EpiDetector: Characterization of Epidemic Outbreak
VAST 2010 Mini Challenge 2

Jaeyeon Kihm§, Jaegul Choo*, Carsten Görg*, Hanseung Lee§, Zhicheng Liu*, Haesun Park*, John Stasko*
Georgia Institute of Technology

ABSTRACT

We created a visual analytics tool called EpiDetector for the VAST 2010 Mini Challenge 2. The system visualizes hospitalization records across different cities involved in an epidemic outbreak. We began our analysis process by cleaning the data and aggregating many different symptoms into eight main syndromes. EpiDetector then presents hospital admittances, mortalities, length of hospital stay, and contributing symptoms and syndromes over time. Using EpiDetector we were able to identify characteristics of the onset, spread, and decline of the epidemic.

KEYWORDS: VAST Challenge, visual analytics, disease outbreak, epidemic, natural language processing

INDEX TERMS: H.2.8 [Database Management]: Database Applications—Data Mining; I.2.7 [Artificial Intelligence]: Natural Language Processing—Language Parsing and Understanding; J.3 [Life and Medical Sciences]: Medical Information Systems

1 PROBLEM OVERVIEW

The Vast 2010 Mini Challenge 2 asks participants to analyze hospitalization records across cities for characterizing the spread of diseases. The provided data are the hospital admittance and the patient death records for eleven locations involved in the epidemic. The hospital admittance records are composed of the admit date, patient gender, id, age, and the symptoms. Death records include the patient id and the date of death.

2 EPIDETECTOR

We created a visual analytics tool, EpiDetector, to explore characteristics of epidemic outbreaks. EpiDetector has three different interactive views:

- Control Panel: provides controls for selecting and filtering the data based on either a city/location or a syndrome.
- Overview of hospitalization records (see Figure 1): the upper graph shows a timeline with the number of admitted patients per date or the number of patients who were admitted and died on a date, along with the associated medical syndrome. The lower graph shows mortalities and the number of days from the patient's admittance to their death.
- Syndrome composition view (see Figure 2): this window shows a list of syndromes and the set of symptoms making up each.

3 METHOD

Since characterizing the spread of the disease was the main focus of this challenge, describing which syndrome occurred where and when is of vital importance. First, we had to "clean" the noisy, free-text medical symptoms that included many abbreviations and misspelled words. We decided to take all the different symptoms and classify them into eight main syndromes monitored by the Real-time Outbreak Disease Surveillance (RODS) [1] at the University of Pittsburgh. We thought that eight simple syndromes would be more useful to recognize the epidemic outbreak than all the different symptoms.

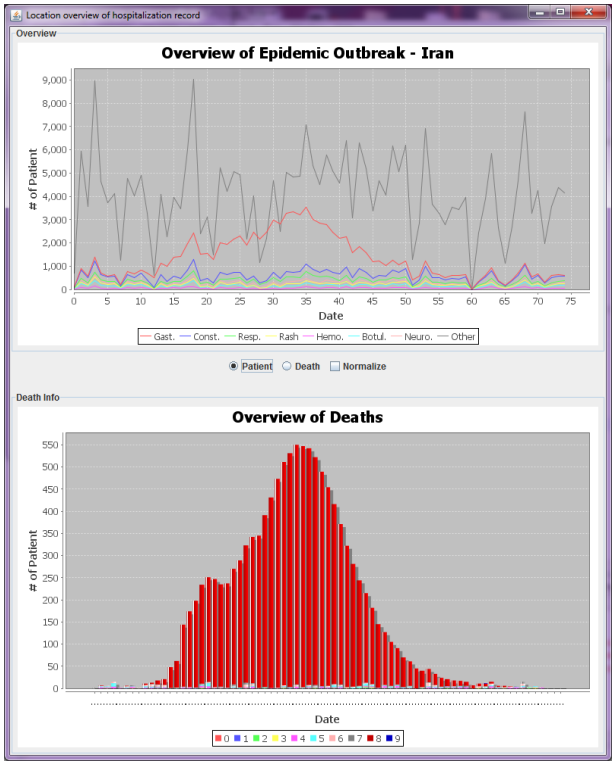

Figure 1. Location overview of hopitalization records – Iran.

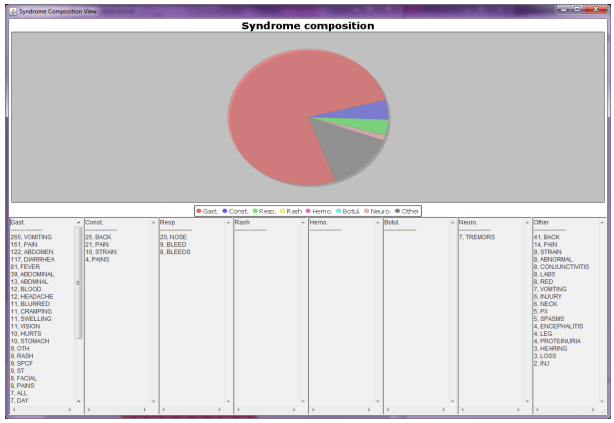

Figure 2. Syndrome composition view – May 18, 2009 in Iran.

* e-mail: {joyfull,goerg,zcliu,hpark,stasko}@cc.gatech.edu
§ e-mail: {jkihm3,hanseung.lee}@gatech.edu

3.1 Text Preprocessing

We used three different natural language processing techniques to clean the free-text symptoms. First, we identified text strings that actually were two separate words merged together into one word, such as "diarrheavomiting", and we split them. We iterated through each symptom, one letter at a time, checking the string so far as being a valid term in a medical dictionary. Second, we detected duplicated words like "headacheheadache" using a similar idea and we removed the duplication. Finally, we expanded abbreviations, such as "ab" or "abd" for "abdomen", by checking known lists of abbreviations. We could clean many ambiguous symptoms by applying these preprocessing techniques.

3.2 Syndrome classification

We then classified all the different symptoms into eight primary syndromes. We created a rule-based classification scheme using RODS syndromic definitions to determine which symptoms belonged to which syndromes.

4 Mortality Rates

We analyzed the mortality rate of each syndrome by showing the number of patients who died over time using a bar chart. The x-axis indicates the progression of days in time. The bar for each date encodes, through its height, the number of people who entered the hospital on that date. The bars are colored in segments according to how many days each person was in the hospital before dying. The chart at the bottom in Figure 1 shows this data for Iran, and we can clearly see the rise in the number of deaths in the middle of the period. Similarly, most of the bars are composed of large red regions which correspond to the person being in the hospital for eight days before passing away. To find out which syndromes are most connected to the epidemic, we compared the pattern of each syndrome in the upper graph with the trend in the lower graph (see Figure 1). Since the pattern of the gastrointestinal syndromes in the upper graph is most similar to the trend of the epidemic in the lower graph, we suspect that the gastrointestinal syndrome and the epidemic are strongly correlated. Specifically, we can see that the number of patients having a gastrointestinal syndrome is dominant among the patients who died in eight days on May 18, 2009 (see Figure 2). The most frequent symptoms on the death records are vomiting (265 occurrences), pain (161), abdomen (122), diarrhea (117) and fever (81) (see left column in Figure 2). We found the same pattern of death records for all other locations except Thailand and Turkey (see Figure 3). Thus, we suspect that an infectious disease broke out and was transmitted to all locations except for those two.

5 Outbreak Patterns Across Cities

We found that the epidemic disease typically makes patients die in eight days. Table 1 shows the outbreak pattern across the different locations.

	Onset	Peak	Recovery
Nairobi	April 20	May 14	June 16
Lebanon	April 22	May 16	June 18
Venezuela	April 22	May 18	June 19
Aleppo	April 24	May 15	June 17
Yemen	April 24	May 17	June 18
Karachi	April 24	May 17	June 18
Iran	April 24	May 18	June 20
Saudi Arabia	April 25	May 18	June 20
Colombia	April 26	May 20	June 19

Table 1. Onset, peak, and recovery dates of the outbreaks.

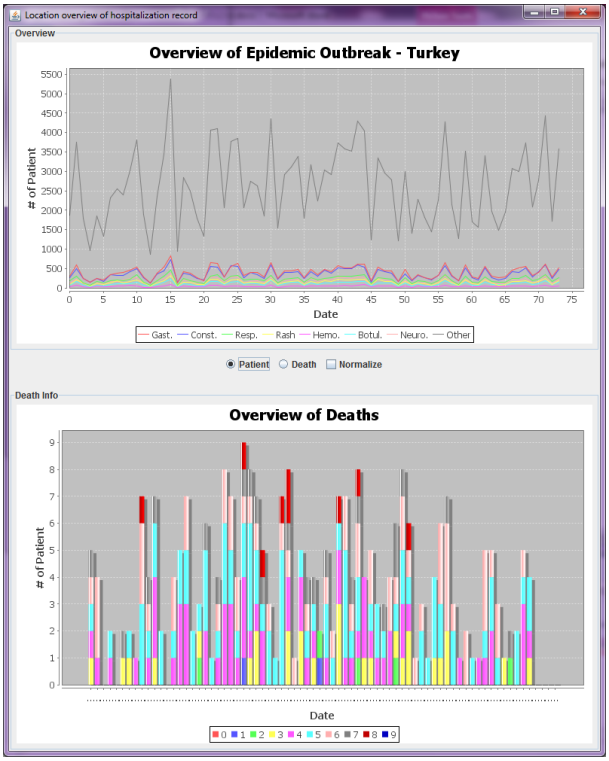

Figure 3. Location overview of hospitalization records – Turkey.

We defined the onset to be the date of the first suspected death, the peak to be the day having most suspected deaths, and the recovery to be the last day a suspected death occurred. Locations in the table are sorted by the onset date.

Examining the data, we noticed that the epidemic appeared to begin in Nairobi, Kenya with early onsets also in Venezuela and Lebanon. It then quickly spread to Syria, Yemen, Pakistan, Iran, Saudi Arabia, and Colombia. One might expect that neighboring countries such as India, Iran, Iraq, Ethiopia, Egypt, or Brazil would soon be at risk.

6 Conclusion

Using EpiDetector, we analyzed the epidemic outbreak and characterized the pandemic spread from the huge hospitalization records dataset. Our process consisted of an initial phase in which we cleaned the data and classified different symptoms into eight main syndromes. We then used EpiDetector to explore the data interactively.

Acknowledgements

The work of these authors was supported in part by the National Science Foundation grants CCF-0808863 and IIS-0915788 and the VACCINE Center, a Department of Homeland Security Center of Excellence in Command, Control and Interoperability.

References

[1] Chapman, W. W., Christensen, L. M., Wagner, M. M., Haug, P. J., Ivanov, O., Dowling, J. N., and Olszewski, R. T. 2005. Classifying free-text triage chief complaints into syndromic categories with natural languages processing. *Artif. Intell. Med.* 33, 1 (Jan. 2005), 31-40.