Model based Interactive Analysis of Interwoven, Imprecise Narratives

VAST 2010 Mini Challenge 1 Award: Outstanding Interaction Model

Victor Chen, Dustin Dunsmuir, Saba Alimadadi, Eric Lee, Jeffrey Guenther,
John Dill, Cheryl Qian, Chris D. Shaw, Maureen Stone, Robert Woodbury

School of Interactive Arts and Technology, Simon Fraser University

Abstract

CZSaw [1] is a visual analytics tool for sense-making across entities, documents, and relations with a focus on supporting the analysis process. It uses a variety of flexible data visualizations to represent and explore networks of entities and relations from different perspectives. CZSaw supports clustering documents and entities into smaller groups to make sense of them and weave individual facts into a complete picture. CZSaw also provides entity refinement functions to support interactive data cleaning. Its dependency propagation mechanism speeds the analysis sense-making loop by automatically synchronizing data and views, and propagating changes to the whole system.

Keywords: Visual analytics, investigative analysis, intelligence analysis, sense-making, analysis process

Index Terms: I.3.8 [Computer Graphics]: Applications-Visual Analytics, I.6.9 [Visualization]: Information Visualization, H.5.2 [Information Systems]: Information Interfaces and Presentation.

1 Problem Description

This challenge has 103 documents, which come from different resources and describe different countries, regions, and people. Within the encompassing story of illegal firearm dealing activity, there are also several sub-threads. Many errors or inconsistencies existed in these documents, for example, miss-spelled names in surveillance reports. It is almost impossible for an analyst to keep track of the whole scenario through reading alone, even if she has enough time to read all the documents.

2 The CZSaw System

CZSaw helps analysts solve large scale problems via flexible data views that provide overviews and details on demand. In addition, CZSaw provides process views to manage the complex analysis process itself.

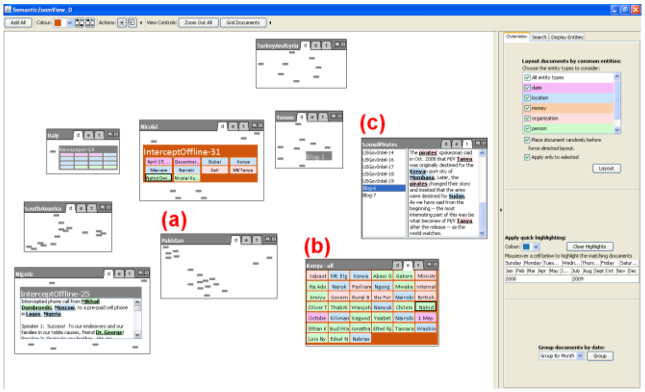
First, CZSaw's data views allow visualization and manipulation of entities, documents, and relations for use in the sense-making process. These visualizations aid in selective reading of documents to make connections between disparate facts. The *Hybrid View* is an enhanced graph visualization of entities (nodes) and relations (edges) where the nodes can be visualized with a variety of techniques. The *Semantic Zoom View (SZV)* examines documents at several levels of detail (overview, document's entities, and detailed text). The *Document View* allows the analyst to read documents and scan their contained entities.

Second, CZSaw provides an interaction model and history mechanism to support the analysis process. User interactions are recorded and translated into a script language at the task level. Analysts can then replay or reuse their analysis steps to help them understand, explore, and reference their analysis process. CZSaw also creates a model of the analysis process in the form of a dependency graph through which changes can be propagated. Driven by the dependency propagation mechanism, data views automatically update themselves to reflect changes in data such as modifications of entities. CZSaw provides users computational power for data query and management. Functions include managing display states and layout, querying/filtering entities and relations, and refining entities on the fly.

3 Supporting Key Analysis Interactions

CZSaw relies on extracted entities, as do many similar systems. Thus we asked our colleagues in the SFU Natural Language Lab to run entity extraction algorithms on the original dataset to generate an XML file containing extracted entities (person, location, date etc).

To take advantage of differing analysis approaches and exercise different CZSaw strategies, we began this challenge in two separate teams. To read and analyze all 103 documents and match the desired solution format, both groups adopted a divide-and-conquer data organization strategy, grouping documents and entities by country and event. One team focussed their investigation within the *Semantic Zoom View (SZV)*, grouping documents by country before drilling down to investigate further. The other team used the *Hybrid View's* node-link graphs and the *Document View* to read details. After this data exploration stage, we integrated the teams' findings and reported the outcome.



Intercepted phone call from Michail Dombrovski, Moscow, to a pre-paid cell phone in Lagos, Nigeria

Speaker 1: Success! To our endeavors and our families in our noble causes, friend Dr. George!

The pirates' spokesman said in Oct. 2008 that M/V Tanya was originally destined for the Kenyan port city of Mombasa. Later, the pirates changed their story and insisted that the arms were destined for Sudan. As we have said from the beginning -- the most interesting part of this may be what becomes of M/V Tanya after the release -- as the world watches.

Figure 1. The *Semantic Zoom View (SZV)* visualizes documents with multiple levels of detail. (a) Group overview showing document glyphs. (b) The set of all entities in the group. (c) The full text of the group documents.

The *SZV* (Fig. 1) shows documents that can be semantically zoomed to three levels: overview, entities in the document, and detailed text. For an overview, it uses a clustering algorithm to

E-mail: { yvchen, dtd, salimada, ela10, jguenthe, dill }@sfu.ca, qianz@purdue.edu, shaw@sfu.ca, stone@stonesc.com, rw@sfu.ca

layout documents – the more entities two documents have in common, the closer they are placed, resulting in clusters of documents about the same set of entities. In this challenge, clusters contained many documents related to arms dealing in the same country. Our first team scanned and searched the clusters and created permanent groupings in the *SZV*, where each group displayed the documents for a country. Each group was analyzed separately to keep the information flow to a reasonable cognitive load. Similar to individual documents, groups in the *SZV* can be displayed as sets of zoom-able documents, combined sets of entities for brushing across the rest of the view, or the full text of each document.

CZsaw users can create sets of entities and relations, and visualize them with custom layouts in the *Hybrid View*. Examples include all entities in the data set, all entities of a given type (e.g. people), entities filtered by value (e.g. the name "Nicolai"), or entities related to previously defined sets. Fig. 2 shows all reports in a *Hybrid View* as a graph where two documents are connected if they contain one or more of the same entities. Fig. 3 shows the social network of people, and how they are connected by code words (e.g. textbooks, farming and drilling equipment), arms deals, and money transfers. To create this view, we listed all the people, searched for codes, bank accounts, and money that connect at least two people, and then displayed these connections. With such a selective display method, we can examine connections among people from different perspectives (e.g. country, date, and arms deals). The force-directed layout automatically pulls related entities/documents closer, producing clusters. Thus, we were able to examine each cluster by reading a smaller number of documents.

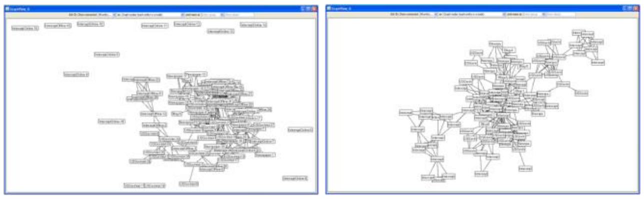


Figure 2. Left: Documents graph before entity refinements. Right: After entity refinements.

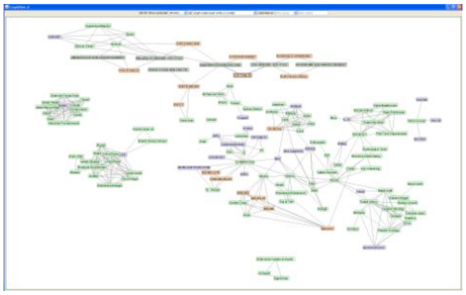


Figure 3. People (green nodes) are clustered into groups defined by code word connections and money transactions.

The left image in Fig. 2 shows many isolated documents. Scanning through them showed that machine entity extraction was neither fully accurate nor complete. CZSaw allows users to interactively manage entities enabling correction of errors and recording of hypotheses (that two entities are the same) during the analysis process. Operations include:

- *Extract new entities*: Users can manually extract new entities while reading a document and also create new entity types (e.g. bank account).
- *Merge entities*: An entity may be misspelled in some documents, leading to two different entities in the system, for example: person "Khemkhaengare" and "Khemkhaengon". The user can merge several entities into one.
- *Alias entity*: A person may appear as a phone number or email address. Merging such entities with their name entities will lose this valuable information. The alias entity function creates an equivalence relationship among such entities so that any document related to one will also relate to its aliased entities.
- *Link entity*: In some cases, such as telephone conversations, identifiers such as "caller1", "I", and "him" can belong to specific people as determined by the analyst. While reading documents, the analyst can create linkages between existing entities and a document when appropriate.
- *Unlink entity*: The user can remove an entity from all documents (deleting the entity) or from specific documents.

The user can manually refine entities while reading a document or working within other views (e.g. similar nodes in *Hybrid View's* entity network bring about entity-merge possibilities). CZSaw's dependency propagation mechanism instantly updates content and layout in views to reflect these changes, which may form new clusters or merge existing clusters (Fig. 2).

By capturing the analyst's interactions, CZSaw creates a model of the analysis process in the form of a directed acyclic dependency graph capable of propagating changes. Nodes in the graph (*variables* in CZSaw) are results generated from user interactions. The user interacts with entity or relation variables in views to create the next step's results. Edges indicate dependency relationships among variables. Any content change to a variable triggers the propagation mechanism to update downstream variables and in turn update data views to reflect the change. The graph's root node represents the entire data set. Thus any change to the data set (such as an entity refinement operation) starts the propagation at the root node, potentially changing all analysis results and updating all data views. As entity refinement proceeds, the document network is transformed into a more understandable image (Fig. 2). The analyst can also reuse parts of the analysis process by assigning new data to one node in the middle of the graph.

Interactions during the analysis process are transformed into a text script. Replaying this script lets the analyst review the whole analysis process. Editing this script supports fine control of the analysis process, for example to quickly change parameter values used in interactions to get better results.

This script also facilitates collaborative analysis. For example, one analyst recorded the steps to create document groups in the *SZV*. Fellow team members were then easily able to recreate the groups in their own instance of CZSaw by replaying the script.

4 Summary

CZSaw provides rich features for analyzing real-world documents via clustering and data cleaning. Interaction, data and visualization are tightly integrated by the underlying script and dependency graph.

References

[1] N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw and R. Woodbury. "Capturing and Supporting the Analysis Process", Proceedings of IEEE Visual Analytics Science & Technology (Atlantic City, NJ, Oct 11-16, 2009), pp. 131-138. Oct 2009.