Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.421 – 428

RESEARCH ARTICLE

Vector Quantization for Privacy during Big Data Analysis and for Compression of Big Data

Dr. D.Aruna Kumari[1], N.Tejaswani[2]

aruna_D@kluniversity.in [1]

[1] Associate professor, [2] IV/IV B.Tech

[1,2] Department of ECM, KL University

AP, INDIA – 522502

ABSTRACT: ***Now a day's, data on the web is increasing like water in the ocean. Big data is one of the emerging areas that deals with large velocity of data. "Big data is not defined in single world data sets that are too large and complex to manipulate or interrogate with standard methods or tools". Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. Data warehouses deals with multiple data sources and getting integration where as Big data covers whole WWW data that can be called as cloud data. The Big Data challenge is becoming one of the most exciting opportunities for the next years. There are many problems and applications in big data. In this paper we are going to discuss privacy issues related to big data and reducing the size of big data using compression techniques.***

Keywords: big data, privacy, volume, compression

1. INTRODUCTION

Big data is one of the emerging research area in computer science. In many applications like business intelligence, military department, health care, social networks the data size in increasing drastically more i.e volume is more and measured in terabytes or petabytes. Big data is collection of data sets from wide variety of sources , it is very difficult to analyze such kind of data. In other words big data is that which many users can be maintained and many databases can also be maintained. Big data is similar to data warehouses but data bigger, consequently requires different approaches like techniques, tools, architecture. Data bigger in the sense, if we take data of facebook or twitter data, we can see millions of posts, photos, videos per day. We can call such data as Big data.

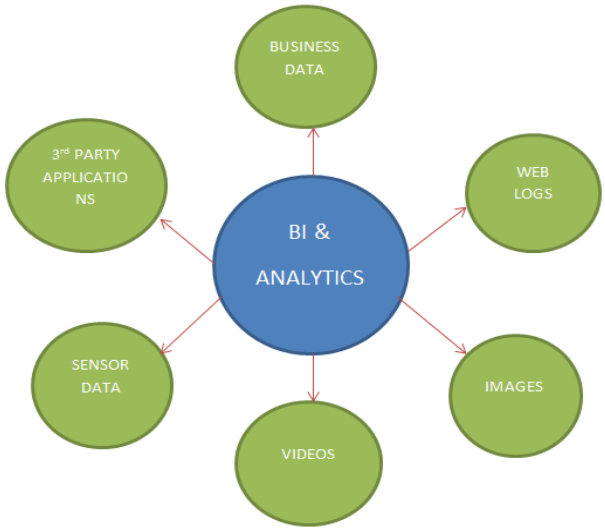
Big data is popular term used to describe exponential growth and availability of growth, both structured and unstructured data. Both big data may be as important to business and society as the internet has become more important. It is unstructured because many texts and images are not stored in structured form and so it is in unstructured format. Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile.
Generally big data is more scalable and it is measured in pet bytes. It s a combination of various data warehouses and massive amounts of different users. Big data systems are evolution of rdbms, it is a new ways to store the data in different forms.

Big data has many different forms, different platforms, tools, and technologies. They are quite difficult to understand and use by analytics. For conversion of simple text data into multimedia data and relational data we need new data types, for that in olden days we have BI ANALYTICS(business intelligence) system.

In that BI ANALYTICS we have business data, web logs, videos, images, 3rd party applications, sensor data.

Business data is that which is action of systems run the businesses like erp system, customer relationship system etc.

3rd party applications is that like twitter, Facebook, marketing chains and large data. we use it for business should be in BI ANALYTIC form only.



1.1 Scenarios of Big Data

Why not Big data:

Big data is not transactional in nature, and not simple or easy, not structured data ware house, not single platform, not easy or fast for analytics.

1.2 Big Data Characteristics:

Some of The key features of big data is scalability, volume, variety, velocity, variability, complexity, veracity, validity, volatility.

1. *Volume*

Big data implies enormous volumes of data. one of the feature of Big data is volume its wide range of data occupation. Data is generated by different sources like social media , satellite data, geographical data, Hospital data , military data , web data,..etc. we can see distributed data base as one of the example source for big data where data is scattered around the world ,so it can contain huge of volumes of web data.

2. Variety

Big data's unique feature is Variety. Variety is nothing different types of data, the data can contains multimedia, unstructured data, audio data…etc. Generally data is structured or unstructured. Previously data is collected from the sources like spreadsheets and databases. Now Big data consists of variety of data like emails, pictures, teaching material, videos, monitoring devices, various documents, relational data bases, multimedia data bases, 3d information. This kind of unstructured data creates complexity for storing the information and retrieving the information. Performing data mining techniques on big data is one of crucial step.

3. Velocity

Another characteristic is velocity. Its massiveness , the flow of data is both continuous and dynamic. But , because of this velocity researchers and business people can take valuable decisions for their profitable purpose . Facebook , twitter data is example for continuous and dynamic data. Maintaining dynamic data requires complex hardware requirements. one of the big challenges is velocity, if we handle dynamic data on the web as well as on the data store, we can get more profits upon analyzing usefull and hidden and proper data.

4. Veracity

As Big Data contains large volumes of variety data, there is a chance for having of noisy or unwanted data. If that noisy or unwanted data doesn't produce any meaningful information during data analysis step , then we can remove that data. Veracity is like unnecessary and unknown data. We can remove such type of data. It is like data preprocessing step in data mining. One of the problem of Big data is volume, we can decrease it by doing data processing and null data elimination.

5. Validity and Volatility

Reasons for having large volumes of data are its validity. If the validity is mentioned like clearing of previous data for some time then volume can be reduced and analysis takes faster in time. Meaning is also important; while fixing the validity meaning should be tested other the data should not be cleared.

1.3 problems in big data:

There are many problems in big data. Some of the problems in big data are:

1. Echo chamber effect
2. Google flu trends, tools in big data can be easily gamed,
3. correlations, especially subtle correlations, prone to giving scientific-sounding solutions to hopelessly imprecise questions,
4. scrubbing data , Etc

1.4 Organizing Big Data:

Organizing the data is one of the very important step in data warehousing which is called as data integration. Data from multiple sources will be collected and put together in one location is called as data warehouse. When coming to the big data , it is a complex task as big data is growing at a very faster rate and it is difficult to measure also. Big data contains large variety of data formats like unstructured and structured.

There are some tools which are useful for organizing the big data . hadoop is new tool that supports large values of data , that can be integrated and organized, so that retrieval and storage tasks can be performed well. Hadoop distributed file is having storage system for maintaining web logs.

2. PROPOSED APPROACH FOR COMPRESSING BIG DATA USING LBG ALGORITHM

It is been proposed to perform clustering on original data and transformed data for evaluating the performance of the proposed approach. Generated clusters will be compared and analysed. The proposed approaches are flexible so that, any data mining task can be applied on transformed data and can expect accurate secure results

Generally, a large training sequence or large number of input vectors will produce a effective codebook, codebook plays important role. Codebook is used for encoding and decoding the codes and can be used in data transmission and compression. Limitation in designing the codebook is optimization problem. There is several number of codebook design algorithms were there like Mean-distance-ordered Partial Codebook Search (MPS), Principal Component analysis (PCA), Generalized Lloyd Algorithm (GLA). Generally GLA yield only locally optimized codebooks. More advanced methods, such as deterministic annealing and genetic optimization have promised to overcome the drawback of local optimal at the expense of greater computational requirements and memory requirements[7].

A VQ is just like an approximator. The Key point is just ``rounding-off" (say to the nearest integer). An example of a one-dimensional VQ is shown below:

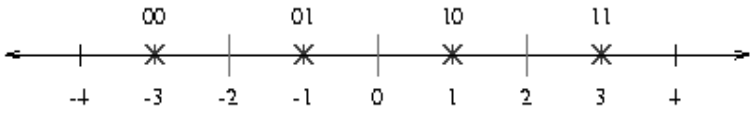


Figure 2.1 **:** one-dimensional Vector Quantizer

Two dimensional vector quantizer is shown in bellow. -3 is represented for all the numbers which are less than -2. (i.e, approximated to -3). Numbers between -2 and 0 are rounded to -1. Numbers between 0 and 2 are rounded to +1. All the numbers greater than 2 is approximated by +3. Note that the approximate values are uniquely represented by 2 bits. It is called as one -dimensional, 2-bit VQ and it has a rate of 2 bits/dimension.

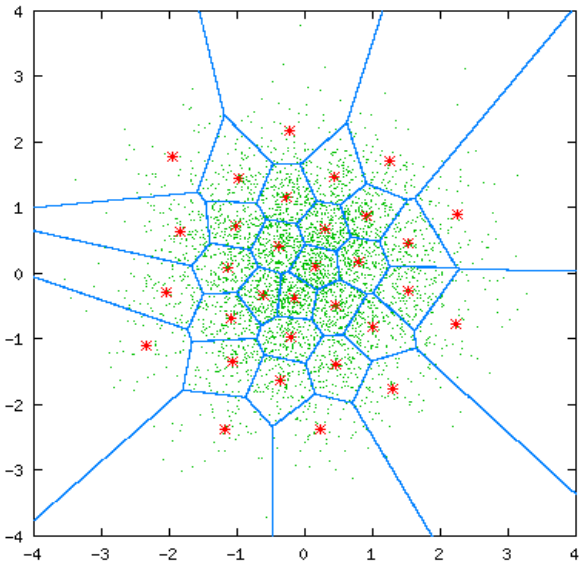


Figure 2.2: Two-dimensional Vector Quantizer

An example of a Two-dimensional VQ is shown above. Here, pair of numbers falling in a particular region is grouped by a red star associated with that region. In the Figure 4.3, there are 32 regions and 32 red stars, each of which can be uniquely represented by 4 bits. Hence, this is a two-dimensional, 4-bit VQ and its rate is 2 bits/dimension[10][9].

Figure:2.1 shows the example for Basic clustering operation . 3 clusters are represented in the figure, cluster one has two objects, cluster two has 6 objects and cluster 3 has 7 objects. K means algorithm can be analyzed as follows;

K-means algorithm:

Step1: Select the value of K , which is nothing but number of clusters.
Step2: Choose k objects in a random fashion. These will become the initial centroids.
Step3: The objects which are closer to the cluster will be assigned or moved to that group (also called cluster) for which the objects are nearest to the centroid.
Step4: Recalculate the centroids for newly formed k clusters.
Step5: Repeat steps 3 and 4 until all the objects have been moved.

The aim of K-means clustering is to place the objects which are closer to its nearest cluster with the help of centroid condition and nearest neighbour search condition.

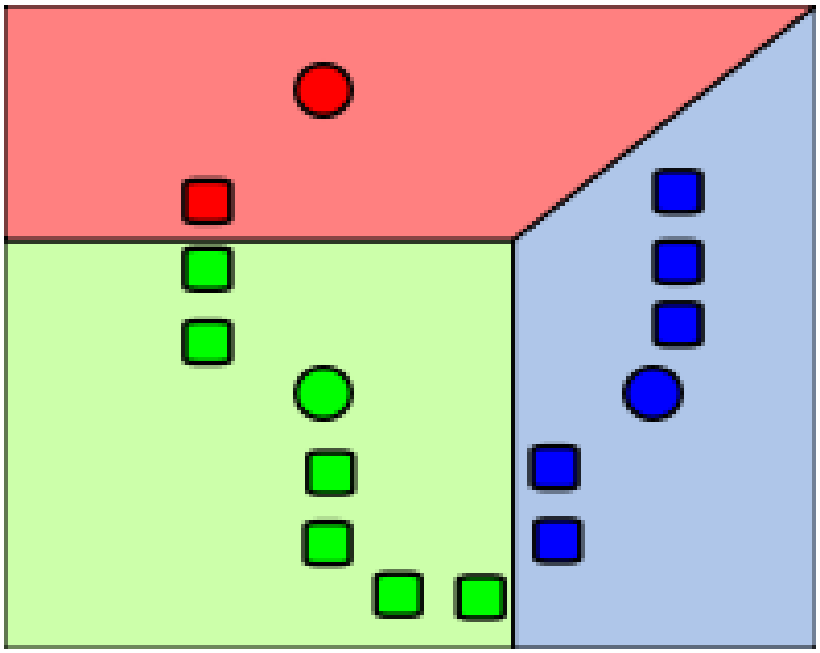


Figure:2.3: example on clustering

3. Proposed Approach for Privacy for Big Data: quantization

When the data is given for analysis, we can analyze the hidden predictive data but sensitive data should not be mined or revealed. Doing so, data transformation is introduced where original data will be transformed to other fashion, from which sensitive data will not be revealed when an analysis is takes place.

Data Transformation

1. encoding
2. Code book generation : Quantization
3. Decoding

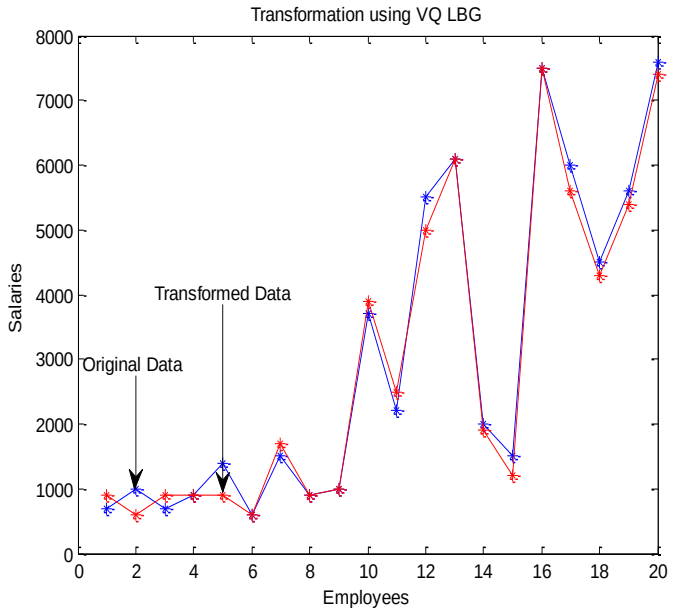


Figure 3.2: Transformation of Original data with VQ32-LBG

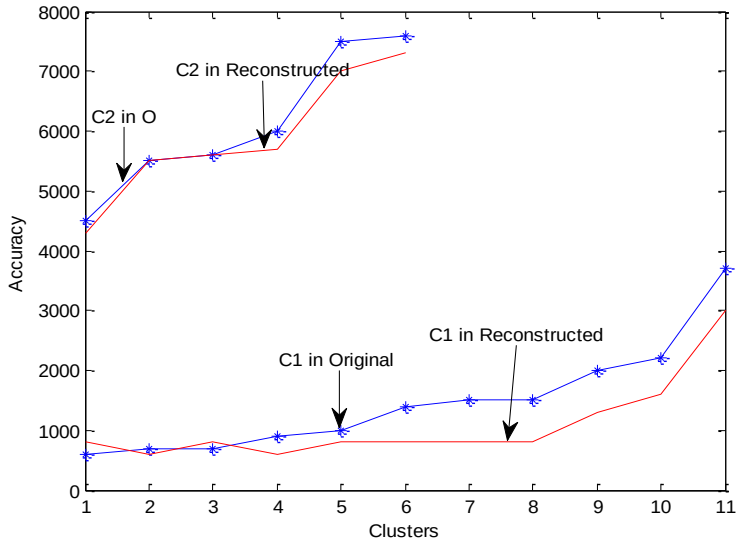


Figure: 3.3 Comparison of original clusters with VQ LBG clusters.

4. Compression of Big Data

Proposed Algorithm:

INPUT: Data set

Step1: data set is taken as input

Step2: LBG algorithm is applied

Step3: output which is in the form of compressed version of original data set which represents centroids of its nearest data .

OUTPUT: Only centroids not complete data, which is enough for analysis

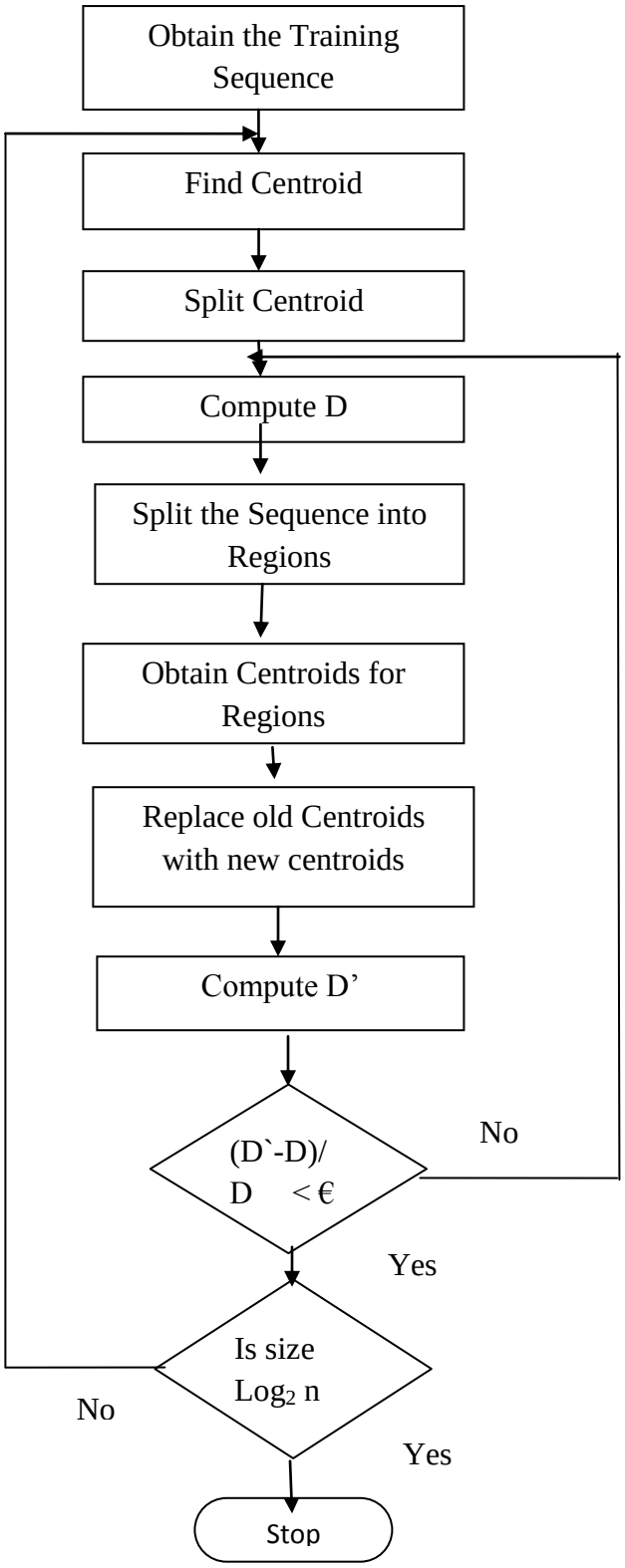


Figure 4.1: LBG Algorithm

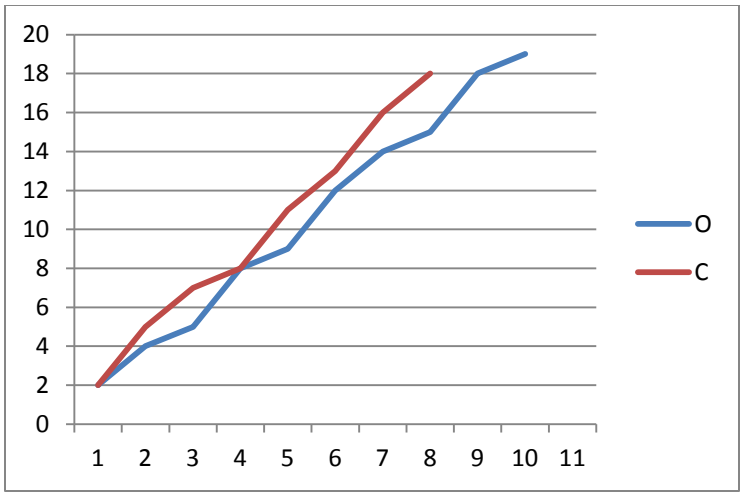


Figure 4.2: Original data and compressed data

O: - Original Data
C:- Compressed Data
C represents the compressed form of original data which contains centroids .

5. Conclusions

As Big Data is having many applications, it became popular and has many research directions. This paper proposed how to preserve privacy during big data analysis and reducing the size of big data which is very important. Another way of reducing the volume of big data is clearing the unwanted or long lasting data.

REFERENCES:

1. **D.Aruna Kumari**, Dr. K.Rajesekhara Rao, M.Suman published a paper on " Privacy preserving data mining using LBG Design Algorithm " in **Springer-Verlag Berlin Heidelberg**, International Conference on CSPE 2012, Chennai , Dec-2012.
2. **D.Aruna Kumari** , Dr.K.Rajasekhara Rao,M.Suman "Privacy preserving data mining: LBG Algorithm" in International Journal of Database Management Systems(IJDMS) ISSN : 0975-5705 (Online); 0975-5985(Print).
3. **D.Aruna Kumari,** Dr. K.Rajesekhara Rao, M.Suman published a paper on "Vector quantization for privacy preserving clustering in data mining" in Advanced Computing: An International Journal (ACIJ -Nov 2012)
4. **D.Aruna Kumari**, Dr. K.Rajesekhara Rao, M.Suman Tharun Maddu Published a paper on"Compression in privacy preserving data mining" in International Journal of Advanced Computer technology(IJACT ISSN : 2320-0790). April 2013.
5. **D.Aruna Kumari**, Dr. K.Rajesekhara Rao and M.Suman Published a paper on "Privacy preserving clustering data mining using VQ code book generation" in **AIRCCJ Computer Science and Information technology (CS &IT) Proceedings**
6. C. W. Tsai, C. Y. Lee, M. C. Chiang, and C. S. Yang, A Fast VQ Codebook Generation Algorithm via Pattern Reduction, Pattern Recognition Letters, vol. 30, pp. 653{660, 2009}
7. K.Somasundaram, S.Vimala, "A Novel Codebook Initialization Technique for Generalized Lloyd Algorithm using Cluster Density", International Journal on Computer Science and Engineering, Vol. 2, No. 5, pp. 1807-1809, 2010.
8. K.Somasundaram, S.Vimala, "Codebook Generation for Vector Quantization with Edge Features", CiiT International Journal of Digital Image Processing, Vol. 2, No.7, pp. 194-198, 2010.
9. Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino State-of-the-art in Privacy Preserving Data Mining in SIGMOD Record, Vol. 33, No. 1, March 2004.
10. Maloji Suman,Habibulla Khan,M. Madhavi Latha,D. Aruna Kumari "Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions " Springer -Advances in Intelligent and Soft Computing (AISC) Volume 132, 2012, pp 379-386
11. Binit kumar Sinha "Privacy preserving clustering in data mining".

