Interactive Visual Analysis for Comprehensive Dataset

Jihyoun Park, Yang Liu(✉), Carlos K.F. Tse, Minjing Mao, Mengte Miao, Kangheng Wu, and Zhibin Lei

Hong Kong Applied Science and Technology Research Institute, ASTRI, Shatin, Hong Kong
{jhpark, yangliu, carlostse, minjingmao, mtmiao, khwu, lei}@astri.org

Abstract. In the big data era, handling the volume, velocity and variety of data is the prime requirement for analyzing an event. This paper presents our work for interactive visual analysis software with comprehensive format data input to solve such issues. There are three subsystems to process different types and formats of public and personal data at the same time. A detailed case study shows that the tool efficiently finds the target people and location from various data sources without any offline training or manual search.

Keywords: Interactive · Visual analysis · Investigation · Comprehensive data

1 Introduction

As the data collected and supplied for analysis of a social event are getting bigger and more diverse, it is increasing the demand of proper tools to assist human analyst to process those vast and various data comprehensively as well as in time. That is the challenge of the big data era that we are already living in. The big data can be defined by three main characteristics: volume, velocity and veracity [1]. The amount of data that we need to scan to reach to all related information of a social event becomes huge (volume). They are updated continuously in online. Moreover, the online issues are volatile and talking about immediate events at the time, so analysts are required to interpret the messages in near-real time (velocity). As they come from various channels, the types and formats of data are all different (variety). Therefore, it is crucial to provide data analysts a right tool that is capable to synthesize pieces of scattered data together and reconstruct the whole truth without much delay, in order to respond the data affluence phenomena of these days. Visually displaying the summary and relationship of collected data can be very useful for users to get comprehensive view of the situation effectively and efficiently rather than a simple classification. The application area of such visual analysis tools encompasses from marketing companies, which study their market and target customers, and to police officers who investigate the crime scenes.

In this paper, we propose an interactive visual analysis tool, which takes a comprehensive set of data as the input. The input data set includes public data such as

X. He et al. (Eds.): IScIDE 2015, Part I, LNCS 9242, pp. 451–461, 2015.
DOI: 10.1007/978-3-319-23989-7_46

online news and public announcement, and personal data such as emails, personal profiles, GPS information, credit card consumption records, and twitter-like text stream. The purpose of this tool as an analysis assistant is bringing up interesting targets for further investigation by combining all the available information. The tool is divided into three interacted subsystems to process different media (time and location-related records, static documents and stream texts) separately. Additionally, the information of one subsystem can interactively trigger the operations of other subsystems. To reduce the response time, we applied pattern recognition techniques to automatically find and recommend interesting targets without any prior knowledge or training. In addition, the related events to the targets are also identified and provided to users. In this way, this tool is expected to save the expensive human labor, which use to be used to manually check the relationship of raw data, for more sophisticated analysis.

The rest of this paper is composed of five parts. In Sect. 2, we review related works that studied heterogeneous input data set issues for comprehensive data analysis. In Sect. 3, the detail design of our visual analysis tool and its three subsystems to deal with a particular data type is presented. Section 4 shows a case study to investigate a crime incident using our tool. Finally, Sect. 6 summarizes our work.

2 Related Works

Visual analytics is defined as the science of analytical reasoning helped by software using interactive visual interfaces [2]. 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 [3].

Public data (online news and micro blogs) and personal data (GPS records, credit card records, and smart phone logs) are studied extensively for separate purposes. Generally, text from public data was analyzed to find topic evolution over time [4, 5], while information of personal data was seized for business such as advertisement and sales [5, 6]. Accordingly, as far as we can see, there is still a practice gap in the integration analysis of public data and personal data for investigation for specific events, such as clues to solve criminal case and civil case. The limited research results we found in investigation is simple, such as using suspect information as index, and focusing on one function (document management, forensic sample) [7, 8]. On the other hand, with the explosion of comprehensive data, more evidence should be available while it is harder to find the gold manually. Therefore, there is a need for an automatic analysis tool to lead the investigators to dig gold efficiently. In addition, it is expected the tool will help people in other areas, such as to understand the social network behaviors better by detecting the abnormal patterns automatically [6].

Regarding text analysis, Hogenboom et al. [9] 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 modeling,

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.

3 Data

The data that this paper used for the case study is referred to the theoretic frame in [14]. The data sources consist of mixed types of both public and private data shown in Table 1. Public data means those data available for anybody, while private data is defined as those that an individual account holder can only access to. We categorized the data in this case study into three groups: time and location based data, static reference documents and short real-time stream texts from online social network services.

Table 1. Case study dataset.

Data type	Public data	Private data
Time and location		GPS, credit card records
Static documents	News articles	Emails, employee records
Stream text	Call center notice	Social network messages

4 Visual Analysis Tool Design

This section explains how we designed three subsystems of our visual analysis tool, which are specialized to handle each type of data identified in Table 1. The subsystems also work together interactively, so that they together can construct a total solution to reconstruct the target event wholly by combining piece of evidences found in different data sources.

Table 2. Subsystem design for different data types.

Data type	Analysis strategy	Tool
Time and location	– Location map – Pattern analysis	GPS analyzer (4.1)
Static document	– Natural language processing – Ontology analysis	Static document analyzer (4.2)
Stream text	– Keyword cloud – Timeline analysis	Stream text analyzer (4.3)

The target data type and analysis strategies of three subsystems are briefly summarized in Table 2. The first subsystem, GPS analyzer, is used to deal with the time and location-related GPS data and credit card transaction records using location map and

pattern analysis methods. The second subsystem, static document analyzer, reads a variety type of different static documents such as Excel and Word and extracts important keywords, relations and events from them. Lastly, the third subsystem, stream text analyzer, identifies the time-sensitive events for further investigations from the streaming data.

4.1 GPS Analyzer

From the given data set, we obtained two types of time and location-related data. One is GPS data from related persons' cars. It contains the date and time, latitude and longitude but does not have the location name. The other is credit card records for shopping and meals. This type of data contains only date, that is, no time is given. We linked up the credit card transaction records with the GPS data to identify locations. If GPS data discontinues for a reasonable amount of time, for example 15 min, we assume that the car is parked at some place for shopping, meals, work or sleep. People normally stay at their houses at night and the place where most people regularly go and stay during the daytime on weekdays is the company. Otherwise, the parking places can be possibly linked to those places with credit card records.

After finding the locations of the restaurants, shops, offices and houses, we analyzed the daily patterns of a person in two ways. First, we simplified the locations appeared in a person's daily life to three categories: restaurant, company and home (Fig. 1). In this way, we could identify a certain person's abnormal traffic pattern compared with usual daily patterns of others and narrow down our targets for further investigation. For example, if a person's location is found somewhere else in the middle of night, it is abnormal.

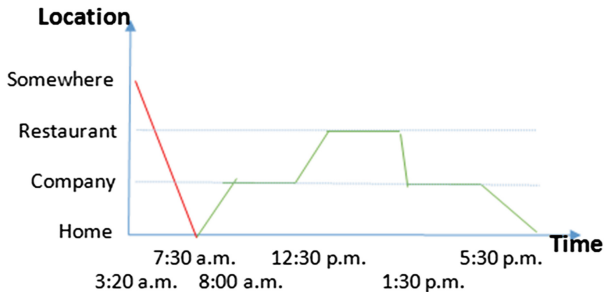

Fig. 1. Location classification and pattern analysis.

Second, we drew the latitude and longitude for all routes of every person in the target group with Python imaging tools. Every day all related person's movements are summarized in one page of map. This helped to find out who accompanied whom to which place. If two people have overlapped routes in the same time interval, they probably meet each other around the given time. Therefore, it is convenient to identify

multiple patterns for further investigation by checking the figures generated. Also, the consumption records at the place have been visualized in the view of consumer and/or location.

4.2 Static Document Analyzer

This module is designed to read static data sources written in common document formats such as Excel, Word, plain text or web contents. The static document analyzer parses the contents and creates an ontology model to construct the link between them. Ontology is a semantic analysis method to derive indirect relations among objects or individuals from facts [11]. The advantage of the ontology model is its flexibility to define object structure, so users can use their expert knowledge more easily to address important terms for a particular data set.

Most of ontology modeling in our tool was done in automatic processes. Besides, we implemented a review and editing interface for users to manually add or correct information. For example, structured documents such as employment records and emails can be handled automatically to analyze the contents. For semi-structured documents such as countries' fact sheets and resumes, which have clearly divided sections but some contents were written in a free-style text format, we applied two steps to complete the ontology modeling. After roughly creating the initial ontology model using the automation process, users are required to review the structure manually. We also had more free-style reports, which are unstructured documents. In this case, we decided to input the ontology model of the data manually for more accurate analysis.

Lastly, we summarized the news 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. To determine the importance of a keyword in a news report, we used Term-Frequency Inverse-Document-Frequency (TFIDF) 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) can give us more meaningful and relevant information than each individual word, we used Part of Speech (POS) tags and custom dictionary to register particular people, organizations and places that we want to trace after [12]. 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. In addition, 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 D3 bubble chart [13] window on the right hand side.

4.3 Stream Text Analyzer

In order to unearth an event and its timeline in twitter like blogs, two major issues need to be considered in the design and implementation of our tool.

One issue is how to abstract the key components of an event from the messages: who, how, where and when. First, we apply the natural language processing method as in the static document analyzer to preprocess every message to eliminate the stop-words, stem and tag the remained words. Second, we calculate the co-occurrence of any two words in every minute to aggregate their relationship. We then measure the importance of each word by counting the number of messages that contain it. This is because the more messages contain the word, the more attraction it receives. After aggregating the data, we put the result into the database for future visual analytics. Finally, we calculate the Levenshtein distance between any two messages, and remove the same and similar messages, to clean up the result for display.

The other issue is how to properly display the key components summarized from the previous processing. To solve this, we mainly use three charts: SUBJECT, ACTION and LOCATION. Specifically, we put time on x-axis and word count on y-axis; the word itself appears on the chart. Notably, all three charts are synchronized. To facilitate the better analytics, we set a marker to indicate the time that the user is currently viewing. Through all the word charts, we can easily tell the high frequent subjects, actions and locations, respectively. By dragging the scrollbar below the charts, we can view all the data that are already processed. The charts are updated periodically when newly aggregated data are available. In this way, we can see the timeline and progress of events. For those words which are overlapped, we can either zoom in the chart, or check the details of messages in the right panel. On the right hand side of charts, there is a panel to list up the messages around the current time being viewed.

5 Interactive Analysis Case Study

The subject of the case study is to investigate a crime incident. The Tethys-based gas company GAStech has made huge profits in the island country of Kronos. However, Kronos people suffered from pollution related to the GAStech commercial activities. On January 20th 2014, several employees of GAStech were missing after the company's management group annual meeting. Protectors of Kronos (POK), one local environment protection organization seemed to be the suspect [14]. This paper focuses on the interaction among three subsystems to understand the incident more comprehensively in the investigator's point of view.

As shown in Fig. 2, the strategy for the analysis starts with the personal data such as GPS and credit card records. In the studied case, by visualizing the GPS records and

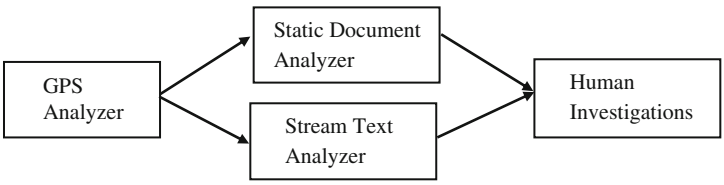

Fig. 2. Analysis flowchart.

consumption records in the GPS analyzer, people with abnormal patterns and the related locations are easily identified. After the abnormal pattern in personal data is determined, the abnormal list is then input in the static document analyzer and the stream text analyzer to find further suspects and related events. Additionally, the related public data offers more information as suspect candidate. Finally, the confirmed suspects and locations are supposed to be submitted to human investigators in this scenario as the clues for their further in-depth analysis.

5.1 Personal Data: Patten Identification

We start the analysis from those employees with abnormal traffic patterns since we quickly found some employees have went out at midnight using our GPS analyzer. For example, Isak Baza, Linnea Bergen, Nils Calixito, Axel Calzas, Lidelse Dedos Adra Nubarron, and Brand Tempestad had single late out at January 10 midnight. But, since logically several meetings are needed for planning the kidnapping event; we consider multiple late outs should be more important. In addition, although Lucas_Alcazar has multiple late outs, the GPS data shows he always stayed at the company, which indicates he is just working overtime. Therefore, the final suspect list has been narrowed down to Isia Vann, Loreto Bodrogi, Hennie Osvaldo and Minke Mies.

To confirm the employees with multiple going out in the midnight are really having meetings, a distance calculator program is also integrated in the GPS analyzer. The input is the employees and the time, and output is the corresponding distance between them. Limited by the discrete GPS data, the calculator can only find out the nearest location of the input time and calculate the distance between them. The calculator tells us that some employees did have meetings in the mid-night.

Accordingly, the places visited in the middle night frequently by the suspects are plotted in Fig. 3. The identified places are also matched to the tourist map given in the data set for further investigation.

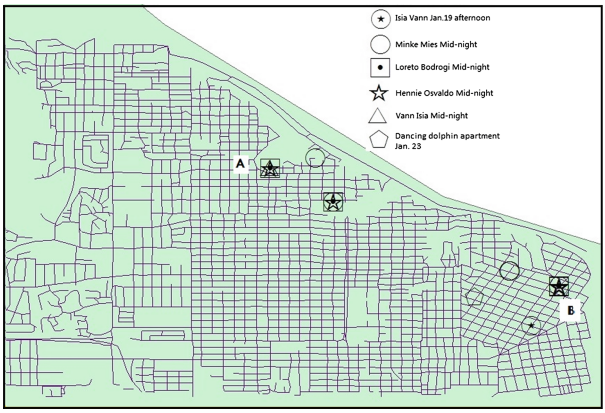

Fig. 3. GPS analyzer interface.

For the consumption pattern, we checked the difference of credit cards and loyalty cards per location. However, the difference is not evident. Therefore, we use the combined records only for simplicity. In addition, with the suspects found in traffic patterns, we studied the corresponding consumption patterns. For example, we see Isia Vann has an abnormal large bill at Frydos Autosupply n' More at January 17. But since he used a company car, the pattern does not clearly indicate if he is a suspect in the kidnapping. Finally, the consumption record charts by location may reveal people gathering information if we assume they are sharing the bill. But, again, it is hard to tell it is a normal office gathering or kidnapping plan meeting. Accordingly, we do not use consumption pattern analysis result for event identification input.

5.2 Personal Data to Public Data: Event Identification

The stream text analyzer is used to find the people's community and hot topics and events at this moment. As shown in Fig. 4, the chart of subject gives us some top frequent subjects: "POK Rally", "Sylvia Marek", "Abila City Park", "Prof. Stefano", "Lucio Jacab", "Victor-E" and "Dr. Newman". The verbs corresponding to the above subjects can be found in the chart of action. Together with using the word net in D3 bubble chart [13], the related actions are: "open", "begin", "introduce", "speak" and "play". From the chart of location, we can see the place of "Egeou St/Parla St". With studying the call centre record that contains the place (as listed on the right panel), we can confirm it is the event location.

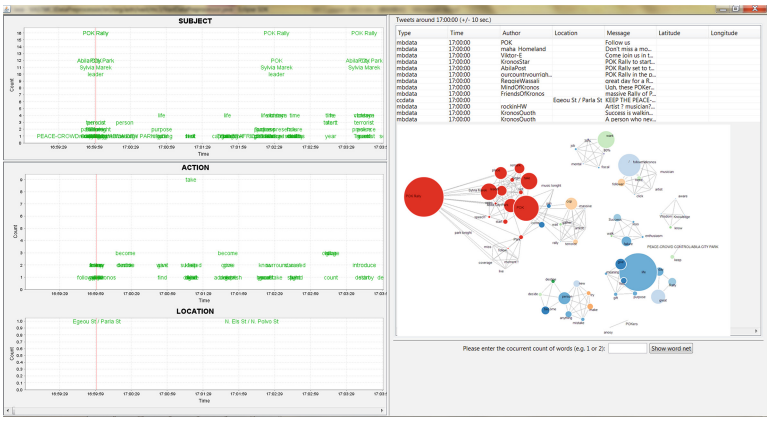

Fig. 4. Stream text analyzer interface.

In this way, we can abstract an important event from the raw data: POK Rally. It started at 17:00:00 and ended around 19:05:54. The progress of the event can be described as follows: the leader of POK, i.e. Sylvia Marek, gave an opening speak to start the rally in Abila City Park; a special guest Dr. Newman was invited to give a talk with Prof. Stefano and Lucio Jacab; the band Victor-E was invited to play music.

Similarly, we abstracted three other important events: the fire at the Dancing Dolphin Apartment, the vehicle accident between a black van and a bicycle, and the gunfire and subsequent events after the black van was stuck at the Gelato Galore.

With the location input from the GPS analyzer, we think the fire at the Dancing Dolphin Apartment is most important to the GAStech disappearances. In the stream text analyzer, we find a witness "dangermice" reporting the whole progress of fire. This people's micro blog data include the longitude and latitude. With such information, we can check the fire location and see that the location of vehicle accident and the fire are both closed to B in Fig. 3.

For the static document analyzer, Isia Vann is our start point since he is considered a strong suspect in 5.1. Figure 5 shows all personal profile, news articles, people and organizations, which are identified as having direct or indirect connections with Isia Vann during the ontology and keyword analysis processes. Accordingly, Isia Vann's email traffic is checked. The suspects list is then extended to the receipts of Email circle "RE: FW: ARISE - Inspiration for Defenders of Kronos": Isia Vann, Inga Ferro, Loreto Bodrogi, Hennie Osvaldo, Minke Mies. In addition, the abnormal multiple middle night traffics check in the GPS analyzer returns a list includes Isia Vann, Minke Mies, Loreto Bodrogi and Hennie Osvaldo.

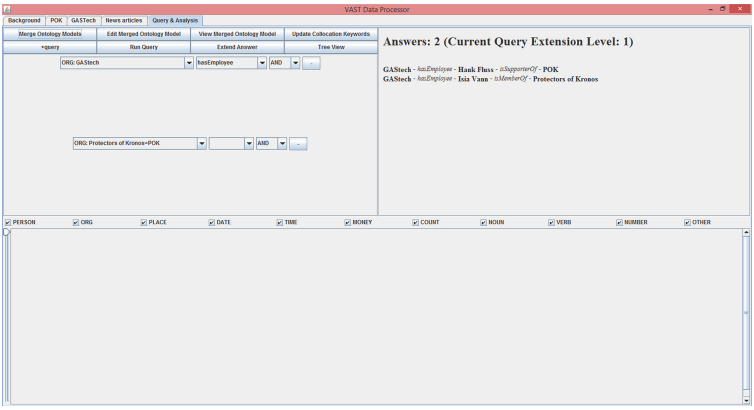

Fig. 5. Static document analyzer interface.

5.3 In-Depth Analysis Suggestions

The final step of the case study is to offer some suggestions to police officers, who will investigate this case further based on our findings.

First of all, we think the disappeared GAStech employees include executive group of GAStech and the kidnapper suspects. The kidnapper suspects include Isia Vann, Inga Ferro, Loreto Bodrogi, Hennie Osvaldo, Minke Mies, as indicated in 5.2. Therefore, we suggest that the police trace down whereabouts of those suspects.

Secondly, we roughly reconstructed the timeline of the kidnapping incidents as follows. The employee disappearance happened at January 20, around 10:00 am.

Before that, the executive group was having the annual meeting at GAStech building starting from 9:00 am. After 10:00 am, the suspects made a call to claim a false fire alarm, and took the victims away in the evacuation. Probably, the executive group and some kidnappers left Abila by private planes in the afternoon of January 20. The suspects may plan the actions several time at the locations listed in Fig. 3 during January 7–January 20. Point A and Point B in Fig. 3 are the most visited places by the suspects. And, at January 23, some of the suspects may try to set a real fire at dolphin apartment (point B in Fig. 3) to remove the evidence. After that, the suspects flee caused a traffic accident, as well as a subsequent gunfire with local police. Therefore, we recommend point A and point B in Fig. 3 for further investigation.

6 Summary

To provide a consolidated single analysis tool, which integrates various data sources to help reduce manual effort of human data analysts in this big data era, we proposed an interactive visual analysis tool, which is composed of three subsystems that takes part in analyzing a particular type of data. Our tool could successfully identify the abnormal patterns in the personal data automatically. Accordingly, the suspect people, location, and events found in the public data set are listed for later investigation efficiently.

Currently, the interaction between subsystems has to be done manually. The future work may include the automatic interaction design, which means the subsystem should work more "smart" to filter the unrelated results. In addition, in a more practical case, both of the public data and the personal data may include more "noise". Therefore, a pre-process on the data for a specific case/scenario should also be considered.

References

1. Laney, D.: 3-D data management: controlling data volume, velocity and variety. META Group Research note, 6 February 2001
2. Thomas, J., Cook, K.: A visual analytics agenda. IEEE Comput. Graphics Appl. **26**, 10–13 (2006)
3. Keim, D., et al.: Visual analytics: how much visualization and how much analytics? ACM SIGKDD Explor. Newsl. **11**(2), 5–8 (2009)
4. Dou, W., et al.: LeadLine: interactive visual analysis of text data through event, identification and exploration. In: IEEE Conference on Visual Analytics Science and Technology 2012, Seattle, WA, 14–19 October 2012
5. Wanner, F., et al.: State-of-the-art report of visual analysis for event detection in text data streams. http://bib.dbvis.de/uploadedFiles/3_submission.pdf
6. Slingsby, A., et al.: Visual analysis of social networks in space and time. Mobile Data Challenge Workshop 2012, Newcastle (2012)
7. Criminal analysis: new prospects for investigation using i2 software. https://visualanalysis.com/Downloads/CaseStudies/ANB-IBASEOCRVP_UK_Q2%202011_ICP_Low.pdf
8. Top law enforcement software tools. http://www.capterra.com/law-enforcement-software/

9. Hogenboom, F., et al.: An overview of event extraction from text. In: van Erp, M., et al. (eds.) Proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web, pp. 48–57, Bonn (2011)
10. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise data analysis and visualization: an interview study. IEEE Trans. Visual Comput. Graphics **18**(12), 2917–2926 (2012)
11. Kietz, J.U., Maedche, A., Volz, R.: A method for semi-automatic ontology acquisition from a corporate intranet. In: EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins (2000)
12. Arazy, O., Woo, C.: Enhancing information retrieval through statistical natural language processing: a study of collocation indexing. MIS Q. **31**(3), 525–546 (2007)
13. Data Driven Documents. http://d3js.org/
14. VAST Challenge (2014). http://www.vacommunity.org/VAST+Challenge+2014