VisWorks Text and Network Visual Analytics

VAST 2010 Mini Challenge 1 Award:"Effective Interactive Visualization of Document Contents"

Lei Shi Weihong Qian Furu Wei Li Tan *

ABSTRACT

VisWorks is a software package for text and network visual analytics. This paper introduces its visualization, analytic process and lesson learned in solving Mini-Challenge 1 of VAST 2010 contest.

1 INTRODUCTION

Analytics over data mixed up by both unstructured and structured information poses a great challenge to the community. The usual pain points include the difficulty to select the appropriate mining/learning algorithms, the time/labor cost to customize analytic process and the lack of methodology to synthesize the insights found at multiple facets. To the end, there is rigid and urgent demand for an integrated visual analytics tool that could combine the human intelligence and machine capability to tackle the above challenges and fasten the analytic tasks.

VisWorks is a software package for text and network visual analytics. It tightly combines state-of-the-art analytics algorithms with interactive visualization for iterative and progressive human analytics. The visualization designs in VisWorks carefully trade-off ease of use, effectiveness for analytics and appeals of the user interface. The package also provides support to common data formats and various interactions for visual exploration and reasoning, hence offers user great flexibility in the analytic tasks.

2 VISWORKS VISUALIZATION

Figure 1 shows an overview of Mini-Challenge 1 data by TIARA visualization [2]. X axis represents document time and Y axis represents document number evolved over time. The vertical layers show topics/categories extracted from text collection and here correspond to different countries. Keywords in each layer are time-sensitive named entitles extracted from documents of specific country. Keyword size indicates its occurrence count (aka hotness) while the keyword color indicates its entity facet type, such as people, place and activity. The facet navigation panel in the left offers control to the entity facets and the legend panel in the main view allow user to select the layers drawn in the visualization.

Figure 3(a) gives a example of VIGOR network visualization. It provides the basic node/edge rendering function that encodes multiple information facets on the node/edge size/thickness/color/shape. Some network analytics algorithms including role analysis and clustering are also embedded in the package. Additionally, we provide advanced filtering functions over both node and edge in the network, based on the information facets encoded.

3 ANALYTIC PROCESS

3.1 Data Pre-Processing

The pre-processing starts by segmenting the raw contest data into text snippets according to title line of each report/message/post inside the aggregated document. Both the time and body of each snippet could be obtained through regular expression matching. Named entitles (NE) are extracted with Stanford NE Parser [1], including people, place and organization. In a further step, places are mapped to countries and activities are extracted from verbs with POS parser.

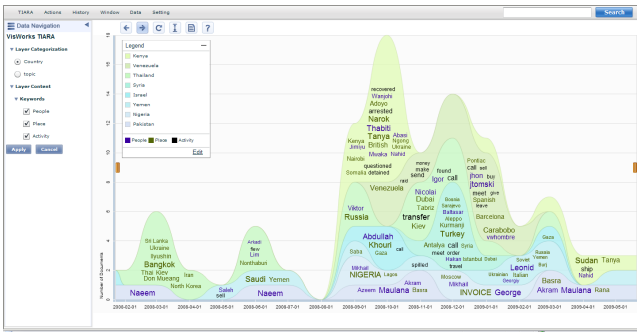


Figure 1: Visual summary of document content by country.

To develop an effective snippet classification, we conduct a clustering over the country co-occurrence graph where node weight is mapped to country occurrence count by snippets and edge weight indicates country co-occurrence within the same snippet. By filtering out nodes and edges with smaller weights and removing some bridging nodes (which should be the intermediate countries), we finally find out eight country groups. They are used as snippet category in the following visualization.

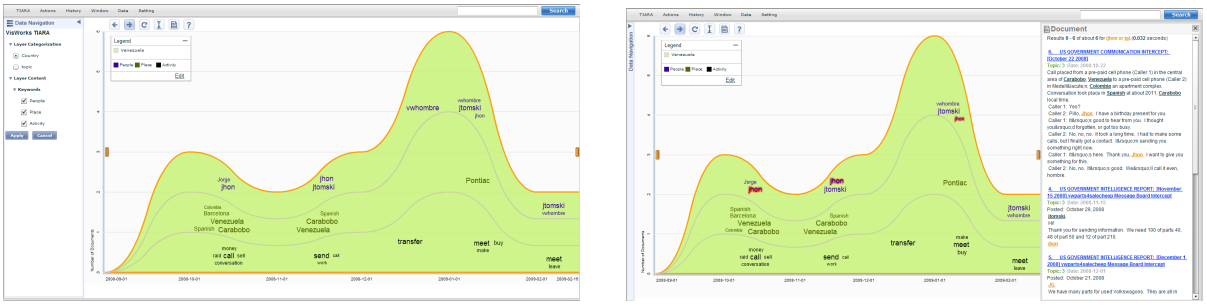
3.2 Visual Analytics

For the first task of Mini-Challenge 1, we compose an overview picture to summarize the activities happened in each country group using TIARA visualization, as shown in Figure 1. As required by the task, we drill-down to the content of each country group and derive insights by navigating the data with our tool. Here we only explain the analytic process over the arms dealing activities of one country group - Venezuela. The analyses of other countries are of the same methodology.

We select the Venezuela layer in TIARA and choose the time period containing snippets of this country group to proceed. The content summarization of Venezuela then shows up, as illustrated in Figure 2(a). The three layers in the view indicate the keywords of different category: people, place and activity. In the activity layer, an evolving trend of call-send-transfer(money)-meet reveals. We hypothesis that the local Venezuela buyer first call to discuss with the dealer, then send something (maybe the arms list), transfer the money and finally arrange to meet. We further locate the key people in these arms dealing events. It is noticed that jhon appears throughout the timeline, so we click on this keyword and retrieve all the snippets containing jhon, as shown in Figure 2(b). After reading these snippets arranged by time, we find out that jhon is an intermediate dealer connecting Venezuela buyer to jtomski. To help summarize the situation by the end of the period, we drill-down to the most recent time and click on the layer to let TIARA extract the top five important sentences. With these output, we conclude that both jhon (the intermediate) and Vwhombre (probably the local buyer) will meet jt (the arms dealer) in UAE on late April, 2009.

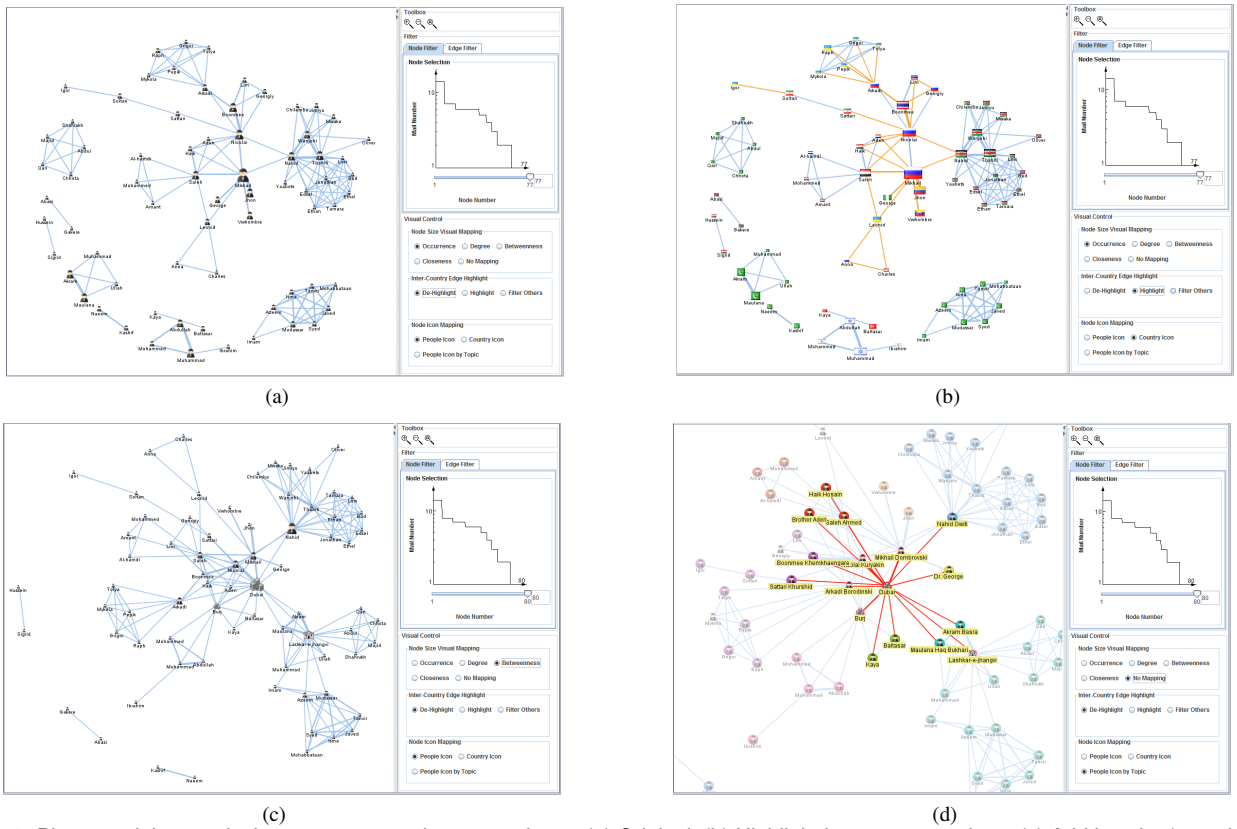
For the second task, we start from the player social network in Figure 3(a). It is found that the player network is composed of several connected components, where Nicolai and Mikhail connect the largest component together. To gain insight of country distribution

*Lei Shi, Weihong Qian, Furu Wei and Li Tan are with IBM Research - China, e-mail:{shllsh,qianwh,weifr,tltan}@cn.ibm.com



(a) (b)

Figure 2: Drilling to content of single country and the snippets containing a keyword.



(a) (b)

(c) (d)

Figure 3: Player social networks by co-occurrence in same snippet: (a) Original; (b) Highlight inter-country edges; (c) Add location/organization nodes; (d) Color-coding by country group category.

over the player network, we encode national flag into each node to indicate his/her country and highlight the inter-country connections by a different orange color. The results in Figure 3(b) demonstrate that Nicolai,Mikhail, Arkadi and Boonmee act as intermediate players for liaison among countries. This could also be detected by measuring the betweenness centrality score of these players. In a next step, we incorporate locations and organizations into the co-occurrence graph. It is interesting to notice that the whole graph snaps together as almost one connected component, shown in Figure 3(c). Dubai acts as the primary bridge. To further understand the situation, we choose to map the country group category into contour color of each node icon. As given in Figure 3(d), Dubai actually connects to all the eight country groups either directly or through one intermediate potential dealer.

4 Lesson Learned and Future Direction

In this visual analytics practice, we identify several shortcomings in the current tool and process, which point out potential directions for our future works: First, the data pre-processing and the visual analytics are hardly glued together. Whenever a issue is found in the visualization related to the data, the analyst should go back to correct the pre-processing and then restart the analysis. Built-in visual editing should be able to reduce this iteration cost by modifying the processed data on-the-fly. Second, in this practice, we frequently switch between the text and network visualization to synthesize the information gained in each approach. However, view change considerably destroys our mental map to the data and degrade the task performance. Visualization mashups, as mentioned by the reviewer, should help to improve the synthesis tasks. Third, throughout the task analysis, we rely on manual recording of the patterns discovered and insight gained. It will be quite useful if the visual analytics tools could introduce some kind of reporting function which allows user to input discoveries and retrieve the summary in the end.

References

[1] Stanford NE Parser, http://nlp.stanford.edu/software/crf-ner.shtml.

[2] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *VAST '10*, 2010.