Information Visualization Analysis of the Vastopolis Epidemic Spread

VAST 2011 Mini Challenge 1

Ramiro Lucero

Universidad de Buenos Aires

ABSTRACT

This paper analyses the Vast 2011 Mini Challenge 1 using data mining and information visualization techniques in order to answer the Challenge questions.

The main objective of the Challenge is to find the origin and form of transmission of the disease that is affecting the citizens of the Vastopolis metropolitan area.

The analysis performed was done using the PostgreSQL database together with the Tableau Desktop software.

KEYWORDS: data mining, flu.

1 INTRODUCTION

The Vast 2011 Mini Challenge 1 - Characterization of an Epidemic Spread, is focused on Vastopolis, a major metropolitan area with a population of approximately two million residents. During May 2011 professionals at local hospitals have noticed a dramatic increase in reported illnesses in just a few days. The observed symptoms are largely flu-like, and there have been several deaths believed to be associated with this outbreak.

Using the datasets provided, the origin of the epidemic spread needed to be found and also a hypothesis on how the infection was being transmitted had to be postulated. Additionally, it needed to be found if the outbreak was contained on the last day of the dataset and if treatment resources needed to be deployed outside the affected area.

In this paper I describe the steps taken using data mining and information visualization techniques in order to solve the Challenge. The principal software used was the PostgreSQL database together with its PostGis module and Tableau Desktop.

2 THEORY

In this section I describe the materials, methods and results for the analysis.

2.1 Materials

The data provided for the analysis were a set of over a million microblog messages collected from various devices with GPS capabilities from the 30th of April through the 20th of May (fields: ID, Date, Time, Text, Latitude and Longitude for each message), a map with information for the entire metropolitan area (showing hospitals, highways, important landmarks and water bodies), and supplemental tables for population statistics (zone, population and daytime population) and observed weather data (Date, Average Weather, Average Wind Speed and Wind direction).

The tools used for working on this challenge were Postgresql 8.4 database with the PostGIS 1.4 module [1], gvSIG 1.11 [2] and Tableau Desktop 6.0 software [3].

2.2 Methods

First, all the tables provided where loaded into PostgreSQL database. The fields of these tables were correctly formatted and the links between tables generated. Afterwards, the Vastopolis map image was geo-referenced and re-projected to UTM Zone 15 North projection to be able to work in metric coordinates instead of degrees. All main features from the image were digitalized using gv-SIG software. These features were then uploaded to PostgreSQL. All the coordinates of the messages (microblogs) provided were re-projected as well to UTM Zone 15 North.

Second, the data was loaded into Tableau. Several visualizations using the raw data were done. A meaningful one was the plot of the messages coordinates over the map, showing only the messages that had the word "flu" in them, and showing one day at the time. The visualization showed that on the 19th of May many people suddenly mentioned the word "flu", and on the 20th of May many people that mentioned this word did it near Vastopolis hospitals. Many false positives were detected: messages from people that were not ill but mentioned the word flu. To better discriminate the people that were ill from the ones who weren't, a better filter was needed.

To detect the origin of the outbreak, it was considered that apart from detecting correctly the messages from people that were diagnosticated with flu, also the messages where people started to exhibit the symptoms had to be found. To do this, first, all the words related to the flu symptoms (*flu, fever, chills, sweats, aches, pain, fatigue, cough, breath, nausea, vomit* and *diarrhea*) were counted in each message. PostgreSQL **tsvector** and **tsquery** text search were use for counting these words. The words *death*, *ill*, *sick* and *killingme* were also counted.

Using PostGis functions the zone from where each message was sent (Downtown, Eastside, etc.) and also their distance to the main features digitalized from Vastopolis image (hospitals, roads, etc.) were calculated.

Two categories were made for the words that were counted: *Symptoms* and *Symptoms_acute*. *Symptoms* grouped all the words counted, and *Symptoms_acute* only the following: *fever, chills, sweats, aches, pain, fatigue, cough, breath, nausea, vomit, diarrhea* and *killingme*. This distinction between words was made because some of the words counted were also used in texts that were not speaking of illness. *Symptoms_acute* have more reliable words in that sense. Each message was assigned the quantity of words counted from each category. Afterwards, for each ID the variable *symp_add* was generated with the total quantity of words sent from the category *Symptoms*, and the variable *symp_acute_add* with the total quantity of words sent from the *Symptoms_acute* category; this was done for all the messages in the database. Based on the distributions of these two variables the following variable for each ID was created:

ramiroalucero@yahoo.com.ar

$$\mathit{had\_flu\_high} = \begin{cases} 1, & \text{when } \mathit{symp\_add} > 2 \text{ or symp\_acute\_add} > 1. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This new variable helped to decide which ID has more chances of really having caught the flu.

The last variables generated for each ID were the date and time of the first message sent with a *Symptom* word.

After having created these new variables, new visualizations were generated with Tableau. By dynamically filtering the IDs and keeping not only the ones having *had_flu_high*=1, but also having their first message containing *Symptoms* words with two or more *Symptoms_acute* words, I arrived to the following plot presented in the next subsection (Figure 1). This plot made more clear what was happening in.

Basically, people with *had_flu_high=1* and that the first time that they send a message talking about the symptoms they write two or more acute symptoms, they have an even higher probability of being really ill. This filter reduces significantly the number of IDs, but highly guarantees that the IDs selected are ill. Once the ill people have been detected, it's easier to see the pattern of the disease.

2.3 Results

Figure 1 shows an outbreak the 18th of May at 8AM mainly in the Downtown area and with less intensity in the Uptown and Eastside areas. The quantity of people affected starts to diminish abruptly the 18th at 8PM, but the 19th at about 2AM, a new outbreak starts in the Plainville and Westside areas. The rate of infections remains quite stable until the last day registered in the dataset.

By analyzing the messages from each outbreak it was discovered that the people infected in the first one presented mainly fever, chills, sweats and respiratory problems, while the people infected in the second outbreak had more gastric problems (vomit, diarrhea and nausea).

During the day of the first outbreak and the day before the wind was blowing from the West.

3 DISCUSSION

As the of messages from people feeling ill in the first outbreak present respiratory problems and they form a plume (Figure 2) that goes from West to East while the wind that day was coming from the West, this suggests that the disease in the first outbreak is spreading through air.

During the second outbreak, messages from ill people are concentrated along the Vast River (Figure 2). The symptoms in this outbreak are different from the first; they are more related to gastric problems (vomit, diarrhea and nausea). This evidence suggests that the method of transmission of the second outbreak is water.

Looking at Figure 2: As the first outbreak seems to be transmitted by air, and the wind on the 18th was coming from the West, its origin should be at the west side of the white trend lines. The second outbreak is transmitted apparently by water through the Vast River, so its origin should be up the river from where the people got sick. If we consider the possibility that both outbreaks are part of the same disease, the origin of the outbreaks would be inside the green circle.

Figure 1 shows an outbreak the 18th of May at 8AM mainly in the Downtown area and with less intensity in the Uptown and Eastside areas. The quantity of people affected starts to diminish abruptly the 18th at 8PM, but the 19th at about 2AM, a new outbreak starts in the Plainville and Westside areas (along the Vast River). The rate of infections remains quite stable until the last day registered in the dataset. This could indicate that the first outbreak is controlled, but the second one isn't.

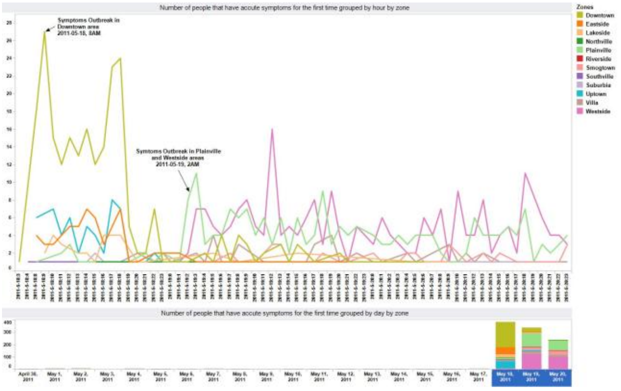

Figure 1. Quantity of IDs sending a message with acute symptoms for the first time, color coded by zone and plotted by hour.

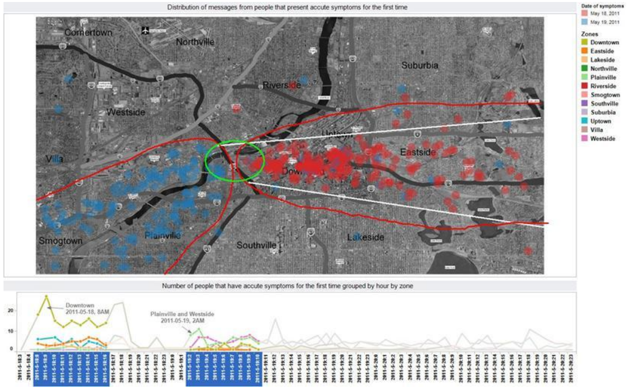

Figure 2. In red the people that were infected in the first outbreak, in blue the people that infected in the second outbreak. Affected areas are outlined in red. The green circle shows the place in which the two outbreaks seem to coincide.

People who got sick along the Vast River presented gastric symptoms. The river flows south outside the area of study, and assuming that the disease is transmitted by water, it would be necessary for emergency management personnel to deploy treatment resources outside the affected area.

REFERENCES

[1] PostgreSQL Database: http://www.postgresql.org.
[2] GvSIG Project: http://www.gvsig.org.
[3] Tableau Desktop: http://www.tableausoftware.com.