

Jigsaw meets Blue Iguanodon - The VAST 2007 Contest

Carsten Görg* Zhicheng Liu† Neel Parekh‡ Kanupriya Singhal§ John Stasko¶

School of Interactive Computing & GVU Center
Georgia Institute of Technology

ABSTRACT

This article describes our use of the *Jigsaw* system in working on the VAST 2007 Contest. *Jigsaw* provides multiple views of a document collection and the individual entities within those documents, with a particular focus on exposing connections between entities. We describe how we refined the identified entities in order to better facilitate *Jigsaw*'s use and how the different views helped us to uncover key parts of the underlying plot.

Keywords: Visual analytics, investigative analysis, intelligence analysis, information visualization, multiple views

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 INTRODUCTION

We worked on the VAST 2007 Contest using the *Jigsaw* system that we have been developing within the Southeastern RVAC. *Jigsaw* is implemented in Java and provides multiple views of the documents in a collection as well as the entities within those documents. Its specific focus is to illuminate connections between entities across the documents. We refer the reader to a regular paper [2] about *Jigsaw* in the VAST'07 proceedings and the project website [1] for details about the system and its views. This article focuses on the process we followed in working on the contest and the changes made to the system based on what we learned in that process.

2 ANALYTIC PROCESS

Jigsaw does not have capabilities for finding themes or concepts in a document collection. Instead, it acts more as a visual index, helping to show which documents are connected to each other and which are relevant to a line of investigation being pursued. Consequently, we began working on the problem by dividing the news report collection into four pieces (for the four people on our team doing the investigation). Each of us skimmed the 350+ reports in our own unique subset just to become familiar with general themes discussed in those documents. We also jotted down notes about people, organizations or events to potentially study further.

Next, we came together and used *Jigsaw* to examine the entire news report collection. *Jigsaw* expects an xml file as input; the file identifies the unique documents and the entities in the documents. We wrote a translator that would change the text reports and the pre-identified entities from the contest data set into the xml form that *Jigsaw* can read. We then ran *Jigsaw* and explored a number of the potential leads that we each identified by our initial

skim of the reports. At first we looked for connections across entities, essentially the same people, organizations or incidents being discussed in multiple reports.

Surprisingly, there was relatively little in the way of connections across entities in the documents. After about 6 hours of exploration, we really had no definite leads, just many, many possibilities. So we returned to the text reports and some team members read subsets of the reports they had not examined before. At that point, we began to identify some potential “interesting” activities and themes to examine further. What also became clear was that the time we spent earlier exploring the documents in *Jigsaw* was not wasted time. It helped us become more familiar with many different activities occurring in the reports. Thus, new more deliberate examinations and readings of the documents began to uncover more promising leads. We began to find connections across some actors and organizations in the data set.

We were curious, however, why those connections did not show up in *Jigsaw* initially. Upon returning to the system, we learned why. Some of the key entities in the plot we uncovered (*r'Bear*, *Madhi Kim*, *Global Ways*, *Cesar Gil*, etc.) were either identified as entities in only some of the documents in which they appeared or they were not identified as entities at all. *Jigsaw* can only visualize the document and entity information provided to it (it presently has no automated entity identification capabilities), so there was little for us to observe (connections-wise) in our first use of the system on the problem.

At this point, we decided that we needed to update the entity information across the document collection. We started with the pre-identified entities and we created software that would scan all the text documents and identify places where these entities simply were missed. This process resulted in adding more than 6000 new entity-to-document matches over the whole collection, and thus the entity-connection-network became much more dense. The drawback of this technique was that we also added more noise by multiplying unimportant or wrongly extracted entities. Therefore, we manually checked the most frequent entities for validation and made a list of false positive entities (wrongly classified or extracted) for each entity type. We excluded these entities from the document collection and we manually added previously unidentified entities that we noticed while reading the documents.

This whole process provided us with a consistent connection network that was mostly devoid of false positives. Since less than one quarter of the entities across the entire collection appeared in more than one report, we added an option in *Jigsaw* that allows the user to filter out all entities that appear in only one report. Doing so allows us to focus on highly-connected entities at the beginning of the investigation and to add further entities when more specific questions arise later during the analysis. We resumed exploring the documents using *Jigsaw* and it was much easier for us to track down different plot threads and explore relationships between actors and events given this refined entity information.

On our second read of the news reports, we noticed one mentioning the rapper *r'Bear* being taken to the hospital with bumps on his face. This seemed suspicious so we decided to explore it further. We issued a query on *r'Bear* which brought his entity into all the views. Expanding his entity in the graph view showed the reports

*e-mail: goerg@cc.gatech.edu

†e-mail: zcliu@cc.gatech.edu

‡e-mail: justneel@gmail.com

§e-mail: ksinghal@cc.gatech.edu

¶e-mail: stasko@cc.gatech.edu

