TextPlorer: An application supporting text analysis

Chi-Chun Pan* Anuj R. Jaiswal† Junyan Luo‡ Anthony Robinson§

The Pennsylvania State University

ABSTRACT

TexPlorer is an integrated system for exploring and analyzing large amounts of text documents. The data processing modules of TexPlorer consist of named entity extraction, entity relation extraction, hierarchical clustering, and text summarization tools. Using a timeline tool, tree-view, table-view, and concept maps, TexPlorer provides an analytical interface for exploring a set of text documents from different perspectives and allows users to explore vast amount of text documents efficiently.

Keywords: Text, Visualization, named-entity extraction, entity-relation extraction, VAST contest

Index Terms: H.4.2 [INFORMATION SYSTEMS APPLICATIONS]: Types of Systems—Decision support;

1 INTRODUCTION

We designed TexPlorer to support the development of an overview of a large set of text documents as drill-down to items of interest. TexPlorer consist a backend data processing module and a frontend data visualization module. The data processing module of TexPlorer consist of named entity extraction, entity relation extraction, hierarchical clustering, and text summarization tools. Processed data then can be visualized using the TexPlorer web portal and ConceptVISTA, an ontology visualization tool.

TexPlorer uses the following tools to process and visualize the VAST 2007 contest dataset:

1. FactXtractor[1] is a named entity and entity relationship extractor developed by the North-East Visualization and Analytics Center at the Pennsylvania State University. FactXtractor processes text documents using GATE and indentifies entity relations with both syntactical and semantic analysis.

2. ConceptVISTA is an ontology creation and visualization tool developed by researchers at the GeoVISTA Center at the Pennsylvania University. We use ConceptVISTA to visualize concept maps extracted by FactXtractor. More information about ConceptVISTA can be found at http://www.geovista.psu.edu/ConceptVISTA/.

3. MEAD[2] is a public domain portable multi-document summarization system original developed at the University of Michigan. We use MEAD to create summaries for text documents and document clusters. More information about MEAD can be found http://tangra.si.umich.edu/clair/mead/.

4. CLUTO is a family of computationally efficient and high-quality data clustering and cluster analysis programs developed by the Digital Technology Center (DTC) at the University of Minnesota. We use CLUTO to compute content-based document clustering. More information about CLUTO can be found at http://glaros.dtc.umn.edu/gkhome/views/cluto.

5. SIMILE Timeline is a DHTML-based AJAXy widget for visualizing time-based events developed as part of the SIMILE project at MIT. More information about the SIMILE Timeline can be found at http://simile.mit.edu/timeline/.

6. WordNET is a large lexical database of English developed at Princeton University. We use WordNET to perform semantic expansions of keywords within our document filtering tools. More information on WordNET can be found at http://wordnet.princeton.edu/.

*e-mail: julianpan@psu.edu

†e-mail:arj135@psu.edu

‡e-mail:jluo@psu.edu

§e-mail:acr181@psu.edu

2 DATA PROCESSING

Since we were working on the RAW dataset, our first step involved preprocessing the data. First, we used FactXtractor to perform name entity and entity relationship extraction. This process allows us to identify people, locations, organizations, date/time entities, and the relationship among them in the dataset. The results were stored into a database for easy retrieving. Second, we applied document filtering with semantic hyponym expansion on all text documents (including news text, support documents, and blogs) where we input a set of keywords related to our problem and expanded them using the WordNET dictionary. The keywords we used including *terror, police, police, bomb, drug, chemical, weapon, arson*, and *activist*. Then we performed content-based hierarchical clustering using Cluto on the filtered text documents. Finally, we used MEAD to produce a short summary for each clusters in the hierarchical clustering tree.

3 VISUALIZATION AND USER INTERACTION

Processed data can be visualized with different components in TexPlorer. The main interface of TexPlorer is a web portal shown in Figure 1. The top panel is a timeline tool where events are arranged in chronological order. Each event is represent with three keywords picked with the TF-IDF algorithm[3]. On clicking the event icons on the timeline tool, an automatically generated summary of that document is shown in a pop-up window.

The bottom left panel is a tree-view of the hierarchical clustering. Each number represents a cluster of documents that contain similar keywords. The parent clusters contain child clusters with similar contents.

The bottom right panel is a table-view for important people, locations, and organizations. By default, each table shows five entities within a selected cluster ordered by importance. The default importance is defined by counting the appearance of each entity. However, users can override the importance by clicking the "+" and "-" links next to the entities. On clicking a "+" link, the corresponding entity is marked as "very important" and highlighted with red. On the other hand, on clicking a "-" link, the corresponding entity is marked as "unimportant" and removed from the table-view. By moving the mouse over a document name, users can get a brief preview of the document. By clicking on a document name, the document will be shown with all types of entities highlighted and color coded.

Figure 1: The web interface of TexPlorer: the top panel is a timeline tool where events are arranged in chronological order, the bottom left panel is a tree-view of the hierarchical clustering, and the bottom right panel is a table-view for important people, locations, and organizations.

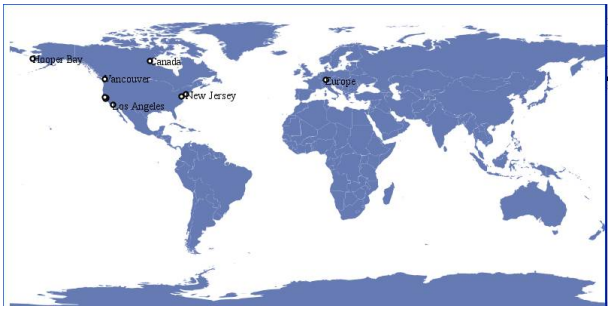

Figure 2: Map visualization showing the 10 most relevant/important locations for cluster 25.

Visualization components on the web interface are coordinated. For example, on clicking an event on the timeline tool, the table-view will be replaced with the corresponding document with color coding for entity types. On clicking a document on the table-view, the timeline tool will be centered to the date when the document is dated. In both cases, the leaf cluster that contain the corresponding document will be highlighted and selected in the tree-view.

In addition to the web interface, TexPlorer can export processed data to external applications. For the location entities in the table-view, clicking the "show map" opens a map display utility where all the important locations in this cluster are plotted. For the people and organization entities, users can then view a concept map for a selected cluster in the ConceptVISTA (Figure 3). Concept maps in ConceptVISTA are based on the Ontology Web Language (OWL) which has significant advantages over traditional data representations such as tables since greater semantics are captured. In addition, we believe the underlying reasoning that could be performed by using concept maps in OWL have immense potential for finding information. One of the next steps in our research is to develop tools that allow this potential to be demonstrated and assessed.

4 Conclusion

TexPlorer was designed to support analysis of relatively large sets of text document. The emphasis is on supporting a full range of overview, filter, and drill-down to details. We integrate some existing text processing tools with creative visualizations allowing users to explore vast amount of text documents. We have used TexPlorer in analysis of the VAST 2007 contest dataset and identified multiple potentially interesting clusters of documents and within them sets of individuals, organizations, events, and places. TexPlorer is one component of a larger research effort to develop a comprehensive suite of visual analytics methods and tools that can support identification of patterns and relationships in heterogeneous sets of information. The next step in our research will be to develop strategies for integrating TexPlorer with a range of geo/information visualization, computational analysis, and knowledge management tools within an analytical workspace that supports a range of analytical tasks.

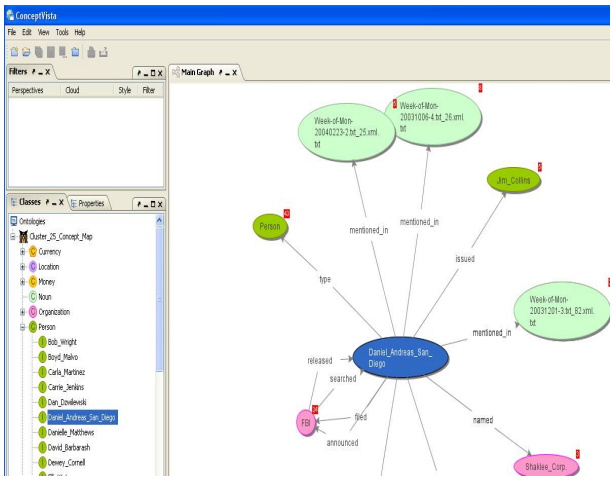

Figure 3: Visualization of concept maps with ConceptVISTA

Acknowledgements

This work was performed with support from the National Visualization and Analytics Center (NVAC), a U.S. Department of Homeland Security Program, under the auspices of the Northeast Regional Visualization and Analytics Center (NEVAC). NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory. We would also like to thank Dr. Prasenjit Mitra, Dr. Alan M. MacEachren, and Dr. Ian Turton for their insightful feedback and comments.

References

[1] C.-C. Pan and P. Mitra. Femarepviz: Automatic extraction and geo-temporal visualization of fema national situation updates. In *IEEE Symposium on Visual Analytics Science and Technology 2007*, 2007.

[2] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May 2004.

[3] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.