IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

An Automatic Big Data Visualization Framework to Plot Chart Using T-SNE Algorithm

R.Banupriya[1*], R.S.Karthik [2]

[1,2]Computer Science, CMS College of Science and Commerce, Bharathiar University, Coimbatore, India

[*]*Corresponding Author: priyant2@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 24/Nov/2018, Published: 30/Nov/2018

***Abstract*—** Data visualization is used to transforms data to image. The data are changed into image format for more understandability of data. Giving a huge dataset for the essential task of visualization is to visualize the data to tell compelling stories by selecting, filtering and transforming the data. It also used to pick the right visualization type such as bar charts or line charts. The ultimate task is to provide more visually effective data representation. A revolutionized system in the field faces the following three main challenges and they are Visualization verification, which it used to determine whether the visualization for a given dataset is interesting, from the viewpoint of human understanding. Visualization search space checks whether the resultant visualization is a "boring" dataset, then it may become interesting after an arbitrary combination of operations such as selections, joins, and aggregations, among others or not. On-time responses is does not deplete the user's patience. The proposed system solves the above challenge by implementing multidimensional scaling or the popular t-SNE algorithm. The t-SNE algorithm is based on non-convex optimization, has become the standard for visualization in a wide range of applications. This work gives a formal framework for the problem of data visualization – finding a 2- dimensional embedding of cluster data that correctly separates individual clusters to make them visually identifiable. Ground-truth cluster is checks the conditions assumed in earlier analyses of clustering while underlying the data. To achieve the goal of data visualization existing system used LambdaMART algorithm to learn rank technique. In proposed system the t-SNE algorithm takes place the role to create more effective visualization with the help of clustering method to group or ungrouped huge data.

Keywords- Machine learning, Computer Science, Artificial, Architecture and Syatems.

I. INTRODUCTION

Data mining is an analytical process designed to explore large amount of data called "big data" to form new subsets of data. It is the process of looking at large sets business or marketing related information in a different way to derive new information from what already exists. In the business environment data mining has proposed with various frameworks to serve blueprints to organize the process of gathering data, analyzing data, disseminating result, implementing results and monitoring improvements. Some complex projects with complex data may require the coordinate efforts of data mining are Import and Export business, Stock market, etc.,

Data mining also define as Searches and produces a suitable result from large data chunks. Data mining comes under the data science. Data science is the field of computer science having statistics, computing, mathematics and several technical processes including different methodologies. Data extraction, data management, data transformations, data pre-processing are some of the different processes involves in data mining. The relationship between different data set and variable are unhidden in data mining.

The ultimate goal of data mining is to predict data. The process of data mining consists of three stages: i) initial exploration, ii) Validate/ verify data through pattern identification, iii) deployment or implement.
Few disciplining where the data mining technique or technology involves are

- Databases and database technology
- Statistics
- Computer science
- Artificial intelligence
- Machine learning

II. ARCHITECTURE OF DATA MINING

Data mining is the process of discovering hidden valuable knowledge by analyzing an extremely large amount of data

which also store in different database. It is reliable. Data preparation and data mining is the major two data mining processing stages followed while perform process over data mining.

Data preparation is the first stage which used to perform while collecting data. It check whether the collection of information is fully filled or not, is the data stored in different database, text files, document are properly integrated or not and Finally check whether the relevant is retrieved from the database.
Different types of data used by data mining to gather data are

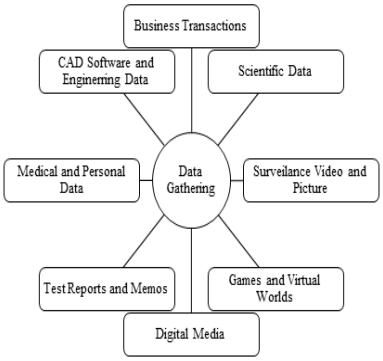


Figure 1: Type of Data Gathered in Data Mining

Some of the updated technologies where data mining explores are artificial neural network is a non-linear data mining predictive model. Decision tree e is a tree shaped structure which used to represent the set of decision to take. The rules generated by tree structure are used to classify the dataset. Genetic algorithm is the design based on the concepts of evolution. It is the optimization techniques used to combine the design.

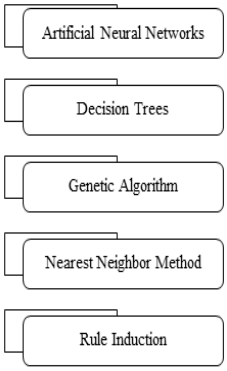


Figure 2: Data Mining Techniques

Nearest Neighbor method is a technique that classifies each record in a database based on the combination of the record class. It also called as k-nearest neighbor technique. Rule induction is the rule decision loop rules from the data based on statistical significance.

The process steps followed to data preparation and data mining are as follows

A. DATA CLEANING

The process to remove the noise data and irrelevant data in data gathered. It used to filling the missing values, combined compute. The output of the cleaning process is adequately cleaned data.

B. DATA INTEGRATION

When data lies in different formats in different location can combine to same place using integration process. The data may store in a database, text files, spreadsheets, documents, data cubes, and so on. The integration method use metadata to reduce errors. It gives more reliable data.

C. DATA SELECTION

The search related data are retrieved from the data collection. It used large volume of historical data for analysis. It integrate data contains much more data than actually required.

D. DATA TRANSFORMATION

The data consolidation method used to selected data transformed into forms.

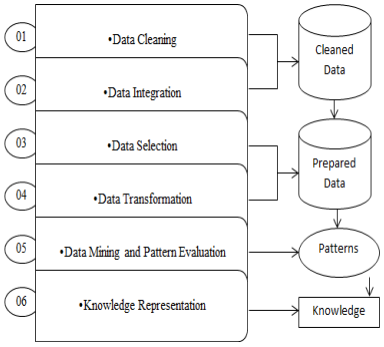


Figure 3: Data Preparation and Data Mining Process

Graphical user interface is the interface used to communicate between the user and the data mining system. It helps to use the system easily and effectively. Due to this process the user does not know the real complexity of the process. It specifies the user request in query form for easily understandable manner.

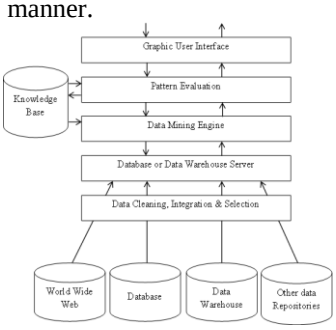


Figure 4: Data mining architecture

Knowledge base is the beneficial process of data mining architecture. It used to guide the search for the result patterns. It contains user beliefs and data from user experiences. It receives input from the knowledge and make the more accurate and reliable result.

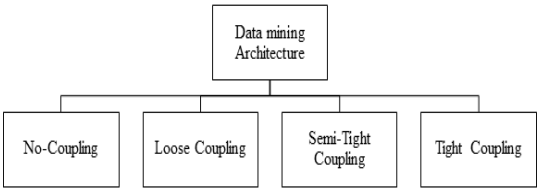


Figure 5: Types of Data Mining Architecture

The major role of data visualization is to effectively visualize the big dataset data using machine learning algorithm to automate the manual process. Existing Deepeye data visualization framework three major challenges like visualization verification, visualization search space and on-time response are concluded with the LamdaMART algorithm. Data Visualization and verification method are in place to identify the interesting visualization type from user point of view.

Visualization search space split the huge data set based on its column to make optimized searching, selecting and filtering specific data by simple click as like query request and response method. It allow user to use aggregation function to group data based on data set column. It is fast processing and more user patient. It provides on-time response. Even though the framework have all few feature it only possible to give fast result only for sample data not for huge data base.

If huge data applied for existing method it takes more time verify the data to move the process from stage one to next. The automatic chart selection method from user point of view not all time gives full accurate result. Sometime it differs by few floating point difference. And it is also only possible to create few types of chart not all types is it another major drawback. Only the basic chart types like Bar chart, Pie chart and Line chart are possible with existing method. With proposed system we use t-SNE non convex optimization algorithm to visualize data within the range. Proposed technique also represents few more scatter model visualization charts. It also provides very huge dataset with in possible short timing. It updates data dynamically to dataset.

III. PROBLEM STATEMENT

Existing approach developed new framework for huge data visualization method. It selects chart type automatically to human vision able form. It gives accurate result only for sample data not more accurate when large data set used. The algorithm used by existing system is not effective one for large dataset. It does not consider cluster point when selecting chart type. It also takes too long time to process when dataset size increased. It automatically understands the human approach and efficiency of human perception. It needs to pass number of check to verify data visualization.

It used top-k model to rank the visualization. To implement the machine Learning technique need wide information. It supervised visualization verification using learning function. It gives precision with an average prediction to precision. The pranking problem used normalization discounting gain method to evaluate our method.

The visualization search space perform simple operated related grouping to create chart. It uses minimal aggregate functions like Sort, Bin, Group-by and Aggregate. It uses x-scale to sort the data through framework. Only minimal functions are possible. More than the listed are not possible with existing framework. All possible form of chart type is segregated first then finalizes the suitable one. Huge dataset takes to high time to complete the verification. Data binning used to partition the value into different buckets.

It works like database tools which retrieve output from dataset or back-end for input given as query in front-end. It is time consuming when entire big data is used for data visualization. It supports only for bar charts and histogram. It used novel data visualization system.

IV. OBJECTIVE

The main intent of this thesis is to find out the clustered data in dataset before visualize data as chart. Existing system focus with either bar chart or histogram chart, we also in need to create scatter, line and so on type of charts to represent data in visual form. The algorithm should be more effective when huge big dataset used. The main objective of proposed system is to reduce the time taken by algorithm to define ranking. Another objective solve is to grouping data in dataset with specific function.

The computerized proposed system automates user work simple by finding and verifying the dataset required models. The proposed system was suitable for both the huge dataset and the simple dataset. The ranking used in existing system supervised learning rank using input space as x axis co-ordinate and output space as y axis co-ordinate. Existing system goal is to perform data search more effective manner. F(x1) and F(x2) are two function in example for algorithm. It feels F(x2) is better than F(x2).

It follows normalization method to group data, which is fast to sample or simple data table but when same method used it takes too large dataset is glow.

V. EXISTING SYSTEM

Using a huge dataset, the essential task to visualize the data to tell compelling stories by selecting, filtering, and transforming the data, and picking the right visualization type such as bar charts or line charts. The ultimate goal of an existing system is to automate the task that currently requires. It needs heavy user intervention in the existing visualization systems. The three main challenges like Visualization verification to determine whether visualization for a given dataset is interesting, from the viewpoint of human understanding or not with some check list, Visualization search space, which helps to satisfy dataset viewer a "boring" dataset may change to interesting. It allow user to perform an arbitrary combination of operations such as selections, joins, and aggregations, among others which effects at chart on viewer point, On-time responses need to user's patience.

A. *DISADVANTAGE OF EXISTING SYSTEM*

Disadvantage in existing system to overcome with proposed system are

i. Centralized visualization method
ii. High cost maintenance
iii. Only sample dataset data get maximum accuracy
iv. Limited chart types only possible.
v. Ranking method calculation time
vi. Take too long time take to find maximum match of user viewpoint.
vii. Search space satisfies with minimum aggregate function.
viii. Time taken to generate chart is too high when little big dataset used, So not user patient
ix. Existing algorithm doesn't consider about neighbor more clustered coordinates while plot.

VI. PROPOSED SYSTEM

If huge data applied for existing method it takes more time verify the data to move the process from stage one to next. The automatic chart selection method from user point of view not all time gives full accurate result. Sometime it differs by few floating point difference. And it is also only possible to create few types of chart not all types is it another major drawback. Only the basic chart types like Bar chart, Pie chart and Line chart are possible with existing method. With proposed system we use t-SNE non convex optimization algorithm to visualize data within the range. Proposed technique also represents few more scatter model visualization charts. It also provides very huge dataset with in possible short timing. It updates data dynamically to dataset.

A. *ADVANTAGE OF PROPOSED SYSTEM*

i. Affordable cost maintenance
ii. Possible to use big data or huge dataset
iii. More chart type available then existing approach.
iv. It is non-convex optimization, which become the de-facto standard for visualization in a wide range application.
v. To increase user view point visualization, 2D embedded of cluster able data that are separated individually.
vi. Search space satisfies with maximum number of aggregate function.
vii. Time consumption

Modeling standards and its references

Model	Reference
IEC 61508-3	Language subset
IEC 62304	Software Unit acceptance criteria
ISO 26262-6	Use of language subsets
DO-331	High-level requirements are accurate and consistent High-level requirements conform to standards Low-level requirements are accurate and consistent Low-level requirements conform to standards

VII. DESIGN AND IMPLEMENTATION

The proposed algorithm t-SNE (t- Distributed Stochastic Neighbor Embedding) is a non-linear dimensional algorithm. It used for exploring high-dimensional data. It works with highly dimensional data. Dimensionality reduction is more important to high dataset when we can plot the data using scatter plot, histogram and boxplot. The pattern sensed using descriptive statistics.

Algorithmic details of t-SNE are an improvement on the Stochastic Neighbor Embedding (SNE) algorithm. Steps to follow are as below.

STEP 1

It starts by converting the high-dimensional Euclidean distance between data points into conditional probabilities.

datapoints: xi, xj,

Conditional probability: pj|i

Here xi would pick xj as its neighbors were picked in proportion to their probability density under a Gaussian centered at xi.

. The nearby datapoints, pj|i is relatively high and widely separated datapoints. It is almost infinitesimal. pj|i is given as follow

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)},$$

Where $^{\sigma}$i is the variance that is centered on datapoint xi.

STEP 2

Low-dimensional counterpart: yi

High-dimensional counterpart: yj

Datapoints: xi, xj,

Conditional probability: qj|i

qj|i is given as follow

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

Set pi|i and pj|j to zero to find model pair wise similarity.

Step1 and step 2 are used to calculate the conditional probability of similarity between pair of points in High Dimensional Space and Low Dimensional Space.

Logically the conditional probabilities pj|i and qj|i must be equal for a perfect representation of the similar datapoints in the different dimensional spaces. It finds the different between pj|i and qj|i to plot in high and low dimensions.

$$Perp(P_i) = 2^{H(P_i)},$$

H(Pi) is the states Shannon entropy of Pi. It measured in bits

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}.$$

It is smooth measure of the effective number of neighbors.

The gradient decent is used to minimization the cost function. It is the resultant force created by a set of springs between map point yi and yj. It used to find the distance between the two map data points similar between two high-dimensional points. The force exerted by the spring between yi and yj is proportional to its length and stiffness.

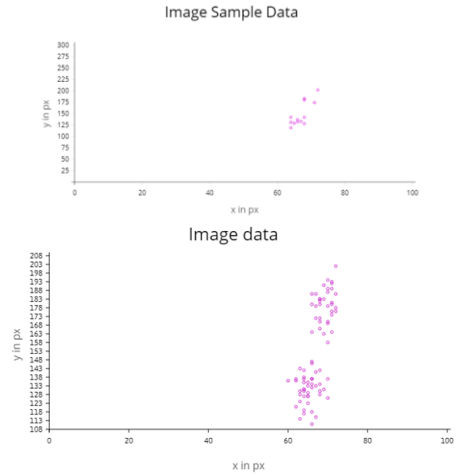
VIII. EXPERIMENTAL RESULTS AND DISCUSSION

Proposed system first examined with image pixel data. Using proposed software tool first we extract pixel value from selected image. First the proposed system applied with image sample dataset. The chart type concluded as scatter point because both the x and y axis data are values. Before plot the value the proposed system check the cluster point.

It also verifies whether the dataset contains the grouped data are not. If not it simple go with the scatter or it moved with scatter with group plot. The sample dataset and its visual representation are figured below.

Sample dataset from huge dataset

height	64	67	64	68	64	68	68	66	68	66	71	72	65
width	131	133	119	142	142	180	183	132	128	137	174	202	129



Above mentioned done for huge dataset with number of pixel data.

72	63	66	70	71	68	63	65	67	66	68	71	70	60	64	64	66	64	72	65	67	72	64	68	66
186	124	134	170	180	130	130	127	141	111	134	189	137	136	130	137	186	127	176	127	115	178	131	183	164

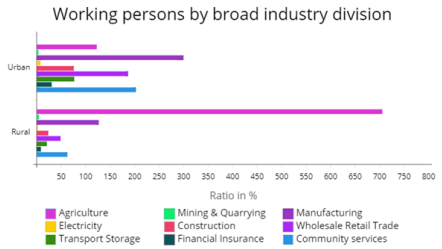
67	70	66	64	70	71	66	71	66	66	63	71	69	70	70	67	65	68	62	70	67	68	62	64	66
186	126	137	138	187	193	137	192	118	180	128	164	183	169	194	172	135	182	121	158	179	170	136	135	147

71	69	64	67	64	68	64	68	68	66	68	66	71	72	65	71	69	69	70	68	65	64	62	66	65	70
176	163	131	133	119	142	142	180	183	132	128	137	174	202	129	181	191	131	179	172	133	117	137	146	123	189

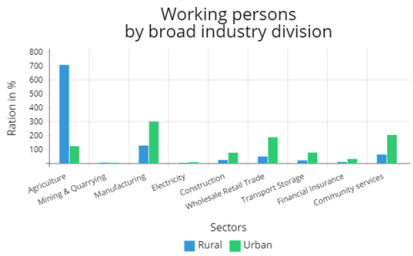
Experienced the grouped dataset visualization for the proposed algorithm the following dataset related to distribution of usually working persons by broad industry division is taken.

Sectors	Rural	Urban
Agriculture	705	122
Mining & Quarrying	4	3
Manufacturing	126	299
Electricity	2	7
Construction	23	75
Wholesale Retail Trade	48	186
Transport Storage	20	76
Financial Insurance	8	30
Community services	62	202

Working persons by broad industry division from rural and urban area – before 2000



Distribution of usually working persons by broad industry division – Chart 1



Distribution of usually working persons by broad industry division – Chart 2

IX. CONCLUSION

The conclusion identifies that the proposed system is more beneficial than existing system by its calculation time, verification type, accuracy, resource needed and so on. It also gives effective visualization to any type of dataset like statistical, grouped, sample data and image representation everything. It also possible to apply any kind of chart type that is suitable for dataset of user. The resource required for proposed system already exists in system, it is not must to go with any new framework to learn and process. The proposed system optimized time taken to verify the huge dataset as like small dataset. To find clustering data gives more effective to huge data visual representation. It supports both label and unlabelled dataset. The proposed system optimizes the search and filtering of dataset based on column grouping. It is more useful for medical field and investigation filed to represent image based visualization.

X. FUTURE WORK

The proposed system algorithm implemented with new system not with existing system framework. It is more effective when we develop new framework or update the proposed algorithm with existing system algorithm to get manual chart type selection. The proposed system also supports with grouping dataset and all kind of filtering mechanism required as per viewer point. In future the new system asset with the proposed algorithm gives more effective result with all kind of chart supports for all verity of dataset. In future the implementation of new framework should have the feature to upload dataset directly without use any advanced backend software. The idea minimizes the resource requirement need to use framework at user end. It reduces cost efficiency.

REFERENCES

[1]. Ms. Lavanya Patil , Dr. Jagdeesh D. Pujari , "Data Visualization: A Handy Plug-In" , International Journal of Engineering Research in Computer Science and Engineering, vol.3,issue.5, pid.351.

[2]. Takuya Kaihatsu ; Shinya Watanabe "A proposal of a low-dimensional approach based on DIRECT method and t-SNE for single optimization problems with many variables" , in the proceedings of the 2017 International Conference on Soft Computing and Intelligent Systems.

[3]. Oluigbo Ikenna V., Nwokonkwo Obi C., Ezeh Gloria N., Ndukwe Ngoziobasi G, "Revolutionizing the Healthcare Industry in Nigeria: The Role of Internet of Things and Big Data Analytics" , International Journal of Scientific Research in Computer Sciences and Engineering , Vol.5 , Issue.6 , pp.1-12, Dec-2017 .

[4]. L.J.P. van der Maaten, G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9" (Nov):2579-2605, 2008.

[5]. Chao-Kuei Hung, "Making machine-learning tools accessible to language teachers and other non-techies: T-SNE-lab and rocanr as first examples", in the proceedings of the 2017 IEEE 8th International Conference on Awareness Science and Technology.

[6]. D.A. Keim, "Information visualization and visual data mining", IEEE Transactions on Visualization and Computer Graphics, Vol. 8, Issue. 1 , Mar 2002.

[7]. Laurens van der Maaten, "Accelerating t-SNE using Tree-Based Algorithms", Journal of Machine Learning Research, 2014 Vol.1, Issue.21.

[8]. "Semi-supervised Learning to Rank with Preference Regularization", Martin Szummer; Emine Yilmaz.

[9]. A.G.Aruna, Dr.M.Sangeetha, C.Sathya, "Impact of Deep Learning in Big Data Analytics", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, Issue.3.

[10].Subham Datta, Gautam, Tapas Saha, "Development of a Rule Based Classification System to Identify a Suitable Classifier for a Particular Dataset", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, Issue.5.

[11]. Xianjun Shen , Xianchao Zhu ,Xingpeng Jiang , Li Gao , Tingting He , Xiaohua Hu,"Visualization of non-metric relationships by adaptive learning multiple maps t-SNE regularization".

[12]. J. Yin, Z. Zheng, and L. Cao, USpan "An efficient algorithm for mining high utility sequential patterns".

[13].X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules".

Authors Profile

Ms.R.Banupriya completed M.Sc. Computer Technology in CIT Engineering College. She was working as Professor in Department of Science & Humanity in Karpagam college of Engineering, Coimbatore, India. She prepared the paper entitiled as"An Automatic Big Data Visualization Framework to Plot Chart Using T-SNE Algorthim".

Mr.R.S. Karthik MCA, M.Phil., SET., completed MCA and M.Phil, in recognized University. He cleared SET Examination. He is currently working as Associate Professor in Department of Information Technology in CMS College of Arts and Science, Coimbatore, India

