An Approach for Visualization of Big Data Using Nanocubes

Asthagupta1, Sachin Majithia2, Dr. Shashi Bhushan3

123Department of IT, Chandigarh Engineering College, Landran Mohali

1astha.gupta26d@gmail.com, 2sachinmajithia@gmail.com, 3 hodit@cecmohali.org

Abstract: The concept of big data begins with the enormous amount of the data after the interpretation of the data warehousing. The practices performed in big data are generally related to issues in the large volumes of data or the unstructured data. The Visualization and structuring of data are the essential techniques required over the data available in the various domains in large amount. There are various techniques which can portray, coordinate, observe and browse the available data. Specially, the visualization of data techniques is associated with the understanding of the available data to plan the structuring of data as well as the analysis of the given databases. In this paper, visualization technique based upon the nanocubes concept is presented, which divide the data into smaller chunks, which make the data more understandable. In this paper, an improved system for the crime data visualization is depicted which consisted of the various layers altogether in order to process the input data and to produce the results in the form of the patterns, color maps, statistics and other such parametric values or the visualization techniques.

Keywords- Big Data, Visualization, Nanocubes.

I. INTRODUCTION

Big Data can be defined as the huge amount of data that might be structured or unstructured. The data is collected from various sources such as traditional enterprise, social or machine generated data which is basically very large and is difficult to process. The big data is the type of data which has diversity, complexity and needs new techniques, algorithms to keep an eye on it and able to extract latent knowledge and value from it. There are various types of data that a big data contains i.e., the data can be relational, textual, structured, unstructured, pictorial data, streaming data and etc. This data can be used for Statistical modeling and Aggregation. It needs advance methods for its processing[10]

Datasets are getting larger, and visualization of data is becoming difficult. Imagine a data set which has billion of entries. We can summaries into small module and then visualize it but that is a difficult task and solution for visualization for huge amount of data cannot be solved. Giving a solution to this problem we build nanocubes or data cubes, which have same meaning. They are the structures that perform various aggregations to every possible module in the database and then represent it. Data cubes are made small enough so that they fit in the main memory and are named as nanocube.[11]

FIGURE 1. BIG DATA.

A. Data Visualization

In simple terms Data Visualization can be stated as an attempt to make people grasp or learn any specific data by putting it in visually. In other words it is a type of communication which is done visually. It deals with formulation and survey of the data which can be or needs to be visually expressed. Visualization of Data or Data Visualization is the representation of the data in pictorial or graphical form. The main objective of Data Visualization is to relate data smoothly and precisely to the user according to his needs.[12] The data can be in any format i.e., numerical, information based, descriptive data and so on and fourth. The complicated data can be approached easily and is more understandable. Data can be created from various numbers of sources like satellites, sensors and internet activity. This data is referred as ‘Big Data’. Hence, Visualization of this data is a difficult task and is the major challenge in the field of Data Visualization.[7]

Visualization of data is a technique for the encryption of the data in optical format. It also follows one of the steps of data analysis. Visualization of Data is the pictorial demonstration of complex information for two objectives: communication and the data analysis. There are many stories contained in the data and to observe and recognize these stories data visualization has a compelling importance to present it to others. There might be thing in the data that are not physical i.e., it is abstract like the information related to the stats.[9] There are various characteristics of effective visualization. Some of them can be stated as below:

- It presents the data.
- It makes the user think about the object not about the technique through which it has been designed.
- It doesn’t alter the data.
- Representation of large data set is orderly.
- Pleasant for the user to see.

B. the Three V’s of Big Data

i) Volume: In simple words it can be referred as the huge amount of data which is generated by any source like
internet, bank, medical, and so on. Over internet, we create and share massive amount of data every day. Social networking site such as Facebook have billions or trillions of like, comment and picture upload each and every day. This data which is increasing rapidly needs to be stored and analyzed with new emerging technology. Now, with the help of Distributed systems we can reserve this high amount of data on different locations.[7,15]

ii) **Velocity**: As defined in Science, it has the same meaning. It is simply related to the speed at which the data is generated i.e., the rate of data generation. For example, speed at which any transaction is performed whether in any bank, any online dealing (purchase of goods), and instant messaging and so on. Big data on the other hand allow us to analyze all data at time of generation and by not affecting the data base.[16]

iii) **Variety**: It refers to the different types of data that is generated. The data can be structured, unstructured and semi-structured or in other words, the data can be any form like pictures, graphs, textual, voice recording, video recordings and so on. Hence, Big Data is the collection of this all such type of data and have a feature to make that data structured and making it easily understandable.[17]

![Three V’s of Big Data](image)

**Figure 2: Three V’s of Big Data**

**C. Big Data Clustering**

The data clustering plays the vital role in the data segmentation, which is achieved in the layered hierarchy. The big data clustering requires the multi-layered analysis because of the data density and the variation in the obtained data entries. Cluster analysis is a method to detect the number of clusters from a given data set. The objects in the cluster are arranged in such a way that they are similar to the objects within the same clusters and are different from the objects that are lying outside the same cluster. Cluster analysis can be used to detect the different patterns, classification, grouping the objects based on similarity or dissimilarity criteria, searching the objects and so on.[14]

**D. Data Classification**

Big data analytics based classification models can be divided into three main classification levels stated as: document level, segment level, and entity level big data analysis. The main goal of Document level whose goal is to classify the trend in the document present in the given data. Segment-level big data analysis expects to group assumption communicated in each one entity and cluster to inquiry linguistic uses. On the other hand, there is no basic contrast in the middle of report and segment level classification in light of the fact that segments are somewhat signifies the data in the multiple classes altogether.[8]

Ordering content at the archive level or at the segment level does not give the important subtle element required conclusions while substance which is required in numerous applications, to get these points of interest and exact result; we have to go to the subtle element level to figure out its aspects. Viewpoint in the given data expects to arrange the assumption concerning the specific parts of individual elements for discover feeling. The first step is to recognize the substances and their angles and their points.[13]

**E. Nanocubes**

Nanocubes are the data structures, which are invented to maximize the speed of the data processing and have been developed by the IV (information visualization) department at the AT&T research labs from the United States. The major reason behind the data extraction in the case of the nanocubes is for the easy visualization of the data. The nanocube visualization in the proposed model is based upon the distributed visualization methodology, which facilitates the computers to process the billions of the records and to visualize them step by step as they underwent the completion of the processing. It explores the patterns from the input data and the visualization results are shown over the web browser using the little memory based modern level processing.[1,2]

**II. BRIEF LITERATURE REVIEW**

As discussed in section 1 there are various techniques which are used for the visualization of the big data. The work done related to our research for the visualization of the big data in the source code is described below:

Deepa Gupta, Sameera Siddiqui (Big Data implementation and Visualization) et al they said that large organizations and government agencies launched research programs to keep view of the challenges presented in big data. Visualization is used as an effective tool for the representation of data. The graphics community researchers discovered big data tool for effective visualization. The data collected from web or mobile devices are of great size and to provide them to be understood. In other words, extraction of knowledge or making better decisions from that data by using various scientific processes is a difficult task. Visualization is a tool which helps to see glimpse of the big data. An information structure named nanocube is discussed and Hadoop for back processing and data discovery tools is used. Visualization helps to explore business data more easily and make it fully understandable.[1]

Seong-hun Park, Young-guk Ha (Visualization of Resource Description Framework Ontology Using Hadoop) et al proposed that existing models used for the big data visualization creates problem ad is a difficult task. He tried to do this while using Hadoop. The system is divided into three parts i.e. the visualization server, data server and the visualization devices. The visualization as
well as the data server uses the Hadoop and user used the web browser. The user in this is provided with the result for the visualization on the web browser.[2]

Sun-Yuan Kung (Visualization of Big Data) et al they proposed about the divergent types of sources from which big data can be collected. They can include data from physical to social and cyber types, which is messy, incomplete and imprecise. Due to its volume and velocity (quantitative) and variety (qualitative) challenges, big data is also considered or related to something like, “the elephant to the blind men”. It has been an important paradigm for the data mining and learning tools. In continuation with it also state about the V paradigm which is for “Visualization”. He discussed that the visualization tools are said as the supplement to the domain expertise and provides a big picture for the user to help the users to formulate critical questions, insightful answers and postulate heuristics.[3]

Daniel Keim, Huamin Qu (Big-Data Visualization) et al discussed the data driven era, where the data is used for a variety of purpose. The power to make decisions timely on the data available for business success, clinical treatments, national security and disaster management. The data is generated from the large scale simulations, observations, experiments or sensors and so on, they help to lead to new discoveries by using best possible tools and extract knowledge from them. Though, most data becomes either too small or too big in its life span. In fact, all the fields of study and practice have problem in big data. Large corporate sector or the government agencies are facing various types of challenges in addressing the big data. Visualization on the other hand, is not essential enough for presenting the vast information and the complex analysis. It discovers new research opportunities for the visualization community. The main highlights shown in this paper is about problems in the big data i.e., visually and also new applications, systems and technology.[4]

Ekaterina Olshannikova, Aleksandr Ometov (Visualizing Big Data with augmented and virtual reality: challenges and research agenda) et al they proposed about an overview of multi-disciplinary in the field of big data and its visualization techniques and tools. The research issues and achievements are discussed. In the existing big data, the main focus is to summarize the challenges that are faced in the visualization methods. Although, the current state of the big data visualization a novel solution for the issues is presented. In the paper, proper classification of all the existing data types, their analytical methods, tools and the visualization techniques which particularly emphasis on the visualization methodology in the past years. Based on these results, the disadvantages in the existing visualization methods are revealed. Even though, the modern world’s technical development, judgment by humans and their involvements or interactions, logical thinking is kept in mind while working with the big data concept. Hence, the involvement of humans is made less on this large amount of information. Therefore, an approach i.e., non-traditional is proposed where the discussion is all about the capabilities of virtual and augmented reality which can be applied in the big data visualization field. Applications in the visualization of big data is discussed which is utility of the mixed reality technology integration. Hence, without the data losses due to human interruptions or issues in a short period of time, the central area of mixed reality is also discussed. The main classification and challenges that might occur in the integration of the two technology i.e., augmented reality helmets and virtual reality displays, impacts are also discussed.[5]

Charles (Chuck) Hansen (Big Data: A Scientific Visualization Perspective) et al proposed that there are modern computers which have high speed and performance who measure data sizes in terabytes and petabytes. These machines offer huge potential for the solving of large scale realistic problems, the effectiveness mainly hinge upon the human ability to interact with that simulation results and extract useful information. In 21st century, the greatest challenge would be of making use of this vast amount of information produced and to effectively understand it. Visual data analysis will be the most important one for the understanding of such large scale information.[6]

III. RESEARCH GAP

In the literature survey, it is acknowledged that the psuedocode to build nanocube can be more optimize. Nanocubes are described as the unique data cubes which do not require large heaps of data for the visualization. Efficient storage, querying of large datasets is offered in the nanocube but without limits i.e., the variety of real world data can be represented effectively by using nanocubes. Web Browser like Firefox, Chrome can be used to implement these limitless data by using some metric feature. Therefore, an approach can be built to visualize the in small screen. In other words, the huge amount of data can be represented on the small screen by using some metrics; the nanocubes can be used for the better visualization of the datasets.

IV. IMPLEMENTATION

A. Implementation Procedure

In this paper, we proposed an algorithm which evaluates the similarity and the relationship within the data entities for the discovery of the perfect trends and patterns within the given data for further data analysis. The big data analysis features like positive, negative, automatic classification and product feature classification. It basically improves the performance for the pattern classification. The proposed algorithm consisted of four basic components that are, Post/Thread Acquisition, Tokenization, Similarity Evaluation and Segmentation based Analysis.

i) Post/Thread Acquisition- It is the first step which is basically used to read the data saved offline. Data acquisition basically reads the data in the true or raw form which needs to be processed further.

ii) Tokenization- In this step, based upon supervised and unsupervised rule, the programs reads the required data entities from the given data. This bag of features (also called bag of words) is obtained from the given data by using the tokenization or trend analysis over the given data. The data, then, loaded to the memory and passed to tokenization process for further computation process. This process of tokenization extracts all of the terms or words from the input data and filters them on the basis of a list of valuable bag of features.
iii) Similarity Evaluation- The input data is analyzed for the inter entity based evaluation for all individual entities in the given data. The similarity based evaluation is utilized to group the data altogether in the one segment. The file contains the ranking for each of the word listed on the list. The rank or weight or strength of the words has been listed in the document within the variable ranges. The bag of words is classified on the basis of their use and its impact in the input data.

iv) Segmentation based Analysis- It will extract the segments according to the various trends and its micro-components through the following steps:

- A document or an individual entity is broken in its basic parts of segments, called clusters, which identify the structural elements of a document, segment or individual entities.
- Segment-bearing patterns and trends in the input data, which are identified through the use of specifically designed algorithms.
- Each segment-bearing the trends and patterns in a document are computed on the basis of the similarity score and similarity data based on a logarithmic scale that ranges within the variable scale.
- Finally, the scores are combined to determine the overall segment analysis of the document and segments. Document scores variably ranges which contain useful and valuable trends for the data segmentation.

Therefore, in the proposed model’s it will extract proper nouns from the data crime type, crime region and other such percentages. Each extracted named entity is classified, tagged and assigned an entity score, which gives meaning and context to each entity that which belong to which category phase according to scores. With named entity extraction you have the access to valuable insight.

Algorithm: Brief Design of Big data analysis algorithm

1. Obtain the data from the input big data thread Tr
2. Extract the list of primary terms from the input data
3. Extract the N number of columns from the input data
4. Load the supervised trend and feature classification (TFC) knowledge data
5. Classify the data in the various trends based upon the similarity evaluation based upon the TFC data
6. Acquisition of the pattern classification (PAC) knowledge data
7. Discover the patterns within the data based upon the PAC data
8. Classify the data entities in the various three primary types based upon the intensity evaluation
   i. If similarity score is too high: Assign the data in the high intensity index based cluster.
   ii. If similarity score is calculated on the moderate level: Assign the data in the moderate intensity index based cluster
   iii. If similarity score is calculated on the lower level: Assign the data in the lower intensity index based cluster
9. Collaborate the priority based indexing and clustering
10. Return the final processed and segmented data

B. Performance Metrics

Total Count Results: This metric will basically count the total number of data input for the evaluation. For example, in this, we have taken a total of 50000 entries. They are stored in the training database and contain the record of the various types of the crimes in the input file. These 50000 entries have been processed under the data evaluation model and utilize the 10000 entries in every round. Basically, it will tell the total number of data set taken and processed it according to the user, i.e., how much data he or she have to process in one go and of which one nanocube is made.

Results of Category Breakup: The category breakup shows the result from all the dimensions in the nanocube. Here, there are four dimensions upon which the results are categorized; they are time, count, crime and location. They all shows the result in simultaneously i.e., the time dimension will show at time period, crime will show the type of crime, the count will show the number of crime and location will show at which location the crime took place. In other words, it will categorize the result based upon the dimensions.

![Figure 3: Result of category breakdown](image)

**Breakdown the count per crime type:** In the original test database file, a number of patterns have been listed, which must be discovered in order to find the useful information. It will basically show the results according to the whole data base. The crime dimension is focused and it reports the counts for each of the possible values. In other words, it will take the result of category breakdown and according to its schema the results are shown.

**Visualization of the Nanocubes:** The crime data visualization is an art to show the intensity of the crimes within the given area or city. The proposed model has been designed for the identical visualization for the variety of the crimes.
V. RESULT AND DISCUSSION

The proposed model is based upon the big data clustering and classification for the analysis of the crime data. It has been applied over the crime data obtained from the Chicago city police department, and it acts the standard dataset for the crime data visualization and processing for the crime intensity evaluation. The proposed model has been designed to split the big data in the smaller blocks known as the nanocubes, which are processed individually under the computing systems for the information discovery in the input crime data. The system improves the crime data visualization which consisted of the various layers altogether in order to process the input data and to produce the results in the form of the patterns, color maps, statistics and other such parametric values or the visualization techniques. It evaluates the variety of the factors for the purpose of the data pattern discovery and then the data is visualized with the higher accuracy.

VI. CONCLUSION

In this paper, the analysis of the crime data is done based upon big data clustering and classification. The big data is divided into smaller chunks known as nanocubes and then based upon various factors the visualization is done. The input data is processed in the nanocubes format. Nanocubes are described as a data structure for data cubes which are in-memory. Nanocubes are so formed in the way that the datasets are explored interactively on the web browser with various factors. In some areas, where huge amount of data cannot be easily interpreted; the nanocubes are used to represent the data which takes slight memory. Therefore, with the help of Nanocubes the big data can be visualized onto the small screen in more interactive and efficient manner.

ACKNOWLEDGMENT

This research paper is made possible with the help and support of my parents, teachers, family, friends, and all the people who guided me throughout my work. Especially, I am thankful and I express my gratitude to the following people who contributed and helped to make this work possible. I would like to thanks Er. Sachin Majithia and Dr. Shashi Bhushan to support and for encouraging that motivated me to write this paper. They kindly read my paper and suggested me advices on grammar, matter, and the title of the paper. I sincerely thank my parents, family, and friends, who gave me emotional and financial support. Without the support of these kind people the product of this research paper would not be possible.