Middlesex University's Invisque Visual Analytics Tool: Supported by Text Analytics Techniques from the University of Leeds

Sharmin (Tinni) Choudhury[1] Claire Brierley[2] Chris Rooney[1] Kai Xu1 Raymond Chen[1] William Wong[1] Eric Atwell[2]
[1]Middlesex University [2]University of Leeds

ABSTRACT

This report describes the joint entry from Middlesex University and the University of Leeds for Mini Challenge 3 for the VAST Challenge 2011. In order to address the challenge question, the primary tool we used was Middlesex University's Interactive Visual Search and Query Environment (INVISQUE), which served as the user interface to the Mini-Challenge 3 news corpus. INVISQUE was supported by corpus text analytics from the University of Leeds, which provided additional information that was visualised on the INVISQUE user interface.

KEYWORDS: Visual Analytics, Information Visualisation, Text Analysis, Human Factors.

INDEX TERMS: H.5.2 [User Interfaces]: Graphical user interfaces – Information Visualization; H.3 [Information Storage and Retrieval]: Information Search and Retrieval - analytics

1 INTRODUCTION

For the IEEE VAST 2011 Mini-Challenge 3 (MC 3) we were asked to investigate a text corpus consisting of 4474 news articles and tasked to identify any imminent terrorist threats in the Vastopolis metropolitan area and to provide detailed information on the threat or threats (e.g. who, what, where, when, and how) so that officials could conduct counterintelligence activities.

The primary tool used to address the mini challenge task was Middlesex University's Interactive Visual Search and Query Environment (INVISQUE) [1 2]. INVISUQE provided the visual search, query and analysis environment into which the MC 3 news article corpus was visualized and provided an environment for analysis of the corpus. INVISQUE was further supported by text analytics provided by University of Leeds [3 4]. In this paper, we present our answer to MC 3 and also explain how we undertook our analysis.

2 FINDING FROM NEWS CORPUS

There were many interesting activities in Vastopolis but most could not be considered "imminent terrorist" threats. However, what may pose an imminent threat involves stolen equipment from the labs of molecular biologist Professor Edward Patino. Prof. Patino has been harassed by the group Citizens for Ethical Treatment of Lab Mice, who in-turn are affiliated with the Forever Brotherhood of Antarctica.

The Professor himself has recently given lectures on the threat of bioterrorism; in addition, the Center for Disease Control (CDC) also released a recent report highlighting the dangers of bioterrorism. Since the robbery of the professor's lab, the Brotherhood and the Citizens for Ethical Treatment of Lab Mice have shown an increased level of activity. Lastly, dead fish have turned-up in Vast River.

Therefore, the conclusion of our analysis was that there may be an imminent threat to Vastopolis metropolitan area from Forever Brotherhood of Antarctica and their affiliates, the Citizens for Ethical Treatment of Lab Mice involving some form of biological weapon created from the equipment stolen from Professor Patino's lab. The list and timeline of the articles supporting our hypothesis is given in Table 1.

Table 1. Timeline of News Articles

Article Date	Event
11-04-2011	Prof. Patino gives lecture on bioterrorism
18-04-2011	CDC releases publication on threats of bioterrorism
26-04-2011	Prof Patino's lab gets robbed
02-05-2011	Mayor's dog gets kidnapped
03-05-2011	Basketball teams mascot goes missing from Vastopolis Dome
09-05-2011	Citizens for Ethical Treatments of Lab Mice send threatening emails to Vast Press
19-05-2011	Dead fish is found is Vast River

The other events in Vastopolis, which were discounted as either being resolved or self-contained, include,

1. Military weapons went missing from Vastopolis Armed Forces on the 26-04-2011 and on the 30-04-2011, military grade weapons were used in a park shootout in Southville. However, the weapons were recovered at the Vastopolis airport on the 20-05-2011.
2. Two mental patients affiliated with the psychobrotherhood escaped the Vastopolis Center for the Criminally Insane on 27-04-2011 but were caught on the 12-05-2011 while trying to make a bomb. No further information was available in the corpus for psychobrotherhood.
3. An Antarctica Airlines plane crashed and traces of explosives were found in the wreckage but this is a past event. In addition, while there were articles about bad security at the Airport, following the crash security was increased.
4. A 60 year old man built an improvised explosive device to kill his neighbor's cat KeeKee, but the incident was resolved.
5. A man with a bomb concealed in a turkey was stopped at Vastopolis Airport but that news article did not provide any hooks for further investigation.
6. The daughter of a military counter-intelligence agent was raped by another soldier and her identity exposed but the article provided no course to follow.
7. F-Alliance a group of Hackers comprised of high-school drop-outs were arrested, thus another resolved issue.
8. Anarchists for Freedom issue daily threats to Vastopolis Officials but they are yet to act.

t.choudhury@mdx.ac.uk, scscb@leeds.ac.uk, c.rooney@mdx.ac.uk, k.xu@mdx.ac.uk, r.chen@mdx.ac.uk, w.wong@mdx.ac.uk, e.s.atwell@leeds.ac.uk

9. Lastly, Vastopolis was included in general threat issued by the overseas terror group Network of Dread.

3 Analytics Process

We used the INteractive VIsual Search and QUery Environment (INVISQUE), a prototype visual analytics interface created at Middlesex University, to visually sift through the news corpus. INVISQUE uses index-card visualization to represent individual information items, in this case the news articles, and arranges them on screen on an X-Y axis. Figure 1 shows the search results from the keyword search "bomb" arranged on the X axis by significance and on the Y axis by date - so that news articles with higher level of significance for the keyword "bomb" appears more to the left and newer articles appear higher up the Y axis.

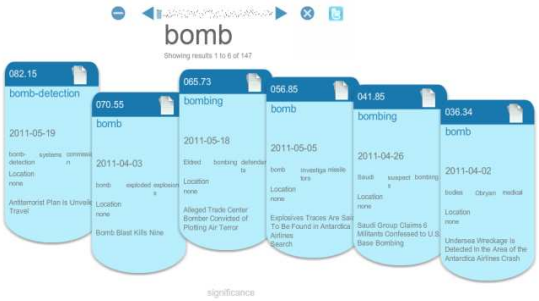

Figure 1. INVISQUE index-card visualization arranged on X-Y axis

The "significance" value was calculated by our collaborators from the University of Leeds who performed keyword extraction on the news corpus. Keyword Extraction is a standard Corpus Linguistics technique for genre classification which pinpoints statistically significant or "key" words for that genre via comparison with a general reference corpus [3 4]. The significance calculation for the MC 3 corpus entailed comparison of word frequency distributions in each of the 4474 news article test sets with their distribution in the entire news article dataset as reference corpus. The Leeds program verifies apparent overuse of lexical items in each article by computing the difference between these observed frequencies and the norm as represented by their expected frequency in the whole dataset, expressed as a log likelihood (LL) statistic. Words with LL scores of 6.63 or above are statistically significant.

Single word searches, e.g. bioterrorism, on the INVISQUE interface is applied against the word list generated by the University of Leeds, see Figure 2, and leads to generation of index-cards that have the matching keyword as the title and the significance of the keyword as the top-left value.

id	word	rank	key_column
2317	and	10.06	1
2317	in	7.07	2
2317	said	17.02	3
2317	Iraq	52.08	4
2317	Minister	47.82	5
2317	Iraqi	43.4	6
2317	government	20.5	7
2317	said.	16.92	8
2317	Foreign	35.45	9

Figure 2. Keyword Extraction Word List

Composite phrases, e.g. "Vast River", are applied against the full article text and in the latter case – the title of the index card becomes the most significant keyword, as calculated by Leeds, in the article. These are illustrated in Figure 3.

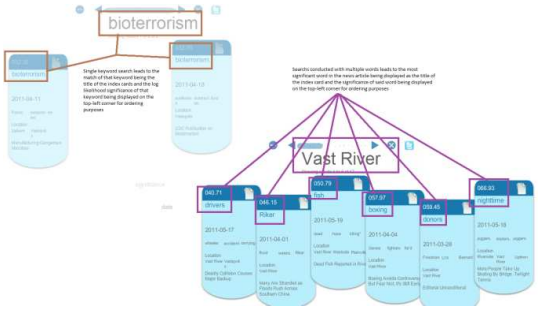

Figure 3. Searching Using INVISQUE

As shown in Figure 4, the cards also show a "gist" of the article by displaying the top three most significant keywords of the article, the article title, Vastopolis locations mentioned in the article, which are extracted and appended to the cards by the middleware based on a pre-compiled list, and the date of the article. The cluster of returned cards can be filtered by any of the card features and the cards also have a shortcut to the full text of the article.

Figure 4. Index Card Fields

Once the MC 3 news corpus was visualized through INVISQUE, the primary technique used to explore the corpus provided was visual searching and filtering. This technique allowed us to explore the corpus very thoroughly, very quickly and we began to get a picture of the happenings in Vastopolis within hours of beginning our exploration.

4 Conclusion

We believe that the threat being faced by Vastopolis is that of bioterrorism and have identified this threat with our tool INVISQUE with supporting text analytics techniques.

References

[1] W. Wong, C. Rooney, R. Chen, K. Xu, and N. Kodagoda, "INVISQUE□: Intuitive Information Exploration through Interactive Visualization," in *ACM CHI Conference on Human Factors in Computing Systems*, 2011.

[2] I. D. C. M. U. Invisque Team, "Invisque," 2011. [Online]. Available: http://www.invisque.com/.

[3] E. Atwell, S. Sharoff, and L. Al-Sulaiti, "Arabic and Arab English in the Arab World," in *Proceedings of CL2009 International Conference on Corpus Linguistics*, 2009.

[4] B. A. Shawa and E. Atwell, "A chatbot system as a tool to animate a corpus," *ICAME International Computer Archive of Modern and Medieval English Journal*, vol. 29, pp. 5-24, 2005.