

VAST 2007 Contest

TexPlover

Chi-Chun Pan*

Anuj R. Jaiswal†

Junyan Luo‡

Anthony Robinson§

Prasenjit Mitra¶

Alan M. MacEachren||

Ian Turton**

The Pennsylvania State University

ABSTRACT

TexPlover is an integrated system for exploring and analyzing vast amount of text documents. The data processing modules of TexPlover consist of named entity extraction, entity relation extraction, hierarchical clustering, and text summarization tools. Using timeline tool, tree-view, table-view, and concept maps, TexPlover provides visualizations from different aspects and allows analysts to explore vast amount of text documents efficiently.

Keywords: Text, Visualization, VAST contest

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

1 INTRODUCTION

We designed TexPlover, an integrated data analysis system the VAST 2007 contest. TexPlover consists of a backend data processing module and a frontend data visualization module. The data processing modules of TexPlover consists of named entity extraction, entity relation extraction, hierarchical clustering, and text summarization tools. Processed data then can be visualized using the TexPlover web portal and ConceptVISTA, an ontology visualization tool.

TexPlover 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 identifies 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 summary for text documents and document clusters. More information about MEAD can be found <http://tangra.si.umich.edu/clair/mead/>.

*e-mail: julianpan@psu.edu

†e-mail: arj135@psu.edu

‡e-mail: jluo@psu.edu

§e-mail: acr181@psu.edu

¶e-mail: pmitra@ist.psu.edu

||e-mail: maceachren@psu.edu

**e-mail: ijt1@psu.edu

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 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/>.

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, location, organization, 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*, *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 short summary for each clusters in the hierarchical clustering tree.

3 VISUALIZATION AND USER INTERACTION

Processed data can be visualized with different components in TexPlover. The main interface of TexPlover 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 represented 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, location, and organization. By default, each table shows five entities within a selected cluster ordered by important. 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.

