VAST Challenge 2014: Mini-Challenge 1

Qiang Song* Rui Li Peng Yin
University of Chinese Academy of Sciences

ABSTRACT

In this paper, we address the Mini Task I of Min-Challenge I, pluging in knowledge and some techniques of data mining and information visualization into our assignment. The background of Mini-Challenge I describes a virtual story of a famous gas company(GAStech), which had suffered employees missing during company celebration, asking for help to find out the suspicious criminals. As an analyst, our task is to use visual analytics to analyze the available data and develop responses to the questions. By using Matlab, Python and D3.js, we gave out a reasonable conclusion and proper answers to the questions.

Keywords: Data mining, information visualization.

1 INTRODUCTION

In the roughly twenty years that Tethys-based GAStech has been operating a natural gas production site in the island country of Kronos, it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship. In January, 2014, the leaders of GAStech are celebrating their new-found fortune as a result of the initial public offering of their very successful company. In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem.

As an expert in visual analytics, we are called in to help law enforcement from Kronos and Tethys assess the situation and figure out where the missing employees are and how to get them home again. Time is of the essence. Mini-Challenge 1 focuses on the disappearance. As an analyst, our task is to bring law enforcement up to date on the current organization of the POK and how that organization has changed over time, as well as to characterize the events surrounding the disappearance. the questions are as follows:

1. Provide a visual representation of the structure of the Protectors of Kronos network;
2. Provide novel visualizations appropriate for communicating key information to the busy leaders of the investigation;
3. Describe the events of January 20-21, 2014;
4. Identify at least two possible explanations why the GAStech employees may be missing;
5. What evidence do you have to support each of these explanations.

2 THEORY

We try to address the Mini Task I of Min-Challenge I, pluging knowledge and some techniques of data mining and information visualization into our assignment. Subsection One presents the proposed Latent Dirichlet Allocation(LDA) algorithm[1] for topics mining and then describes our novel visualization of the events happened in January 20 and 21, 2014. Subsection Two uses a various visualization models to analyze the available data and develops responses to the questions respectively..

2.1 Subsection One

Latent Dirichlet Allocation(LDA)[1] is a generative probabilistic model that uses a set of "topics", distributions over a fixed vocabulary, to describe a corpus of documents. In its generative process, each document is endowed with a Dirichlet distributed vector of topic proportions, and each word of the document is assumed drawn by first drawing a topic assignment from those proportions and then drawing the word from the corresponding topic distribution.

$$p(w|d) = p(w|z)p(z|d) \quad (1)$$

where z is the topic latent variable.

Among those several available data given, there are some relevant current and historical news reports from multiple domestic and translated foreign sources, which in text file format. Thus, we try to preprocess these text files, cleaning up the irrelevant ones and generating proper topic vector format for the rest. After that we fine tune the topics manually to well-understood words.

Here, we express the topics sequence as a vector.

[“Waiting for interview”, ”The capital building”, “A fire alarm”, “People flee out”,“Fire Department is coming”, “A bomb threat”, “Refuse to our interview”, “Police Involved in the investigation”,…]

Besides, our novel visualization of the events is as follows: we use a vivid show to represent the events happened on the news report in time sequence. On Figure 1, the top-left corner is a scrolling news box which shows the up-to-date news of the incident. The top-right corner is a scrolling image box which visualizes the content as images, thus helping readers understand what is going on more easily. We found that a lot of 'things' happened one by one on the very day, and even who are not “experts” can more or less conclude that some of what happened are diversions which lead people's attention to somewhere else. We can say that there must be some truth hidden behind the missing incident.

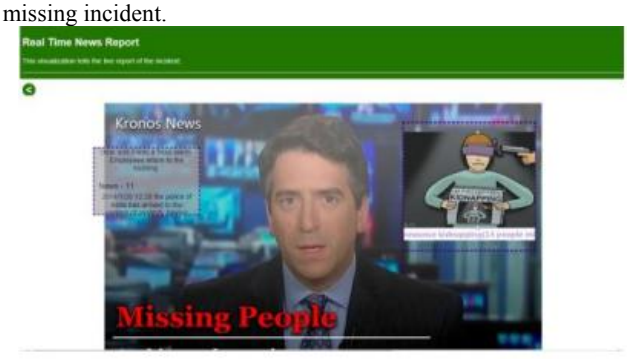

Figure 1: A live News Report

* qiang.song@nlpr.ia.ac.cn

LEAVE 0.5 INCH SPACE AT BOTTOM OF LEFT COLUMN ON FIRST PAGE FOR COPYRIGHT BLOCK

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

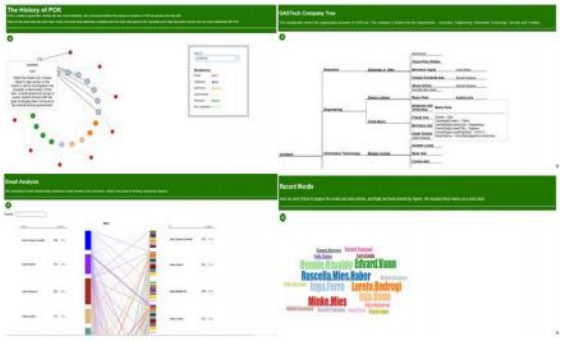

Figure 2: Various Models for Visualization

2.2 Subsection Two

The exploration of large data sets is an important but difficult problem. Information visualization techniques may help to solve the problem.[2] Various basic models have been used for task visualization separately, including network, multi-way tree, bipartite graph, word cloud, etc. The network describes the history of the Protectors of Kronos (POK). From the visualization above we may find that there are two "circles". The nodes on the outer circle represent the detailed events in clock-wise order, and the inner circle expresses key figures and organizations which have some kinds of relationships with POK. The different colors of the nodes represent different roles. According to the Notation, We can clearly learn that these roles include Event, Originator, GASTech Members, Governors, some Local Communities, etc. We can move mouse on each node to obtain the detailed information, or use the "Retrieval Button" to query key words and the matched node will be highlighted. Besides, here we would like to mention that some additional information, such as "The death of Elian Karel" and "POK labeled a Public Threat" is obtained from the News report. Moreover, we also use the multi-way tree to visualize the structure of GASTech Company, from which the detailed profile can be shown clearly. And we visualized email relationships between those email senders and receivers in bipartite graph, which may help in finding suspicious figures. The bars on the left sides represent email senders, and on the right email receivers. When we click on a bar, the bar will be highlighted, showing the connection relationship between senders and receivers about that bar. We found that the stuff who sent everyone email may not be suspicious in that those emails are just administrative emails with no value for investigation. What paid our attention is those who sent emails to several people regularly. We checked those emails manually and found that those emails were indeed related to the company's benefit. With analysis and visualizations above, we found some suspicious characters in the company. We visualized these characters as a word cloud, with larger font size the more importance. Those characters were chosen according to the following clues:

1. Whether did their names have any relationships with POK members. For example, Both Edvard.Vann (GASTech Employee) and Mandor.Vann (POK Leader) have character "Vann". Thus, we can suspect that maybe they have blood relationship.

2. Whether did their email communications have content about the company's benefit? We check the email communication and dig out the suspicious figures.

3 Discussion

After analyzing the given ingredients, we got a series of clues. For instances:

1) We obtained clues of kidnapping from the news reports. By mining the news articles, we found that the POK released a statement, claiming responsibility for the kidnapping of GAStech employees and demanding a $20 million ransom.

2) In 2009, Kaerl died and Mandor Vann became new leader of POK. From then on, POK had taken more radical action to protect Kronos and protest GASTech. In 2011, POK was labeled as a public treat.

3) Some employees of GAStech are members of POK organization. We found the suspicious people have blood relationship with POK.

4) GASTech has made a big success on gas mining in recent years. It was the CEO's decision to have a celebration, and the date was chosen by himself. However, On the very day, the CEO was not appeared and 'fortunately' avoided being 'kidnapped' by POK. We may infer from the above situations that the CEO had got rumors that POK might take radical actions on GAStech. Nevertheless, for his own benefit, he didn't prevent this kidnapping, for he might get more benefits from the case.

When comes to the available data given, we may suffer a large quantity of information in various formats. A vital problem is how to mine the data properly and visualize friendly. In this paper, we address the task with knowledge and techniques obtained from our study research. The tools we used here include Python(A Python is a constricting snake belonging to the Python (genus), or, more generally, any snake in the family Pythonidae (containing the Python genus). Here we use for data mining) and D3.js(a JavaScript library that uses digital data to drive the creation and control of dynamic and interactive graphical forms which run in web browsers. Here we use for results visualization). We also use some friendly HCI, such as "retrieval button" , " mouse click" , "highlight bar" to help the users well understood. All the visualized modules are unified into a single interface. And a drawback of our paper is that we just use some basic models for visualization without any hierarchical or mixed ones.

4 Conclusion

By mining and analyzing the given data, we concluded two possible reasons to explain why the GAStech employees may be missing as follows:

1. Those missing employees might be kidnapped by POK.
2. Those missing employees might be set up by CEO.

Detailed information can be seen from the video and the submission file.

The ultimate goal is to bring the power of visualization technology into problems to allow a better, faster and more intuitive exploration of very large data resources. This will not only be valuable in an economic sense but will also stimulate and delight the user.

References

[1] D.Blei, A.Ng, and M.Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003..

[2] Daniel Keim, "Information Visualization and Visual Data Mining", IEEE Trans. on Visualization and Computer Graphics, Vol. 8, No. 1, Jan-Mar 2002..