IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Scope and Challenges in Data Visualisation: Presentation of Data in a Graphical Format

Asoke Nath[1*], Tejash Datta[2], Faisal Ahmed[3], Nitin Gupta[4]

[1,2,3,4]Department of Computer Science, St. Xavier's College, Kolkata, India

Corresponding Author: asokejoy1@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.15961601 | Available online at: www.ijcseonline.org

Accepted: 25/May/2019, Published: 31/May/2019

Abstract— Data visualization calls to mind the old saying: "a picture is worth a thousand words." Data visualization techniques exploit this fact: they are all about turning data into visual form by presenting data in pictorial or graphical format. This makes it easy for decision-makers to comprehend the information contained within vast amounts of data at a glance to understand and draw inferences from it. In the present paper the authors have presented techniques and approaches for how huge quantities of data can be represented visually. The authors have designed an intuitive interface to make it easier for an end user to plot data and interact with it. It also demonstrated how data presented visually can be used to draw meaningful inferences from datasets representing real-world scenarios.

Keywords— Big data, Data visualisation, Interactive data representation, Data analytics

I. INTRODUCTION

In today's modern technological landscape, data is both more abundant and more important than ever. "Big Data" is a catch-all term used to describe all types of large sets of data, be it structured or unstructured. Even though collecting such large volumes of data is easily accomplished, drawing meaningful inferences from it is a non-trivial task. The information garnered from its correct analysis can provide valuable insights that can be applied in the real world for tangible benefits.

The benefits and insights from big data are not limited to any particular industry or location. Data analysis can be applied in virtually any sector since there are data metrics that can be collected and examined in almost every field.

For example, consider the currently most widely used application of big data: user targeting for advertisements. In this sector, examining the correlation between the type of content a person consumes on the internet and the type of products and services that they shop for online has enabled advertisers to create campaigns that target very particular demographics. This results in higher user engagement with ads which, in turn, produces higher sales.
Next, consider the capricious world of international finance. With how mercurial the stock market can be, even professional fund managers often fail in choosing profitable securities. Indeed, it is quite an impossible task to predict winning stocks. However, analysis of data can reveal minute patterns that can be taken advantage of to obtain even the slightest of a competitive edge. Most large banking firms employ professionals known as "quants" whose sole responsibility is to examine data to find such patterns and opportunities.

Applications of big data are not limited to corporate endeavours; it has been used to improve lives in the medical field as well. Now that most hospital records are digitalized, it is easy to spot patterns and trends in the medical records of patients to discover those who may face a potential risk and initiate preventative care. Big data applications are also well suited for medical research endeavours such as those involving voluminous DNA data.

Given the importance of big data in today's modern technological landscape, it is of paramount importance to apply the best techniques to analyse it and derive meaningful insights from it. Visualising data is one such approach that synergistically combines machine and human intelligence. Data visualisation transforms raw textual information into an intuitive and interactive visual representation. This technique leverages the computing power of modern computers and the intuitive intelligence of human users.

The goal of this research work is to examine and demonstrate how data visualisation techniques can be used to aid an end user in analysing and making discoveries from a huge dataset. Towards this end, an easy to understand and operate system is built that can create detailed and varied visual

representations from datasets. Finally, the efficacy of data visualisation in solving real-world problems and drawing meaningful inferences is demonstrated through its use on real-world data that pertains to various public sectors of India.

The paper is organized as follows: Section I contains the introduction to data visualisation and the needs and uses of it, Section II covers a literature review of related work in the field so far, Section III elucidates the methodology of the project and the approach taken to create a system for visualising data, Section IV describes the results and discussion on the basis of several case studies involving real-world data, Section V presents the conclusion of the research work and mentions its limitations and future prospects.

II. LITERATURE REVIEW

In recent years, applications of data visualisation have increased by leaps and bounds with new visualisation techniques constantly being developed to assist users in deriving useful insights from seemingly meaningless data.

However, scholarship in this field has not kept pace. This has resulted in a situation where the current popularly implemented practices are based upon pre-supposed assumptions which may be found to have no hard basis in reality.

An important finding is that the principle of "less is more" holds true. This means that while visualising complex datasets, it is important to choose characteristics and features that are important to plot. This reduces the complexity of the resulting visualisation and makes it simpler to comprehend.

It has been noted that interactive data is easier to draw conclusions from. The effectiveness of data visualisation techniques is also found to be greatly dependant on the nature of the intended target audience. For example, trained medical professionals are able to plot and read complex healthcare data easily, whereas users without any formal training struggle to create complex data representations using advanced tools.

III. METHODOLOGY

The final product is a website that integrates all the data visualisation functionality into a single coherent interface that is easy to navigate and plot graphs with.

Data that is to be used is first uploaded to the site and stored onto it. This data can then be used to create multiple different types of graphs for the purpose of visualising the data and highlighting different aspects of it. These are also stored on the server. Storing datasets and their corresponding graphs on the site itself enables them to be accessible from anywhere.

The type of graph to be created is first taken as input. Next, the fields or features which are to be represented and used to create the graph need to be selected. Finally, on the basis of these parameters, the server creates the data visualisation and presents it to the user.

Processing huge datasets is a computationally expensive and memory intensive task. Thus, by using a server-based architecture, as opposed to a client-based one, even users operating the software though weak hardware, such as mobile devices, are able to take advantage of the benefits conferred by being able to visualise huge datasets. A server-based approach makes data visualising capabilities widely accessible, over a wide spectrum of devices.

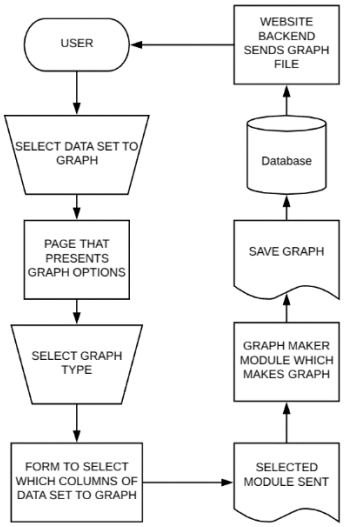

Fig.1 Control flow diagram of the system

The interface was created to be simple, fast and responsive; and guide users in the correct direction of operation using simple design cues. Datasets and their corresponding graphs are organised in an efficient and orderly manner which makes them easy to find, read and access quickly.

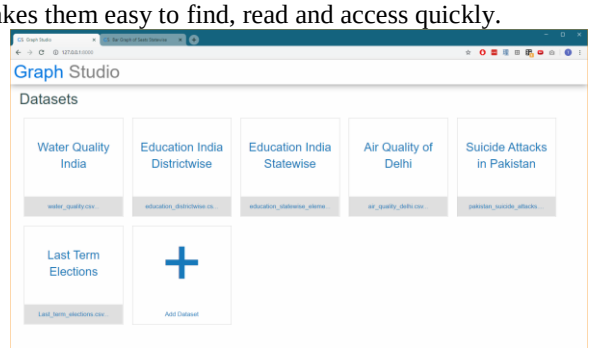

Fig.2 Website homepage

IV. RESULTS AND DISCUSSIONS

The following are case studies which demonstrate the functionality and practical application of the project on real-world datasets that describe the situation in various different public sectors of India.

IV.I. Water Quality in India

The dataset we are using here has the details of the water quality in every state of India for the years 2009 to 2012. The dataset specifies which blocks of which states are affected by a particular pollutant. For example, if there are 100 blocks in West Bengal affected by iron pollution in water in 2009, iron as a pollutant would have a value of 100 in the graph/chart.

To visualize this data, we make a pie chart where we can see the overall number of cases for each pollutant in the country. Similar pie charts are created for each state to get an idea about which are the major pollutants that affect a particular place. For each chart, we use the cumulative data from 2009 to 2012 to obtain the number of cases per pollutant. This information could help us to combat water pollution by understanding the nature of pollutants region wise and taking actions to overcome the pollution for a better and cleaner environment.

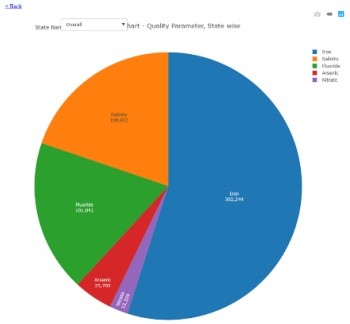

Fig.3 Pie chart of water quality in India

From the overall country level pie chart, we can see that iron is the major pollutant in the country by a large margin. Iron is followed by salinity and fluoride which are close to about 100,000 cases each. Arsenic and nitrate are another two pollutants that are far fewer in their number of cases but are also far more harmful than the other major pollutants.

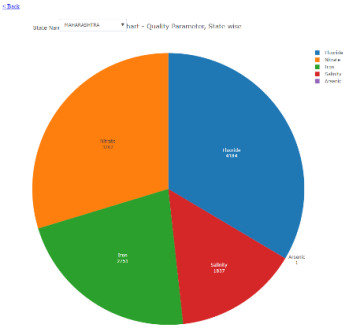

Fig.4 Pie chart of water quality in Madhya Pradesh

Examining further state wise, we can see that cases of iron pollution have taken place in almost every state, barring a few. Salinity and fluoride show a similar pattern, but are absent from more states than iron. Cases of nitrate and arsenic pollution, on the other hand, are concentrated in only a few states.

Now to get an idea of trends in water pollution year after year, we are going to make year wise line graphs for each pollutant in every state and also one for the overall country. This would show us the extent to which each pollutant is affecting water bodies each year and if there is an increase or decrease in the number of cases related to these pollutants.

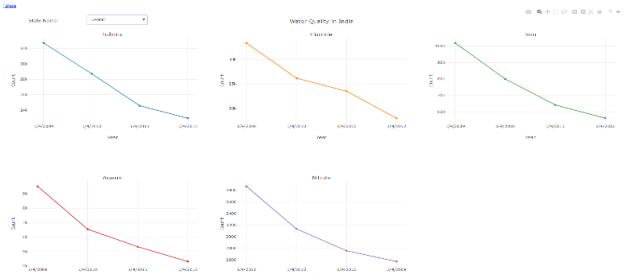

Fig.5 Multiple line plots of water quality metrics

The above graph shows that we are moving in a positive direction when it comes to the number of cases of water pollution. Cases of each pollutant, namely, salinity, fluoride, iron, arsenic and nitrate have been dropping over the years. There has been a positive drop every year after 2009 till 2012. Fluoride and arsenic pollution have seen the biggest drop percentage wise, close to 50%, in the period between 2009 and 2012.

IV.II. Air Quality in Delhi

For this case study, we used publicly available data about the air quality in Delhi over a long period of time, from 1997 to 2016. This covers almost a decade of data about the air quality in Delhi. This is such a massive data set that it would be next to impossible for someone to infer something valuable by just reading it. Our application makes the job easier by generating interactive graphs which make it simpler and more intuitive to read and infer valuable information from the given data.

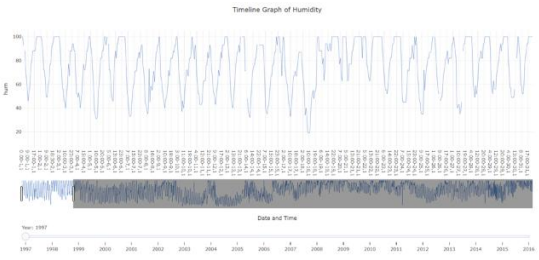

Fig.6 Line graph of humidity in Delhi

We generated interactive graphs that allow the user to understand the air quality data over a period of 10 years by using just sliders to change or narrow down the considered timeframe. In this graph (humidity), one can select the year for which the graph needs to be seen and then use the slider above it to choose the range.

Hovering on the points in the line graph also shows a pop up with a list of all the additional information, if needed.

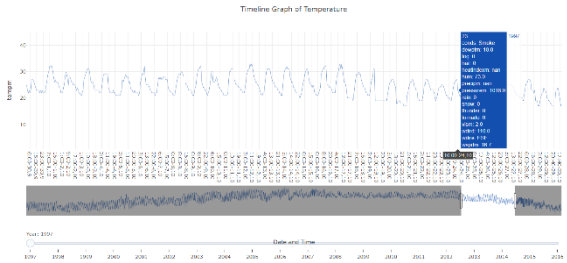

Fig.7 Line graph of temperature in Delhi, with additional information presented in a hover box

In conclusion, we can see that urban areas are generally more problematic when it comes to visibility. This does not come as a surprise because urban areas have more sources of smoke and other particulate matter which directly affect visibility.

A direct correlation could also be seen between temperature and how developed the city was. Temperature was also seen to steadily increase over the years which can be attributed to global warming.

IV.III. Education in India

For this case study, we took data that was available publicly about the literacy rates in India. It was a well-researched and extensive dataset which included not only information about the literacy rate in each district of the country, but also other relevant information like the total population, the name of the district, number of blocks, number of villages, number of clusters, growth rate in literacy, sex ratio, male and female literacy rates and area in square kilometres.

Being such a vast data set of over 600 districts, it would have been extremely difficult to comprehend anything from it without the help of visual information.

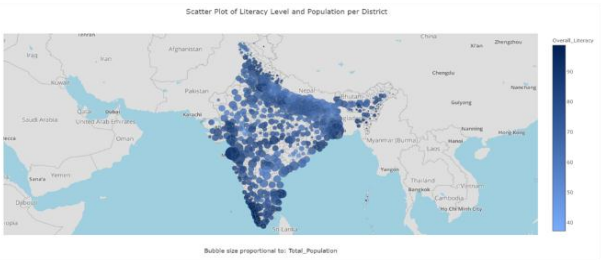

Fig.8 Scatter plot of primary education in India, district wise

This is the scatter plot. The bubbles are coloured according to the overall literacy in the district - the darker the bubble, the higher the literacy rate. The bubble sizes are also proportional to the total population of the district - a smaller bubble means the district has a smaller population and a bigger bubble means the district has a larger population.

The map can be zoomed into for a clearer view of the districts as shown below. The district names are also shown on the map as you zoom in.

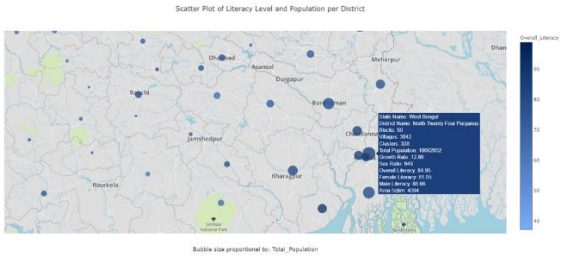

Fig.9 Zoomed in scatter plot

Hovering on a particular bubble shows a popup which has all the information about its corresponding district such as its name, male and female literacy rates, area, total population, etc.

A choropleth was also generated. The states are colour coded depending on the overall literacy of the state - the higher the literacy, the darker the state.

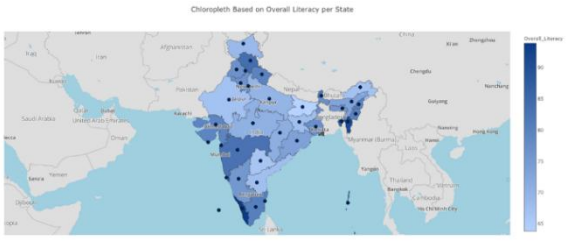

Fig.10 Choropleth of primary education in India, state wise

A few things are evident from studying these graphs:

• There is a direct correlation between how urban a district is and its literacy rate. Urban areas are more educated than rural ones. This should not come as a surprise since education is considered unimportant in most rural areas and there are not enough amenities available for a proper education system to be built there.

• Regions with lesser population seem to be better educated because it is easier to provide facilities to a smaller number of people. This is not always true, however, since there are regions which might be less populated because of reasons like socio-political disturbance, terror, lack of habitable land etc. Those regions, of course, will not be as educated.

IV.IV. Elections in India

Because of the country's really high population, the datasets about the Lok Sabha elections are too big to be understood by a layman. At the same time, it is incredibly important to understand what that data implies to be able to, for example, predict election results from opinion polls.

For this case study, we took data about the previous 2014 Lok Sabha election results and consolidated them to create our own concise dataset which was then visualised for clear understanding.

Future predictions also become of extreme importance in this case and to showcase that, we considered a recent opinion poll carried out by Times Now and created our own consolidated compact dataset, visualised it and then compared it with the graphs generated for the previous Lok Sabha elections.

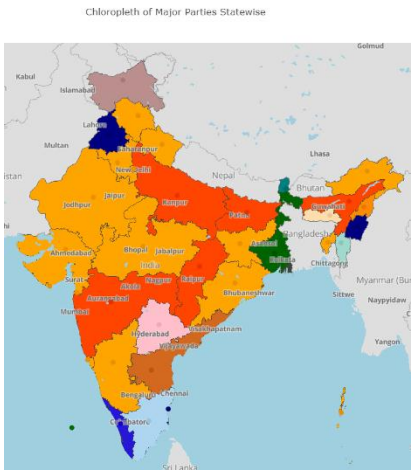

Fig.11 Choropleth of 2019 Lok Sabha election prediction

From the choropleth, one can clearly understand which party was the most popular party in a particular state just by noting the colour of that state. The parties are colour coded according to their alliances. That makes it easier to find out which state's majority party is in alliance with which of the two major alliances - the NDA or the UPA, with the NDA ones being in various shades of orange and the UPA ones being in various shades of blue. The rest of the colours represent other smaller regional parties who are not in alliance with the big two.

A bar graph was also generated as it aids in understanding the seat distribution in a more familiar way. It helps us visually understand which states have more seats and the distribution of seats among the parties of a state. The colours of the bars represent different political parties, following the same colouring convention as was used for the choropleth.

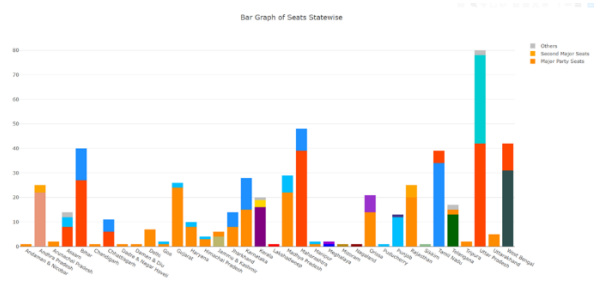

Fig.12 Bar graph of 2019 Lok Sabha election prediction

IV.IV.I. Comparing Predications with Actual Results

The prediction data showed that the NDA would win the 2019 Lok Sabha general elections with a clear majority. It was predicted that the BJP, and the parties allied with it, would gain the greatest number of seats in the majority of states. Comparing this with the actual results of the election, it is found that the predictions have generally held true. NDA candidates won the majority of constituencies in each state. They form a clear majority in the Lok Sabha.

The map of the state wise predictions of the parties with the greatest number of seats in each state is compared with a map of the actual constituency wise election results.

In the former, states where BJP won are colour coded in orange, with states where other NDA parties won denoted in dark orange. Meanwhile, states where INC and its allies are dominant are expressed in various shades of blue. The map shows most of India being covered in orange and dark orange.

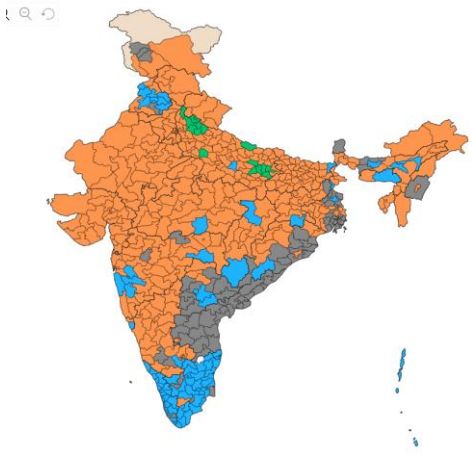

Fig.13 2019 Lok Sabha election results in each constituency

On the other hand, the map of the actual election results shows the seats won by the NDA in orange, those by the UPA in blue and others in grey.

Comparing the two maps shows a strong resemblance in the distribution of seats. NDA and UPA parties were found to be

dominant in the states they were projected to win in by the prediction data.

The overall prediction of which party would dominate in which state is found to be accurate to a considerable degree, other than in the states of North East India (barring Assam) where the number of seats contested is just 1 or 2. However, the number of seats predicted to be won by the NDA was pegged at 279. This is significantly lower than the actual number of seats won by them, which stands at 353.

V. CONCLUSION

Data in the form of text is hard to interpret and fully comprehend. Interactive graphs and diagrams, on the other hand, are much more intuitive and efficient in educating people on any subject. In this project we have seen how huge, seemingly unusable data sets can be made usable by visualising them in an appropriate manner. Taking different datasets, from different public sectors, and making suitable graphs out of them helped us in drawing valuable inferences about the condition of that sector.

This kind of learning is essential in today's world where data is readily available everywhere and the intelligent interpretation of such data is needed more than ever. Such application of data is important for learning and understanding problems, cause, behaviour and outcomes. We believe that such interactive and intuitive learning should be promoted for better understanding of the world around us.

V.I. Limitations

1. The graphs can be only made with structured data in the form of a CSV file.
2. The choice in the types of graphs that can made is limited to the ones implemented in the case studies.
3. The conclusion drawn from the visualised data is dependent on the user, i.e., there is no concrete inference or conclusion made by the software itself.

V.II. Future Scope

At its current stage, the software only works with structured data. The next step would be to implement a way to use unstructured data by converting it into structured data, as the majority of data on the internet is unstructured in format.

Another further addition would be to use machine learning and Artificial Intelligence to predict outcomes by learning from data.

REFERENCES

[1] Mayer-Schonberger, V. & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. *Houghton Mifflin Harcourt.*

[2] Favaretto, M., Clercq, E. & Elger, B. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data.*

[3] Singh, D. & Reddy, C. (2014). A survey on platforms for big data analytics. *Journal of Big Data.*

[4] Healy, K. (2018). Data Visualization: A Practical Introduction. *Princeton University Press.*

[5] Fang, X. & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data.*

[6] K. Parimala, G. Rajkumar, A. Ruba, S. Vijayalakshmi, "Challenges and Opportunities with Big Data", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.5, pp.16-20, 2017

[7] Rakesh. S.Shirsath, Vaibhav A.Desale, Amol. D.Potgantwar, "Big Data Analytical Architecture for Real-Time Applications", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.4, pp.1-8, 2017

[8] 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 Science and Engineering, Vol.5, Issue.6, pp.1-12, 2017

Authors Profile

Asoke Nath, Ph.D., D.Litt is Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. Apart from his normal teaching assignment, he is also actively involved in doing research work in the field of Cryptography and Network security, Steganography, Visual Cryptography, Quantum Computing, Big Data Analytics, Data Science, Green Computing, Li-Fi Technology, Mathematical Modeling in Social Networks, MOOCs, etc. He has published more than **245** publications in International Journals and Conference Proceedings.

Tejash Datta graduated in Computer Science honours from St. Xavier's College, Kolkata in 2019. He has experience developing in various domains, such as games, websites and mobile applications.

Faisal Ahmed graduated in Computer Science honours from St. Xavier's College, Kolkata in 2019. He is currently enrolled in SP Jain School of Global Management, Mumbai, as an MBA student in Marketing, batch of 2019.

Nitin Gupta graduated in Computer Science honours from St. Xavier's College, Kolkata in 2019. He is currently employed as a Google Ads Consultant at Regalix, Hyderabad, since 2019.

