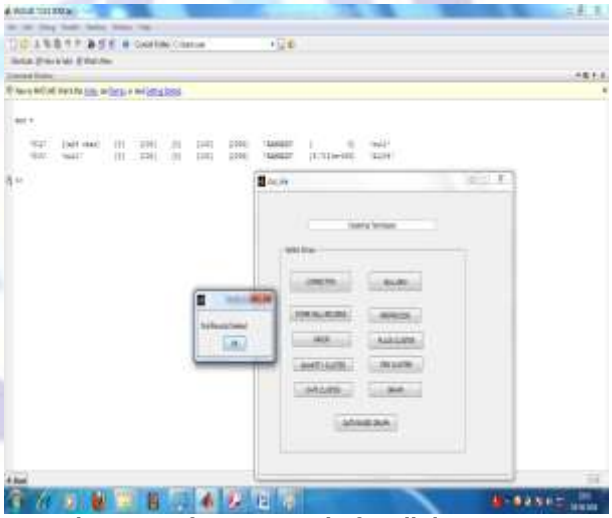


Figure 4: After Removal of null data

4. K- Means Clustering

K- Means clustering algorithm is a partition based cluster analysis method [7]. First select k objects as initial cluster centers then calculate the distance between each cluster centre and each object and assign it to the nearest cluster, update the averages of all cluster repeat this process until the criterion function converged [9,10,11]. It can also be used as a initialization step, for more computationally expensive algorithm like searching vector Quantization or Gaussian mixtures, this giving an approximate separation of the date as a starting point and reducing noise present in the data set.

ISSN 2277-3061

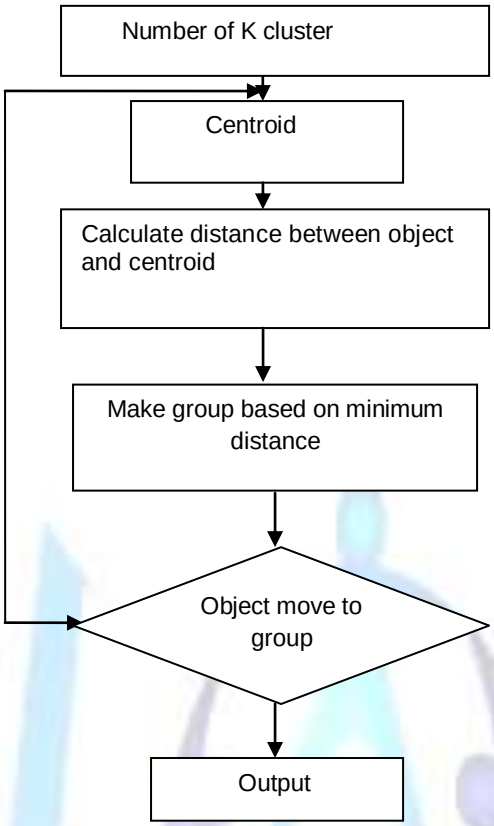


Figure 5: Flowchart-k- means algorithm

The flow chart is explained as follows:

Step 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group of centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all object have been assigned recalculate the positions of the K centroids.

Step 4: Repeat steps 2 and 3 until the centroids can longer move. This produce a separations of the objects into groups from which metric to be minimized can be calculated.

Different matrix can be used to calculate this similarity towards the Euclidian distance d£= $\sqrt{\Sigma i}$ ∧= [(cx - xi)]. Here C is the cluster centre, x is the case it is compared to, i is the dimension of x or c being compared and K is total number of dimensions [5, 3].

Algorithm

Input: N objects to be clustered{x1, x2, .xn} and the number of cluster K.

Output: K cluster and the sum of dissimilarity between each object and its nearest cluster center is the smallest.

Select K objects as initial cluster centers (m1, m2, … mk)

Randomly initialize C

```
{
        For each xiϵ X
        {
                m(xi)=argmin distance(xi,ci)
                jϵ {1..n}
        }
        While m has changed
        {
```

```
                For each I ϵ {1…n}
                Recomputed Ci as the centroid of
                {x|m(x)=i}
        }
        For each xi ϵ X
        {
                M (xi) =argmin distance (xi,ci)
                jϵ {1..n}
        }
}
```

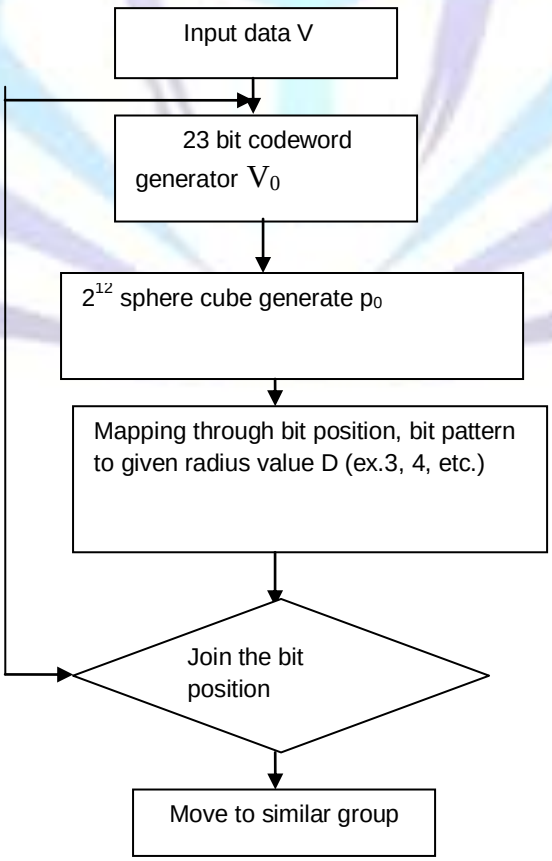
KMeans clustering algorithm is very fast, robust and easily understandable. If the data set is well separated from each other data set, then it gives best result. K-means also produce tighter clusters, if the clusters are globular.

Issues with K- Mean clustering

- Difficult to predict k value.
- With global cluster, it did not work well.
- Different initial partition can result in differed final cluster.

5. Golay code Clustering

Golay code is superior to other conventional clustering [12]. Golay code [8, 9] clustering has two types first one is Binary Golay Code and it represented by 0's and 1's Second one is Ternary Golay code. Binary Golay code [9] has divided into Extended Binary Golay Code and Perfect Binary Golay code. We proposed to use only Perfect Binary Golay code G_{23} is coded as 12 bit of data in a 23-bit word in such a way that any 3-bit error can be connected or any 7-bit error can be detected along with Hamming distance. Golay code clustering algorithm [8, 13] is able to classify all 23 bit codeword into a large number of cluster. Maximum Hamming distance for each cluster is either 7 or 8 and the total numbers of bit positions that have common bit values for each cluster are either 16 or 15. This is particularly important, as bit positions may have low, Hamming distance alone does not mean that two codeword's are similar with such a clustering algorithm.



ISSN 2277-3061

The flowchart is explained as follows:

Step1: Input value

Step2: 23 bit codeword V. V resides in a codeword sphere P and center vector of P is P_0

Step3: Hamming distance D between v and P_0 for each of the codeword. Here D is 3.

Repeat step2 and step3 until all bit positions are mapped. D is equal to V cluster corresponding Golay code group.

In case of Golay code has the total number of possible outcomes is 2^{23}[7]. The entire set of 23 bit vector considers them as the vertex of 23 dimensional binary cubes is partitioned into 2^{12} **s**pheres**.** Each sphere has the radius of 3. Thus 23 bit streams are mapped to 12 bit centers of this sphere giving the ability to create some dissimilarity in the 23 bit stream [14]. Hamming bound is define as for any code c = (n, k, d) with d<2e+1. Hence n is the block code length, k message length (partition), d is the distance different matrix can be used to calculate this similarity to wards the Euclidian distance [15].

Algorithm

Input: N objects to be clustered {s1, s2, ….sn} the number of cluster K.

Output: D cluster and the sum of dissimilarity between each object and its nearest cluster center is the smallest. Select D objects as initial cluster centers (ch1, ch2 ,..chm)

```
Def hamming _distance (s1, s2):
#Return the hamming distance between equal-length sequences
  If len(s1)!= len (s2);
    {
      Raise ValueError("Undefined for sequence of unequal length")
    }
  Else Return sum (ch1! = ch2 for ch1, ch2 in lis( s1, s2))
```

In error correction schemes, of Golay code algorithm a number of parity bits are added to create code word. If some number of distortions happens during transmission, then the redundant parity bit can be used to restore the original data and Golay code clustering provides fast access compared to the existing clustering algorithm [16].

Issues with Golay code Algorithm

- The Golay code is obviously not able to encode a large amount of data is one codeword.
- Replication in storage

6. Performance Evaluation

We take some of the parameters to compare K-means and Golay code algorithms and tabulated as follows:

Algorithm	Type of clustering	Dimensionality of data	Time complexity	Distance measurement
K- means	Local cluster	spherical	$O(n^2)$	Euclidean
Golay code	Local and Global cluster	cube	O(n)	Hamming

Table 1: Comparative Analysis

Here we experiment on 3000 records to Enhance sales promotion in data mart towards clustering the data through place and item sales and depicted in figure7. X axis represent as place and Y axis represents as item sales. The data sets and experiment result achieved using K-Means and Golay code clustering are described below, and created three groups of these data with each group having 1000 records each of similar sales item at different places. In this work Pentium dual core processor with 2GB RAM and 300 GB Hard disk system is utilized and it is loaded with windows operating system and MATLB software to analyze the performance of K-means and Golay code algorithms. Analysis of sales data sets is shown in Fig. 7.

ISSN 2277-3061

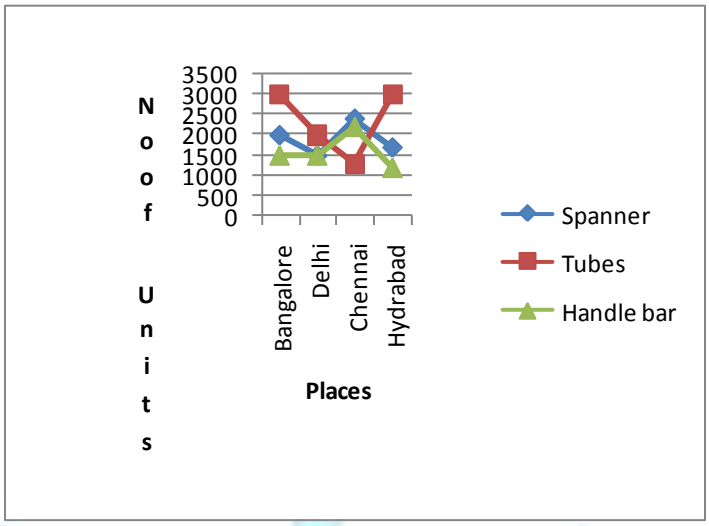


Figure 7: Analysis of sales data sets

The precision accuracy is calculated by diving the number of relevant data points correctly put in a particular group (by running the cluster algorithm) by the total number of all data points put in that group and the results are shown in Table2 and Fig.8.

Data sets	Average precision Accuracy in %	
	K-means	Golay Code
500	52.75	56.7
1000	57.35	62.75
1500	62.03	67.56
2000	59	68.25
2500	63.45	70.42
3000	65.23	75.45

Table 2: precision accuracy measure between K-means and Golay code

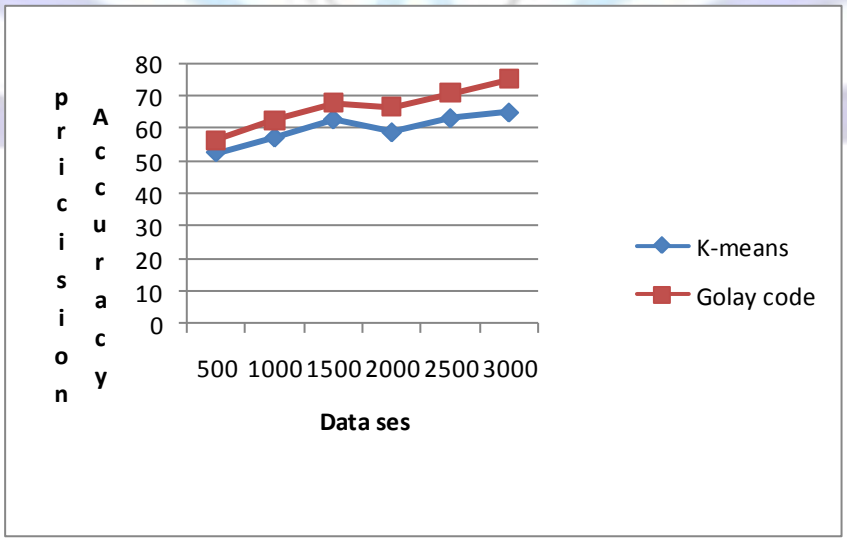


Figure 8: Analysis of Precision Accuracy

We have also analyzed the time taken by K-means and Golay code for clustering the sales data. The results are shown in Table 3 and in Fig.9.

Record sets	Time in ms	
	K-Means	Golay code
500	487	22
1000	542	23.28
1500	602	24.53
2000	686	26.19
2500	706	26.58
3000	798	28.25

Table3: Analysis of Execution time

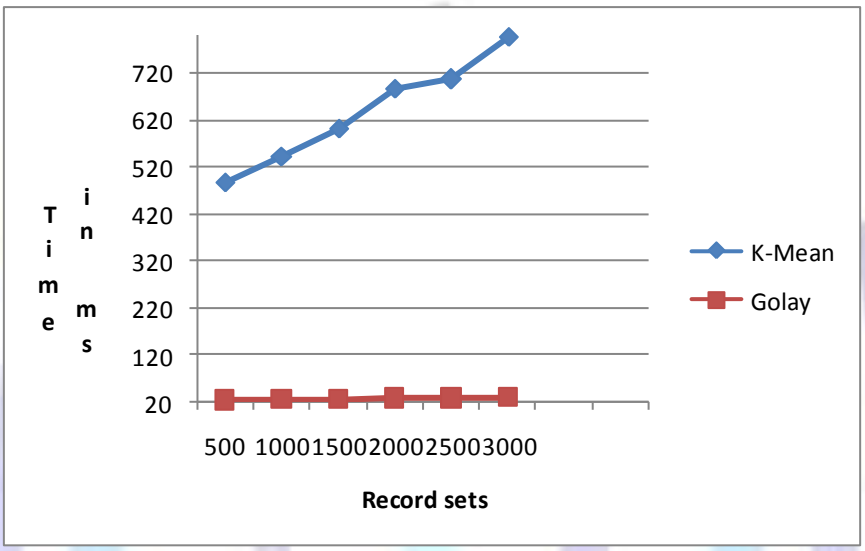


Figure 9: Analysis of Execution time

Our analysis shows that the Golay code algorithm outperforms than K-means algorithm in terms of execution time and accuracy for sales data analysis.

7. Conclusion

We first pre-processed the sales data by removing all null and empty attributes, eliminating redundant, and noise data. The K-means and Golay code clustering algorithms are used to cluster the sales data in terms of place and item to enhance the sales promotion. Finally, K-means and Golay code algorithms performances are compared in terms of accuracy and execution time. We proved that Golay code algorithm outperforms than K-means in all the factors. In future, big data tools Hadoop and Mapreduce are proposed to apply for business data domain that involves huge amount of data.

References

[1]. http:\\ www.clusterindex.in

[2]. Poites vera, Bhavesh oza, "**A survey on K-Mean Clustering and Practice Swam Optimization",** International Journal of Science and Morden Engineering, Vol: 1, Issues: 3, Feb-2013.

[3]. http:\\\www.clustering role in business.htm

[4]. Ganapathy mani, Nima Bari, DuoDuo Liao, Simon Berkovich, **"Organization of Knowledge Extraction From Big Data System"**, IEEE 2014 International Conference computing for Geospatial Research and Applications.

[4]. Laurence Morisselte and Sylvain Chartier, "**THE K-Means Clustering Technique General Consideration and Implementation in Mathematica",** Tutorial in Quantitative Methods for Psychology.

[5]. Shalove Agarwal, Shashank Yadavand Kanchan singh, "**K -Mean Verssus K-Mean++ Clustering Techniques ",** in IEEE 2012.

[6]. Hong jun yu, Tao Jins, Dechang Chen and Simon, Y.Berkovich, **"Golay Code Clustering for Mobility Behaviour Similarity Classification in Pocket Switched Network",** Journal of Communication and computer 2012.

[7]. Laurence Morisselte and Sylvain Chartier, "**THE K-Means Clustering Technique General Consideration and Implementation in Mathematica",** Tutorial in Quantitative Methods for Psychology.

[8]. http:\\ www.preprocessing Wikipedia.index.html

[9]. Nima Bari, Duoduoliao simon, Bekovich, " **Organization of Meta knowledge in the form of 23 bit Templates for Big data processing**", IEEE 2014 International Conference computing for Geospatial Research and Applications.

[10]. R.Manikandan, "**Improving Efficiency of Textual Static web Content Mining Using Cluster Techniques",** Journal of Theoretical and Applied Information Technology, vol: 33, Nov 2011.

[11]. Nerti Arora, Magesh Motwani, "**Sum of Distance Based Algorithm for Clustering Web Data**", International Journal of computer Applications Feb2014.

[12]. http:\\www.Golay code and example.html

[13]. Dr.Chandra. E, Anuradha V.P, **"A Survey on Clustering Algorithm for Data In spatial and Data Base Management System",** International Journal of computer Applications, vol : 24, Issue : 9, July 2011.

[14]. Navjot Kaur, Jaspreet Kaur, Sahiwal Naveet Kaur, "**Efficient K-Mean Clustering Algorithm using Ranking Method in Data Mining",** International Journal of Advanced Research in computer Engineering and Technology , Vol : 1, Issue :3, May 2012.

[15]. Eyas E1-Qawasmeh, Maytham Safar and Talal Kanan, "**Investigation of Golay Code(24,12,8) Structure in Improving Search Techniques ",** International Arab Journal of Information Technology, Vol: 3, Jan 2011.

[16]. http:\\ www. Hamming Distance index.html

.