

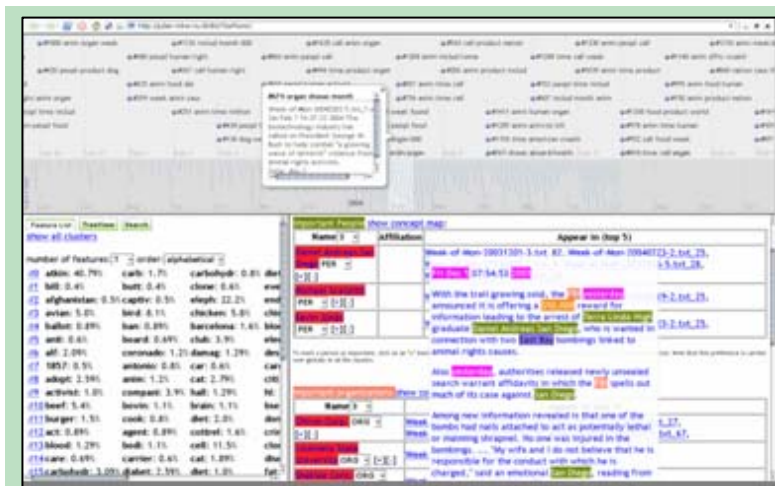
TexPlover

An Integrated System for Text Analysis and Visualization

Benefit: *TexPlover enables academic and government information analysts to find entities and connections in text documents (agency reports, news stories, etc.). A visual work-space allows end-users to cluster documents based on publication date, topical content, and focus location. Named entities (e.g., organizations, persons, and geographical locations) are identified automatically and color-coding highlights each by category.*

TexPlover is an integrated visual analytics application for exploring and analyzing large numbers of unstructured text documents (e.g., news stories, reports). TexPlover's data processing module extracts important locations, events, people, organizations, and keywords by using named entity extraction, entity relationship extraction, hierarchical clustering, and text summarization tools. Processed data is then visualized using the TexPlover information visualization interface that includes attribute, time, and location views. TexPlover provides an analytical interface for users to explore large numbers of text documents efficiently, from multiple perspectives, including place and time.

TexPlover combines components developed by several research teams: (1) FactXtractor is a named entity and relationship extractor developed by NEVAC; (2) ConceptVISTA is an ontology management and visualization application from the GeoVISTA Center at Penn State; (3) MEAD is a portable multi-document summarization system from the University of Michigan; (4) CLUTO is a family of data clustering and cluster analysis programs developed by the Digital Technology Center at the University of Minnesota; (5) SIMILE Timeline is a DHTML-based AJAX widget for visualizing time-based events developed at MIT; and (6) WordNET is a lexical database of English developed at Princeton University.



The main TexPlover interface: *The timeline tool (top panel) has events, each represented by 3 keywords, arranged chronologically. Clicking an event opens a pop-up window with an automatically generated summary.*

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

The table-view (bottom right) depicts important people, locations, and organizations extracted from the text. By default, entity importance is defined by frequency of occurrence in the document, but users can promote or demote individual entities. Mousing over a document name shows a preview, while clicking it shows the full document with entities color coded by category.

The main TexPlover interface (left) contains three component panels, each coordinated with the others based on the user's actions. In addition to the primary web interface, TexPlover can export processed data to external applications. A map utility displays important location entities extracted from a chosen document cluster. For the people and organization entities, ConceptVista is accessed to display a concept map for a selected cluster.

TexPlover 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.

Funded by: Department of Homeland Security

Early Development

Lab Prototype

Commercial Product

Jan 2008

For more information, contact:

Alan MacEachren, (814) 865-7491, maceachren@psu.edu

www.geovista.psu.edu/NEVAC