IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Big Data Visualization Techniques of Social Media: A Survey

Komal Javalkoti[1*], Vipul Joshi[2], Pooja Shah[3]

[1,3]Dept. of Computer Engineering , Shankersinh Vaghela Bapu Institute of Technology , Gujratat Technological University, Ghandhinagr, India
[2]Dept. of Information Technology, Shankersinh Vaghela Bapu Institute of Technology , Gujratat Technological University, Ghandhinagr, India

Corresponding author: komaljavalkoti12@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.591594 | Available online at: www.ijcseonline.org

Accepted: 07/Mar/2019, Published: 31/Mar/2019

***Abstract*—** Big data will be transformative in every sphere of life. But Just to Process and analyze those data is not enough, human brain tends to find pattern more efficiently when data is represented visually. Data Visualization and analytics plays important role in decision making in various sector. Many Visual analytics methods have been proposed across disciplines to understand large-scale structured and unstructured social media data. Current Big data Visualization approaches often reduce high dimension data to low dimension, and omit some data trends or relationships. In exploratory analysis of multivariate datasets, performing an analytical task is often necessary. Such tasks may include extracting characteristics subsets and comparing them.In social network, thousands of people produce data at the same time, and huge amount of data will be produce in seconds. In this paper, survey on a Real time Information Visualization and Analysis framework– RIVA[2]. RIVA to collect data from the social networks, such as Twitter, by using Spark Cloud computing platform to discover popular topics around the world.

Keywords*— *visual analytic method, unstructured data, structured data, social media data, big data visualization, Apache spark, BladeGraph.

I. INTRODUCTION

In recent times social media has become more than a place to socialize. It has become a platform for the common people to express his/her opinion and to form online Communities. The Unprecedented availability of social media data offer. Substantial opportunities for data owners, system operator, solution providers and end users to explore and understand social dynamics. However the exponential growth in the volume, velocity, and variability of social media data prevents people from fully utilizing such data.

Exploratory analysis of multivariate dataset is often performed by focusing on a part of the dataset or certain attributes. Many Visualization methods exist for expressing the multivariate aspect of multivariate datasets. The purpose of this study is to assist the search and comparison of characteristics subsets when analyzing multivariate datasets. To assist in comparing two or more subsets, designed Blade Graph ,which is Visualization technique that easily identifies difference in data distribution. This visual representation allows analysts to notice the differences between both plural subsets and variables.

Number of users produce data in social media at same time, Vast amount of data will be generate in Seconds. In social network users will be increasing. Therefore, Use a RIVA Framework to collect the data from Social Media.

II. RELATED WORK VISUALIZATION TECHNIQUES FOR MULTIVARIATE DATASET

A. Parallel Coordinated Display

Parallel coordinate plot[3] is a multidimensional data visualization method that makes use of parallel coordinate axes. Based on the leaning and degree of lines, PCP represent both the distribution of variables and relationships between adjacent variables. An advantages of PCP is that obtaining and intuition understanding an overview of the data is easy.
This method visualizes clustered value in each variable by mean of circle representation. This allows users to compare the rough distribution of each variable. However, identifying differences in the variable is not easy regarding the circle representation exactly is difficult. An analysis tool that analysis tool that visualizes the quantitative multivariate dataset to be comparatively analyzed. Their tool visualize a data set by connecting variables by means of cluster bands based on Parallel Sets. In addition, this tool enables an analyst

to compare the differences of the bands by changing the color of each. However , dividing the dataset into subset based on records and comparing them visually is difficult.

B. Multiple Coordinated Display

GPLOM is a technique for visualizing multivariate data using a matrix. In GPLOM, some multivariate dataset can be is compared with other portion by means of representation. However, GPLOM does not have representation or function relating to the extraction and comparison of subsets[6]. Domino is an analysis tool that enables analysts to extract and combine meaningful subsets. Three block types exits as the basic visual unit in Domino, and the user can analyze the dataset by combining these blocks according to analytical purposes. Although the tool is useful for searching a subset, Domino does not assist in comparing subsets[7].

C. Technique for Comparative Analysis

A comparative analysis tool using the expression of small multiples. Their tool adopts the hierarchical structure of small-multiple displays for flexibly meaningful comparisons between graphics. Using their to, an analyst can create a subset only from qualitative information, not from quantitative information[8].

Diversity Map[9] is a technique to visualize differences between many subsets of multivariate datasets. By assigning noticeable darker colors according to the size of difference, Diversity Map support the discovery f differences in the subsets. The user of Diversity Map can identify differences, but cannot read the details from the representation.

III. BLADEGRAPH:VISUAL REPRESENTATION FOR COMPARING QUANTATIVE VARIABLES

Blade Graph as a technique for visually comparing data distribution in a single variable, that is, as an extension of the braided graph. In the braided graph, the distribution of one variable is represented by one area graph, and two or more distributions are represented by superimposing he area graphs. The braided graph has the advantages that differences in datasets can be determined from a single graph on a single axis. However, the coloring of the braided graph may be misleading. Thus, reading the distribution of each dataset and discovering differences between the distributions is difficult. Therefore, developed a blade graph to visualize distribution of one quantitative variable of dataset in order to compare them. Quantitative variables of multivariate datasets can be represented by blade graphs placed on parallel axes[10].

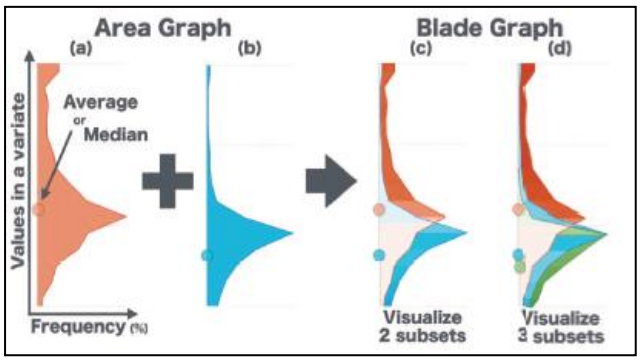

Figure 1 .Procedure for generating a blade graph [1]

IV. RIVA-REAL-TIME INFORMATION VISUALIZATION AND ANALYSIS PLATFORM

A. What is RIVA?

The number of social network users is in the scale of hundreds of millions, and social network nowadays has been one of the most prominent media for receiving and sending information. Twitter data can be generated during a short time, for an example 5 second, and then the data can be processed and analyzed in milliseconds on cloud computing platform. In RIVA, collect the real time data in Twitter through APIs, and utilize the real-time processing tools using Spark platform to receive and process the data. These data are processed according to their hash tags to gather the information about what topics they are referring to and how often these hash tags have been twitted, indicating how popular topics are. Also analyze news headlines websites and compared them with result from twitter feed analysis. In the end, RIVA the analysis results between Twitter and web news topics and visualize them through a web browser interface.

B. System Framework

The architecture of RIVA is illustrated in figure 2, divided into spark cluster, interactive data visualization and analysis tool with apache Zeppelin, and finally sentiment analysis module. On the spark cluster, spark run on Hadoop Distributed File System. There are one master node and currently implemented maximum six worker node and a native standalone scheduler in the cluster. In RIVA, driver program run on the master node to assign tasks for worker nodes, in order to receive Twitter and web news data, and start the executors to process data on each worker node return the results to the driver program and results are stored in database.

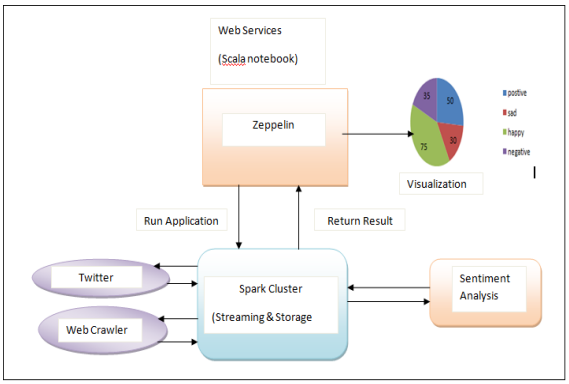

Figure 2. The architecture of RIVA

The interactive data analysis and visualization tool shown in Fig 3 are used with a browser based interface to demonstrate visualized information. Users can also add their own script programs through the web interface and submit their applications to platform. Interpreter as plugins of Apache Zeppelin can support multiple programming languages and handle tasks in the background. The sentiment analysis module analyze message from social networks with all kinds of different viewpoints and opinions, thus the world comparison technique to find the sentiment behind them providing the basis for sentimental distribution visualization[2].

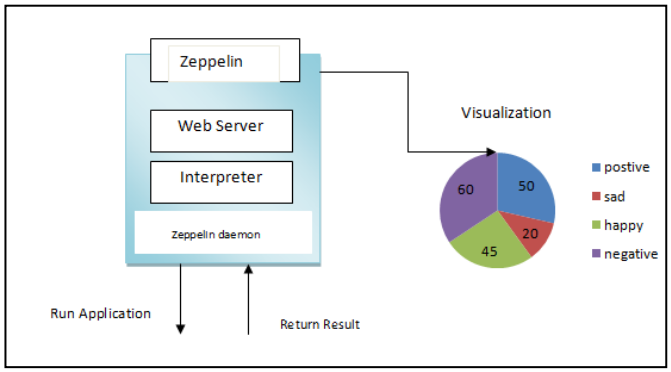

Figure 3. The interactive data analysis and visualization

C. Process Flow of RIVA

The process flow of RIVA shown in Fig 4. There are four stages: data collection, real-time processing, data storing, and visualization and analysis.

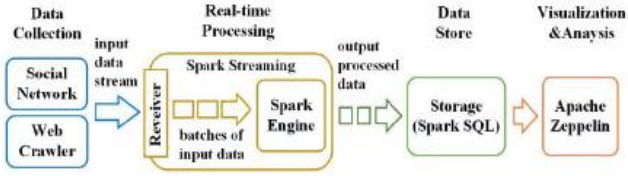

Figure 4. Process Flow of RIVA

1)Data Collection

The data collection stage uses two web services as data sources, one is twitter and other is web news. Spark cluster request data and receive the streaming data from those data sources in fixed batch period. In Twitter, tweet stream data through Twitter APIs and receive the data in real-time. In web news, a web crawler server to collect the contents from each web news.

2) Real-time Processing and data store

In real-time processing stage, Spark streaming to receive and transfer stream data into batches data. Batches data will be processed in Spark Engines to filter the hash tags in twitter and web news.

3) Data Storing

The process data will be stored in the spark template data table. The full texts string are also stored in the data storage for sentiment analysis.

4) Visualization and analysis

To visualize the data processed hash tags and word, use SQL to query the data. Then spilt out words from the full text stored in storage, and compare those word with 266 positive and 225 negative word for finding out the individual sentiment distribution in Twitter and web news. the default value of sentiment weight(SW) is zero, and when the comparison of corresponding word gets a positive result, the SW value will be added by one. If SW is greater than zero than the sentiment analysis is positive.

V. CONCLUISON

Big data visualization improve in visualize of data in social media, which data are multivariate. There are some visualization techniques for multivariate dataset. They work with multivariate of dataset within multiple sources of dataset. In this paper carried out the information about the RIVA. RIVA was a real-time information visualization and analysis platform. In RIVA ,collect the real time data from twitter and web news. After collecting the data, analyze the data using the Apache Zeppelin tool. Then store the data in database in spark template table and visualize the data.

ACKNOWLEDGMENT

The authors gratefully acknowledges the contributions Komal Javalkoti, Prof vipul joshi and Prof pooja shah for their work on the original version of this document.

REFERENCES

[1] Hiroaki Kobayashi, Hiroko Suzuki, Kazuo Misue, "A Visualization Technique t Support Searching and Comparing features of Multivariate Datasets". 201519th International Conference on Information Visualization.

[2] Yong-Ting Wu, He-YenHsieh, Xanno K. Sigalingging, Kaun-Wu Su, Jenq-Shiou Leu", RIVA: A Real-time Information Visualization and Analysis Platform for Social Media Sentiment

Trend." 2017 9th international Congress on Ultra Modem Telecommunication and Control System and Workshop.
[3] A .Inselberg " ThePlane With parallel coordinates", The Visual Computer,Vol. 1, No. 4, pp. 69-91, 1985.
[4] D.B.Carr, R.J.Littlefield. W.L. Nicholson, and J.S. Littlefield, "Scatterplot Matrix Technique for Large N", In Journal of the American statistical Association, Vol. 82, No.398,pp. 424-436, 1987.
[5] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg,"Comparative analysis of multidimensional, quantitativedata", IEEE Transactions on Visualization and Computer
Graphics, Vol. 16, No. 6, pp. 1027–1035, 2010.
[6] J.-F . Im, M. J.McGuffin, and R.Leung, "GPLOM:The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data", IEEE Transactions on Visualization and Computer Graphics, Vol. 20, No. 12 , pp. 2023-2032,2014.
[7] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit,"Domino: Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets", IEEE Transactionson Visualization and Computer Graphics, Vol. 20, No. 12,pp. 2023–2032, 2014.
[8] J. Kehrer, H. Piringer, W. Berger, and E. M. Gr¨oller, "AModel for Structure-Based Comparison of Many Categoriesin Small-Multiple Displays", IEEE Transactions on Visualizationand Computer Graphics, Vol. 19, No. 12, pp. 2287–2296,2013.
[9] T. Pham, R. Hess, C. Ju, E. Zhang, and R. Metoyer, "Visualizationof Diversity in Large Multivariate Data Sets",IEEE Transactions on Visualization and Computer Graphics,Vol. 16, No. 6, pp. 1053–1062, 2010.Letter Symbols for Quantities, ANSI Standard Y10.5-1968.
[10] W. .Javed, B. MDonnel, and N. Elmqvist,"Graphical perception of multiple time series", IEEE Transaction on Visualization and Computer Graphics, Vol. 16, No. 6, pp. 927-934, 2010.
[11] T. Pham, R. Hess, C. Ju, E. Zhang, and R. Metoyer, "Visualizationof Diversity in Large Multivariate Data Sets",IEEE Transactions on Visualization and Computer Graphics,Vol. 16, No. 6, pp. 1053–1062, 2010.Letter Symbols for Quantities, ANSI Standard Y10.5-1968.
[12] Inc. Cisco System, "Cisco Visual Networking Index: Global MobileData Traffic Forecast Update,2015-2020," February 3, 2016.

Authors Profile

Miss Komal B Javalkoti has been study in Master (software Engineerining) at Shankersinh Vaghela bapu Institute of Technology. She had Completed Diploma in Computer Engineering From Gujrat Technological Institute,Gujrat in 2014. Bachelor of Technology in Computer engineering From Gujrat Technological Institute,Gujrat in 2017.

Prof. Vipulkumar V. Joshi has been working as an assistant professor in Information Technology Department at Shankersinh Vaghela bapu Institute of Technology. He had completed his Diploma in IT from Ganpat University, Gujarat in 2004& Bachelor of Technology in IT from Ganpat University, Gujarat in 2008. After he had pursued his master of technology in ITfrom Ganpat University, Gujarat in 2013. Also He had Joined Ph.D. Program in 2017 in the area of IoT from Ganpat University. He is having total 6 years of teaching experience at a bachelor and master level and 2 Years of Industrial Experience. He is having keen interest in the field of big data, IoT, Data Mining andDTN.

Mrs. Pooja Shah has been working as an assistant professor in computer engineering department at Shankersinh Vaghela bapu Institute of Technology. She had completed .her bachelor of engineering from Gujarat University, Gujarat in 2005. After she had completed her master of technology in computer science engineering from Jodhpur National University, Rajasthan in 2012. She is having total 13 years teaching experience at a bachelor and master level along with that she had authored 2 National and 3 International Research papers as a primary and secondary author. Her research areas include networking, information security.

