VACI: Towards Visual Analytics for Criminal Investigation

Rahul Kamal Bhaskar, Julia Paredes, Zahra Shakeri, Zahra Sahaf, Haleh Alemasoom
Craig Anslow, Frank Maurer, Mario Costa Sousa, Faramarz Samavati
Department of Computer Science, University of Calgary
Email: {rbhaskar,jparedes}@ucalgary.ca

Abstract

This paper presents our approach for solving the IEEE VAST 2014 Mini Challenge 1. To solve this challenge we performed different tasks related to data processing and utilized different visualization techniques to explore the data. In this paper we present our application VACI: Visual Analytics for Criminal Investigation and discuss the design and the features.

Keywords: Visual analytics, information visualization, criminal investigation

Index Terms: H.5.2 [Information Interfaces & Presentations]: User Interfaces – Evaluation/methodology

1 Introduction

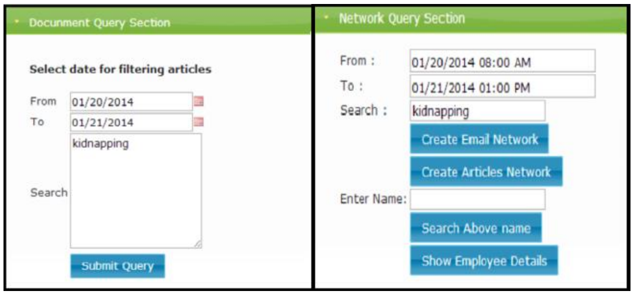
The VAST challenge is a fictitious data set used for cyber-security analysis and is comprised of four challenges. Mini Challenge 1 relates to criminal activities that happened in Abila. The aim of the challenge is to understand the events about kidnapping of several GASTech employees and to find the potential people involved. To address this challenge we initially explored the problem and datasets using various tools including RapidMiner and Tableau. After this exploration stage we concluded that these tools did not provide appropriate features for parsing files, performing analysis, and visualizing the data to solve the questions in the Challenge. Instead we developed our own application VACI: Visual Analytics for Criminal Investigation to find solutions to the problems presented.

2 VACI: Visual analytics for criminal Investigation

VACI is a visual analytics application developed to solve Mini Challenge 1. The data for this challenge comes in the form of newspaper documents and other related data sets to support the documents. VACI is comprised of two independent dashboards: Document Analysis Dashboard and Network Analysis Dashboard (Figures 2 and 3). The Document Analysis Dashboard is used for solving questions related to text search like finding event, time, dates and reasons for the incident. The Network Analysis Dashboard is used for solving questions related to connections between different people working in GASTech and POK.

The dashboards contain different visualizations, which help security analysts to extract insight from the data. While designing our dashboard we were concerned that all the components of the analysis must appear on the same screen as changing or scrolling the screen can make security analyst forgot about previous data.

VACI is a web based visual analytics application implemented with different visualization toolkits including D3.js and Highchart.js, and Stanford NLP for text mining [1,2,3].



Select date for filtering articles
From 01/20/2014
To 01/21/2014
Search kidnapping

From : 01/20/2014 08:00 AM
To : 01/21/2014 01:00 PM
Search : kidnapping
Enter Name:

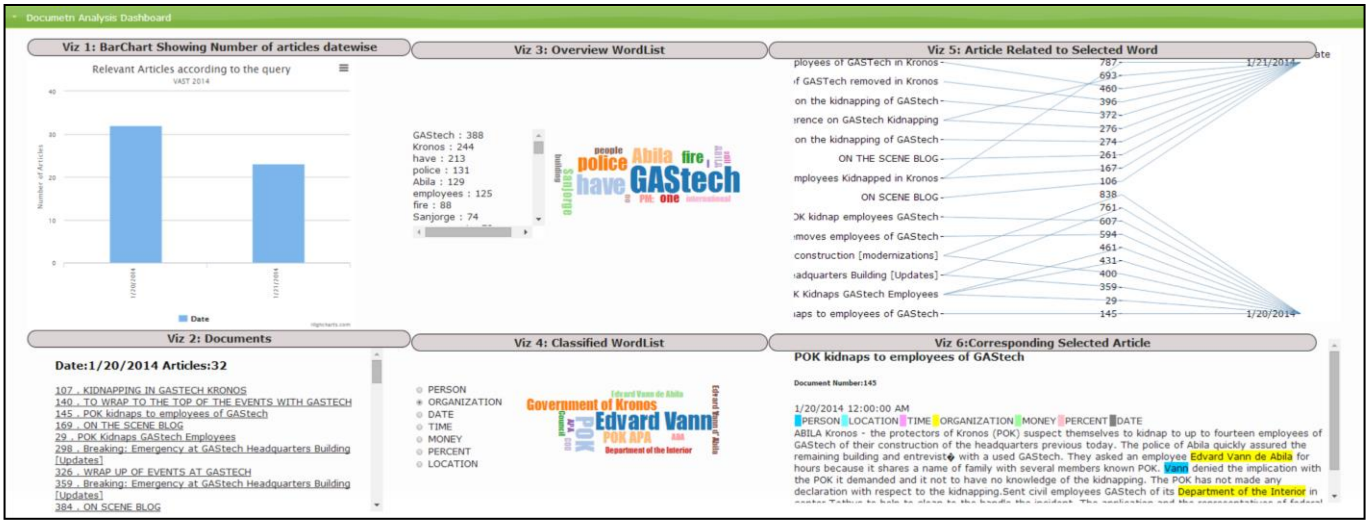
Figure 1: Search Section

2.1 Search

To begin solving the problems in the challenge an analyst must first perform a search. There are two separate search sections in VACI for each dashboard: Document Query Section and Network Query Section (Figure 1). The search features help analysts to search and filter data in the visualizations in the dashboards. Suppose in the Document Analysis Dashboard an analyst wants to search documents that contain events related to the kidnapping during 20 and 21 January 2014. To perform this task the analyst will go to the Document Query Section and enter the date range in the “To” and “From” fields. The analyst can then add key words to search for documents that contain words such as “kidnapping” by entering the words into the “Search” field. The results of the search query populate the data in the visualizations within the dashboard. The Network Query Section also has search features for analyzing data on the Network Analysis Dashboard.

2.2 Document Analysis Dashboard

The Document Analysis Dashboard shows visualizations related to text mining of documents. The visualizations are populated once an analyst submits a search query. Following the question from the Search section an analyst wants to get insight about the word “kidnapping”. The first four visualizations in the Document Analysis Dashboard (Figure 2) show group analysis on the documents. A bar chart shows the number of documents date wise where “kidnapping” word has appeared more (Viz1). The document section shows all documents in the selected date range (Viz2). A word cloud shows the most frequently used words in documents with the “kidnapping” keyword (Viz3). Another word cloud shows the most frequently used words in a categorized way (i.e. name of person, organization, money, location, time and date) (Viz4). If an analyst found any suspicious words they can then apply filters to adjust the visualization to provide clues for solving the crime. For exploring detail about a specific word an analyst can select the word from either of the word clouds (Viz2 and Viz3) and display a parallel coordinates visualization (Viz5). The parallel coordinate shows all documents that contain that word. The purpose of this visualization is to group documents on the basis of different aspects (i.e. date and headlines). Suppose there is multiple documents with the same headlines then exploring a


Documetn Analysis Dashboard
Viz 1: BarChart Showing Number of articles datewise
Relevant Articles according to the query
VAST 2014
Number of Articles
40
30
20
10
0
1/20/2014
1/21/2014
Date
Highcharts.com
Viz 3: Overview WordList
people
police
Abila
fire
have
GAStech
$anjorge
building
PM
one
international
Viz 5: Article Related to Selected Word
ate
ployees of GASTech in Kronos
f GASTech removed in Kronos
on the kidnapping of GAStech
erence on GAStech Kidnapping
on the kidnapping of GAStech
ON THE SCENE BLOG
mployees Kidnapped in Kronos
ON SCENE BLOG
OK kidnap employees GAStech
moves employees of GAStech
construction [modernizations]
adquarters Building [Updates]
K Kidnaps GAStech Employees
aps to employees of GAStech
787
693
460
396
372
276
274
261
167
106
838
761
607
594
461
431
400
359
29
145
1/21/2014
1/20/2014
Viz 2: Documents
Viz 4: Classified WordList
○ PERSON
● ORGANIZATION
○ DATE
○ TIME
○ MONEY
○ PERCENT
○ LOCATION
Government of Kronos
Edvard Vann de Abila
Edvard Vann
POK
POK APA
APA
Council
COG
AIM
Department of the Interior
Edvard Vann d' Abila
Viz 6:Corresponding Selected Article

GAStech : 388
Kronos : 244
have : 213
police : 131
Abila : 129
employees : 125
fire : 88
Sanjorge : 74

Date:1/20/2014 Articles:32

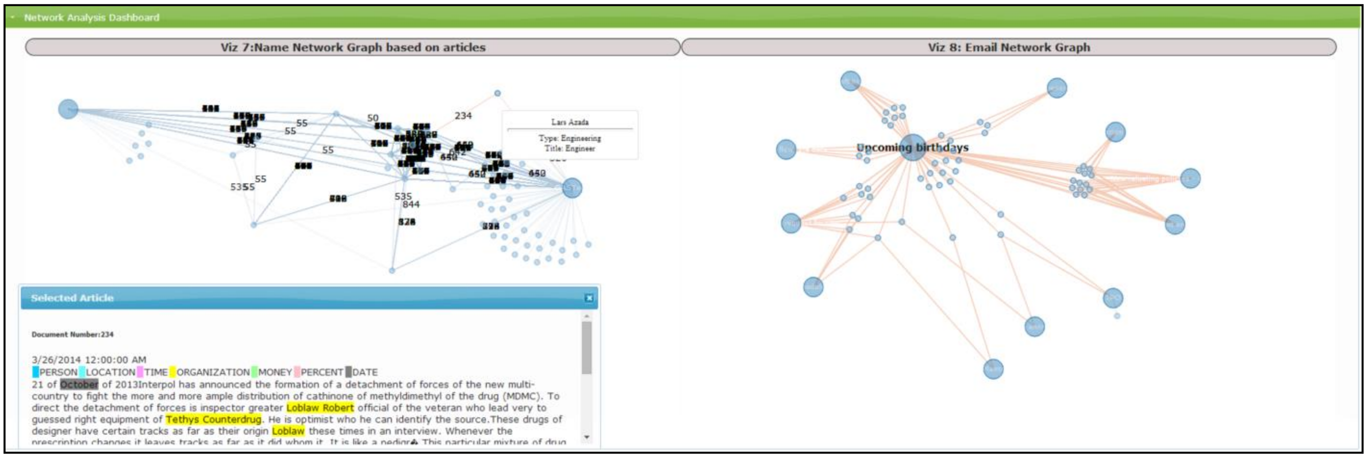
107 . KIDNAPPING IN GASTECH KRONOS
140 . TO WRAP TO THE TOP OF THE EVENTS WITH GASTECH
145 . POK kidnaps to employees of GAStech
169 . ON THE SCENE BLOG
29 . POK Kidnaps GAStech Employees
298 . Breaking: Emergency at GAStech Headquarters Building [Updates]
326 . WRAP UP OF EVENTS AT GASTECH
359 . Breaking: Emergency at GAStech Headquarters Building [Updates]
384 . ON SCENE BLOG

POK kidnaps to employees of GAStech

Document Number:145

1/20/2014 12:00:00 AM
PERSON LOCATION TIME ORGANIZATION MONEY PERCENT DATE
ABILA Kronos - the protectors of Kronos (POK) suspect themselves to kidnap to up to fourteen employees of GAStech of their construction of the headquarters previous today. The police of Abila quickly assured the remaining building and entrevist� with a used GAStech. They asked an employee Edvard Vann de Abila for hours because it shares a name of family with several members known POK. Vann denied the implication with the POK it demanded and it not to have no knowledge of the kidnapping. The POK has not made any declaration with respect to the kidnapping.Sent civil employees GAStech of its Department of the Interior in

Figure 2: Document Analysis Dashboard



Document Number:234

3/26/2014 12:00:00 AM
PERSON LOCATION TIME ORGANIZATION MONEY PERCENT DATE
21 of October of 2013Interpol has announced the formation of a detachment of forces of the new multi-country to fight the more and more ample distribution of cathinone of methyldimethyl of the drug (MDMC). To direct the detachment of forces is inspector greater Loblaw Robert official of the veteran who lead very to guessed right equipment of Tethys Counterdrug. He is optimist who he can identify the source.These drugs of designer have certain tracks as far as their origin Loblaw these times in an interview. Whenever the prescription changes it leaves tracks as far as it did whom it. It is like a pedigr� This particular mixture of drug

Figure 3: Network Analysis Dashboard

single document with the same headline will give details on demand about a document with this headline. Finally an analyst is provided with highlighted documents on the dashboard for the selected documents in the parallel coordinates visualization (Viz6). Documents are highlighted to help understanding on the basis of the type of words like name of person, organisation, money, location, time and date.

Visualizations on this dashboard are synchronized together so changes in one visualization update the data in other visualizations and allows an analyst to perform stepwise exploration of the data. Initially the dashboard has no visualizations on display but once an analyst submits a search query Viz1-4 are displayed and updated on subsequent queries. Furthermore in dashboard Viz1 is synchronized with Viz2-4. If an analyst selects a bar in Viz1 representing a particular date, the result of selection updates and shows data for that particular date in Viz2-4. Viz3 and Viz4 are synchronized with Viz5. When an analyst selects a word from the word cloud (i.e. Viz3 and Viz4), Viz5 then shows the document related to the selected word. Finally, Viz6 is updated after selections of data from Viz5.

2.3 Network Analysis Dashboard

The Network Analysis Dashboard (Figure 3) focuses on tasks related to the analysis and finding potential connections between GASTech employees and POK members. To accomplish these tasks we implemented network graphs (Viz7 and Viz8). Visualizations on this dashboard are displayed after submitting a search query from the Network Query Section. Both visualizations are independent.

Viz7 represents the relationships extracted from the documents as a graph where the nodes represent names of people from GASTech and POK. The edges between the nodes show that both names appeared in the same document. Edges between nodes are labeled with the document number in common. The document number was provided with the purpose so that an analyst can click on the number and examine the content of the document and explore details related to potential connections.

Viz8 is a graph used to visualize suspicious emails and people who are related to that email. The larger nodes in the graph are emails and smaller nodes are people who are related to that email. Selecting the node displays the email. An edge between the nodes represents an email ID and a node can represent either the sender or receiver of the email. Once a suspicious email and group of people have been identified an analyst can then further explore the names of people in the Document Analysis Dashboard.

REFERENCES

[1] Stanford NLP - http://nlp.stanford.edu/software/
[2] D3 - http://d3js.org/
[3] Highcharts - http://www.highcharts.com/