CSE 578: Characterization of Pandemic Spread VAST 2010 Mini Challenge

Rohit Roongta
rroongta@asu.edu
1217947757
Arizona State University

Abstract

The significance of data visualisation is to communicate information and data in an effective and efficient way. There are various different types of graphical representation such as bar/bubble graph, line chart etc which can be used to deliver the information. These representations not only help in analysing the data but also helps in gaining insight about the data. In our project, we have built an interactive web-based dashboard for "IEEE Visual Analytics Science and Technology (VAST) 2010 Mini Challenge 2 - Hospitalization Records: Characterization of Pandemic Spread" to answer the associated Mini-Challenge questions.

1 Introduction

IEEE Visual Analytics Science and Technology (VAST) Challenge is an annual contest which is open for individuals and teams across organisation, academia, and industry to participate. (4). The 2010 VAST Mini-Challenge 2 provides hospital epidemic outbreak records of year 2009 for several cities across the world. For this challenge, we built an interactive data visualisation dashboard which has the capabilities to answer the questions stated in the problem statement. The main goal of the visualisation tool is to analyse the data and provide useful insights from the epidemic outbreak information. (3)

For the dashboard, we have used ReactJS (5) framework to connect various components (graphs) where each component is developed using D3.js (2). For data storage and pre-processing, we have used MongoDB and python scripts.

2 Mini Challenge Overview

The VAST 2010 Mini-Challenge 2 provides epidemic outbreak data set for 11 major cities across the world. The data set is in the form of death records and hospital admittance. The objective of the challenge is to help health officials to analyse the illness across the cities by primarily addressing the below two questions:

1. **Analyze the records to characterize the spread of the disease [MC2.1]:** Symptoms of the disease, Mortality Rates, and Temporal patterns of the onset, peak and recovery of the disease.
2. **Compare the outbreak across cities [MC2.2]:** Timing of outbreaks, Number of people infected, Recovery ability of the individual cities and Anomalies found

3 Dataset

The dataset consists of hospital admit and death records for the 11 cities across the world over a specific period of time. The structure is:

- **Admit Records**: *Patient ID, Admit Date, Age, Gender, Symptoms observed*
- **Death records**: *Patient ID, Date of Death*

The dataset is provided in CSV format where each row represents the patient information such as gender, age, symptoms etc. Overall, it includes both categorical data such as gender, city and ordered data. In total, there are over 6 million records spanned over the time period 16 April 2009 to 29 June 2009.

Preprocessing: For quick rendering and processing, the data was stored in MongoDB as JSON along with derived attributes such as number of cases per age group, cases per day, recovery percentages.

4 Visualisation Design

The various interaction and components of the dashboard (Figure 1) are as follows:

1. **Filter Panel**: User can filter data on the basis of age group(s) and cities as per his interest. User can select one or more options from each filtering category.

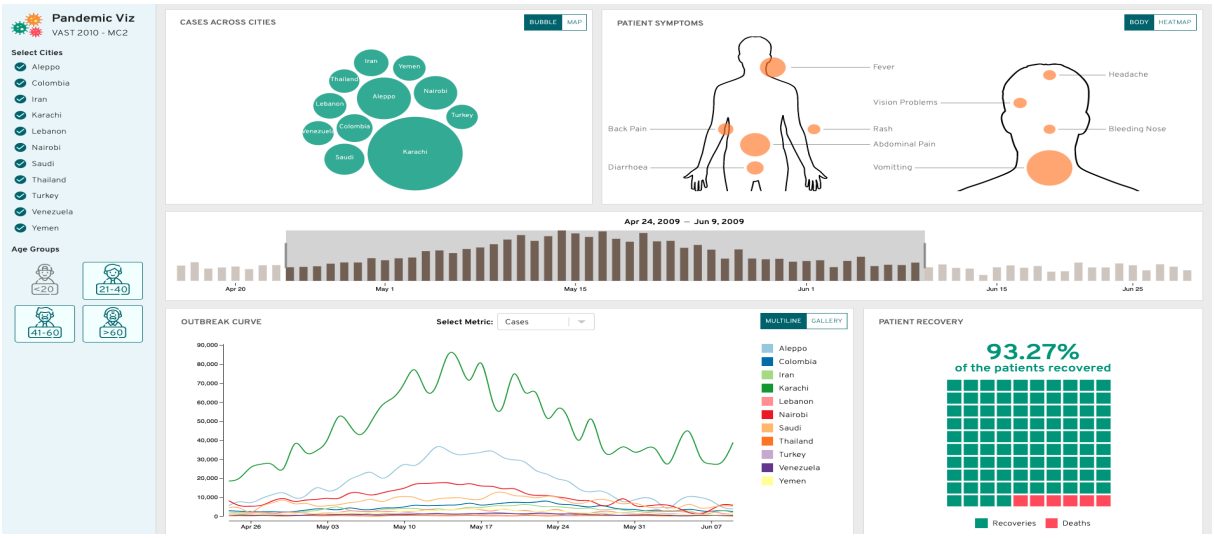

Figure 1: Pandemic Visualisation: Dashboard Overview

2. **Time-frame Sliding Window**: The time frame sliding window is primarily used to filter the data over the specified time range by dragging the window sides. Additionally, it consists of a bar-graph where the height represents the number of patients admitted on a particular day (1).

3. **Cases Across Cities**: Bubble graph is used to depict the number of patients admitted for the selected cities and time period. The size of bubble is relative to number of admittance. A tooltip is displayed on hovering which shows the total cases and total deaths. User can switch to world-map view (Figure 2) where the bubbles are placed in accordance to their geographical location.

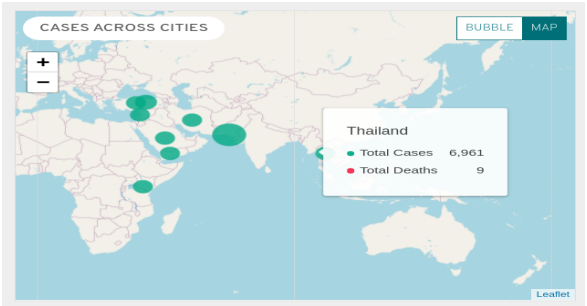

Figure 2: Cases Across Cities: Map-view

4. **Patient Symptoms**: Patient Symptoms are represented using a human body diagram and a heat map. The bubble on human body portrays a particular symptom where the size emphasis on the number of admitted patients for the symptom. For instance, bubble near stomach signifies abdominal pain. On hovering, it displays the number of cases. On the other hand, the heat map (Figure 3) is used for temporal representation where it display the reported cases for each symptom on a particular day.

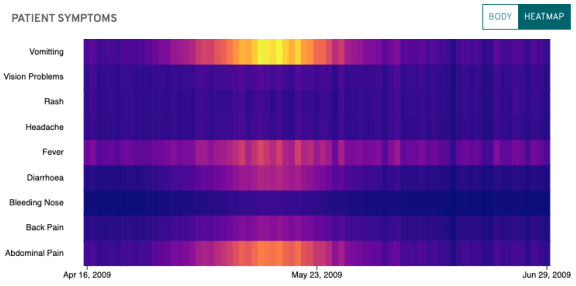

Figure 3: Patient Symptoms (Heat map)

5. **Outbreak curve**: Outbreak curve is represented by a multi-line chart along with filter metrics such as fatality rate, deaths and cases. Each line represents a particular city and on hovering over the line, it highlights and dehighlights the cities. On the other hand, user can toggle to gallery view which consists of outbreak curves for each city. This can be used for comparative studies and analysis.

6. **Patient recovery rate**: Dot matrix is used to display the patient recovery rates. The matrix provides a relation between the number of death with represent to number of cases. It also convey scale, and recovery percentage.

5 Result

The dashboard can be used by health officials for the following analysis and tasks:

5.1 MC2.1

1. **Symptoms**: It has been observed that the most evident symptoms are abdominal pain and vomiting. The heat map also confirms the fact that these symptoms clearly increases along with the pandemic outbreak.
2. **Mortality rate**: Also known as death rate is the number of deaths over a population. This information can be inferred by using the fatality rate and patient recovery rate matrix. The mortality rate for the overall data is 5.29%
3. **Temporal patterns**: The bubble and outbreak graph clearly shows the onset, recovery and peak temporal patterns where the data can be filtered out with the help of sliding window bar graph.

5.2 MC2.2

1. **Timing of outbreaks**: The gallery view of the outbreak curve can be used for comparative studies between the cities. From the onset pattern, it can be observed that there is potential domino effect in the following order: Nairobi $\rightarrow$ Aleppo $\rightarrow$ Lebanon $\rightarrow$ Karachi $\rightarrow$ Yemen $\rightarrow$ Saudi Arabia $\rightarrow$ Iran $\rightarrow$ Venezuela $\rightarrow$ Columbia.
Thailand and Turkey remain unaffected. However, nothing can be concluded about the reason behind the pandemic spread but the potential reason could be travel where the middle eastern region were mainly affected.
2. **Number of infected people**: This information can be inferred from the bubble graph.
3. **Recovery**: With the help of dot matrix, it can observed that Saudi, Iran and Lebanon have the best recovery rate.
4. **Anomalies**: With the help of dot matrix, it can observed that the younger people ($<$20 age) and older people ($>$60 age) have lower recovery rate which can determined by filtering out the data with represent to age groups.

6 Contribution

While the project was mainly the entire team effort, tasks were distributed among the members as per their interest. I proactively participated along different stages of the project. Initially, I analysed the CSV data files and performed pre-processing activities such as preparing JSON file for each graph component so that the data can be rendered quickly and efficiently. Later on, I worked on the integrating the city filter on the sidebar, this gave the functionality of filtering the data across the dashboard with respected to selected cities. After that, I worked on the implementation of the gallery view for the outbreak curve for comparative analysis. Finally, I helped in integrating the all components together and system deployment.

7 Learning

1. I learnt the various front-end libraries ranging from React to D3.js where React was used to structured the project in the form of components and data visualisation logic for each component was implemented with the help of D3.js.
2. Analysing and pre-processing the raw data for efficient data rendering for the visualisation. MongoDB was used to store and query the data accordingly to the various components of the dashboard.
3. As the development progressed, learning the basic UI/UX design and visualisation principle played a crucial role in making the dashboard more user-centric.

8 Team Members

1. Gourav Agrawal (**gagrawa3**)
2. Kshitij Jayprakash Sutar (**ksutar**)
3. Navaneesh Kumar Bajakudlu (**nbajakud**)
4. Rohit Roongta (**rroongta**)
5. Soujanya Ranganatha Bhat (**sranga16**)
6. Sughosh Jahagirdar (**srjahagi**)

References

[1] BOSTOCK, M. d3/d3-brush: Select a one- or two-dimensional region using the mouse or touch., 2020.

[2] BOSTOCK, M. D3.js - data-driven documents, 2020. Available at https://d3js.org/.

[3] HCIL, U. O. M. Visual analytics benchmark repository, 2010. Available at https://www.cs.umd.edu/hcil/varepository/VAST%20Challenge%202010/challenges/MC2%20-%20Characterization%20of%20Pandemic%20Spread/.

[4] IEEE. About ieee vast challenge, 2020. Available at https://vast-challenge.github.io/2020/about.html.

[5] INC., F. React - a javascript library for building user interfaces, 2020. Available at https://reactjs.org/.