Integrated Visual Analytics Tool for Heterogeneous Text Data

Jihyoun Park*

Hong Kong Applied Science and Technology Research Institute

Abstract

Our self-developed java-based visual analytic tool reads a variety of different text data sources and extracts important keywords, relations and events from them using ontology and natural language processing methods. Finally it provides an integrated and interactive search interface to users to facilitate their effective and efficient investigation for the large and complex data set.

Keywords: VAST 2014, mini challenge 1, visual Analytics, ontology modeling, natural language processing.

Index Terms: [Visualization]: Visualization application domains—Visual analytics

1 Introduction

Visual analytics is defined as the science of analytical reasoning helped by software using interactive visual interfaces [1]. As a combination of visualization and automatic analysis, visual analytics can be more effective and efficient than each solution alone, if the data set provided to solve a problem contains many different types of data, thus requires a different level of human and machine involvement to properly deal with them [2].

The mini-challenge 1 asked us to investigate the background of two major organizations related with the recent kidnapping incident in Kronos, POK and GAStech, to find out facts, which may be related to or affected this incident. We were also required to summarize events for two days when 14 GAStech employees were found missing by reading a variety of data sources. The challenge provided 6 different types of text data to tackle this issue. Firstly, we created software to automate the analysis and aggregation of data in order to help organize and extract information from the massive data. Besides, we added a user interface to manually verify and modify the automatically analyzed results so that users could check the basis of their decision making. Then the final aggregator and query interface supported human investigators to find out the evidence to answer questions.

The rest of the paper is organized as follows. In section 2, the text data analysis methodologies adopted in this paper are explained. Section 3 introduces the visual analytics tool that we created to solve the mini-challenge 1. Section 4 concludes with the summary and findings.

2 Data Analysis

Hogenboom et al. [3] classified methods to extract events from text roughly into two categories: data-driven and knowledge-driven. The latter has again two sub-methods: lexico-syntactic patterns and lexico-semantic patterns according to the characteristics of text. As our data sources have different formats and types, we employed the hybrid method. For example, we used the ontology modelling, which is a lexico-semantic pattern, to build relations of organizations and people, and lexico-syntactic-based keyword extraction methods and data-driven methods to summarize news with a set of important keywords.

Before applying those methods, we analyzed the types of given data (Figure 1). Structured data such as GAStech employment records and email headers of internal GAStech employees can be handled automatically to analyze the contents. We classified countries' fact sheets and resumes as semi-structured data, because they had clearly divided sections but some contents were written in a free-style text format. After our tool roughly created the ontology model based on the title of each section, users were required to review the structure manually. We also had unstructured data such as POK history reports. In this case, we decided to input the ontology model of the data manually for more accurate analysis. Lastly, for the 845 news reports, we needed a more systematic way to summarize the contents. The semi-structured parts like published date and journal name were gathered separately, and then we applied natural language processing algorithms to automatically extract important keywords from the news text.

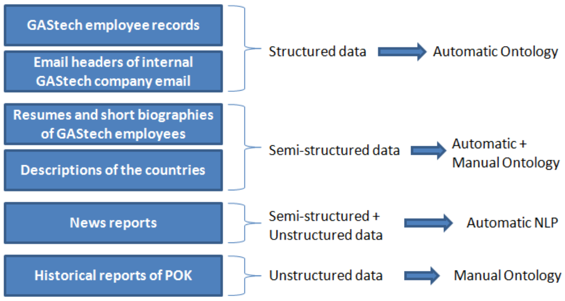

GAStech employee records
Email headers of internal GAStech company email
Structured data
Automatic Ontology
Resumes and short biographies of GAStech employees
Descriptions of the countries
Semi-structured data
Automatic + Manual Ontology
News reports
Semi-structured + Unstructured data
Automatic NLP
Historical reports of POK
Unstructured data
Manual Ontology

Figure 1: Data type and text analysis method.

2.1 Ontology Modeling

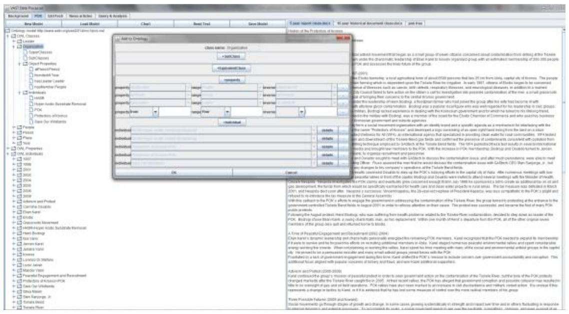


Figure 2: Ontology modeling interface.

Ontology is a handy method to enable semantic analysis on data. An ontology model is consisted of objects and their properties and individuals. It is useful especially when you'd like to find out indirect relations among objects or individuals. And this model is flexible to define object structure unlike database, so you can use

* jhpark@astri.org

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

your expert knowledge more easily to address important terms for a particular data set. Most of ontology modelling was done in automatic processes. Our software reads documents, parses the structure and creates ontology classes, object properties and individuals. In addition, we provide a review and editing interface for users to manually add more information (Figure 2).

2.2 Natural Language Processing

To determine the importance of a keyword in a news report, we used Term-Frequency Inverse-Document-Frequency measure. This measure highlights unique events of a certain news contents from all the other news contents by eliminating common terms. Furthermore, as a combined word (collocation keyword) like Isia Vann can give us more meaningful and relevant information than each individual word, Isia and Vann, we used Part of Speech tags and custom dictionary to register particular people, organizations and places that we want to trace after (Figure 3). Important persons, places and organizations identified from the previous ontology analysis are saved in our custom dictionary to extract keywords from news articles more accurately.

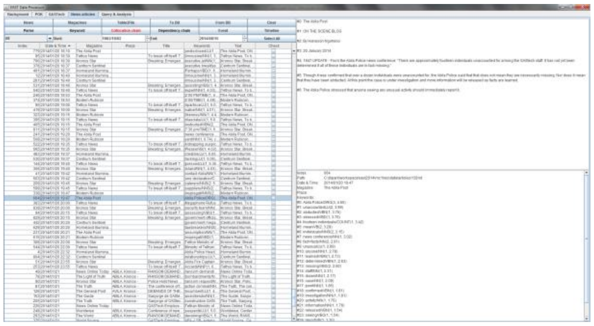


Figure 3: Collocation keywords analysis.

3 Visual Analytics Tool

We developed a java-based visual analytic tool for the comprehensive data analysis and investigation of all available data. The main design concept was to build a consolidated view and interactive interface for users to conveniently query the basic analysis results to compose the final conclusion.

3.1 Ontology Analysis Visualization

The ontology model of country fact sheets, GASTech employee data and POK history reports are represented in a tree structure (Figure 4).

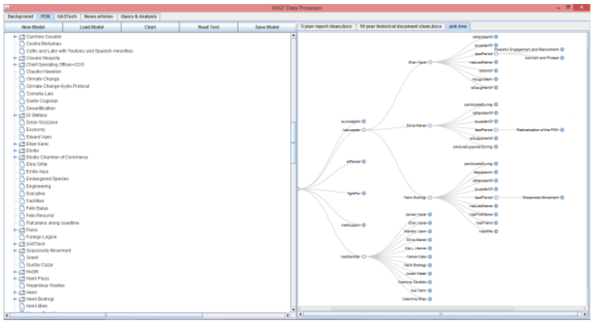
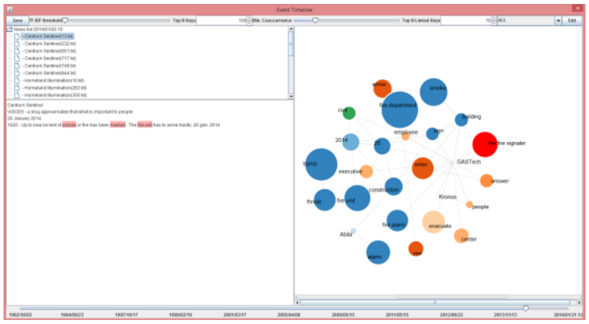


Figure 4: Ontology model tree.

3.2 Keywords Visualization

Keywords are aggregated according to the date of when the original news was published and calculated the co-occurrence count among keywords to find the linked set of keywords. Users can review the list of news in a certain day in the left window and quickly check the important events of the day in the right window (Figure 5).



Centrum Sentinel

VOICES - a blog approximately that what is important to people

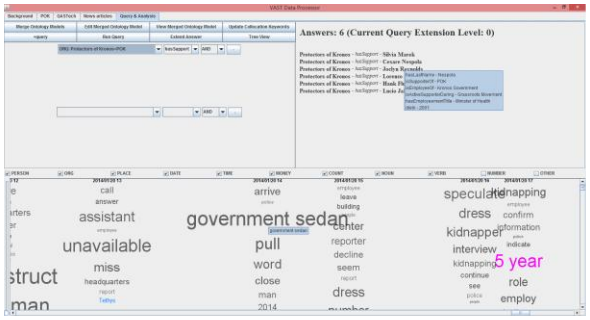
20 January 2014

1025 - Up to now no test of smoke or fire has been marked. The fire unit has to arrive hardly. 20 gen. 2014

Figure 5: News keywords bubble chart.

3.3 Query and Analysis

Finally, the Query and Analysis module combines all the previous investigation results and provides users an integrated interface to dynamically search relations among organizations from their internal links and to be able to look at all the related information like published news articles for them (Figure 6). To assist the data investigation, we added a couple of more helpers such as a pop-up dialog of original news texts with highlighted keywords when users click a keyword on the screen, and more visualization options using D3 charts [4].



Answers: 6 (Current Query Extension Level: 0)

Protectors of Kronos - hasSupport - Silvia Marek
Protectors of Kronos - hasSupport - Cesare Nespola
Protectors of Kronos - hasSupport - Jodyn Reynolds
Protectors of Kronos - hasSupport - Lorenzo
Protectors of Kronos - hasSupport - Hank Fa
Protectors of Kronos - hasSupport - Lucio Ja

Figure 6: Consolidated query view.

4 Conclusion

Visual analytics helps us find the right information from huge and unorganized data quickly and easily. In this mini-challenge, we exerted to provide users an integrated tool to link all the available heterogeneous data in order to give them a comprehensive view of the problem and help them reach to the right answer after all.

References

[1] J. Thomas and K. Cook. A Visual Analytics Agenda. *IEEE Computer Graphics and Applications*, 26:10–13, January/February 2006.

[2] D. Keim et al. Visual Analytics: How Much Visualization and How Much Analytics? ACM SIGKDD Explorations Newsletter, 11(2):5-8, December 2009.

[3] F. Hogenboom et al. An overview of event extraction from text. In M. van Erp et al., editors, *Proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011)* (Bonn, Germany, October 23, 2011), pages 48-57, 2011.

[4] Data Driven Documents http://d3js.org/