The CZSaw Notes Case Study

Eric Lee, Ankit Gupta, David Darvill, John Dill, Chris D Shaw, Robert Woodbury
School of Interactive Arts & Technology
Simon Fraser University

ABSTRACT

Analysts need to keep track of their analytic findings, observations, ideas, and hypotheses throughout the analysis process. While some visual analytics tools support such note-taking needs, these notes are often represented as objects separate from the data and in a workspace separate from the data visualizations. Representing notes the same way as the data and integrating them with data visualizations can enable analysts to build a more cohesive picture of the analytical process. We created a note-taking functionality called CZNotes within the visual analytics tool CZSaw for analyzing unstructured text documents. CZNotes are designed to use the same model as the data and can thus be visualized in CZSaw's existing data views.

We conducted a preliminary case study to observe the use of CZNotes and observed that CZNotes has the potential to support progressive analysis, to act as a shortcut to the data, and supports creation of new data relationships.

Keywords: Text Analytics, Annotations, Note-taking, Sensemaking, System Architecture

1. INTRODUCTION

One of the goals of visual analytics is to build computational tools to support data analysis. The data collected and analyzed by the intelligence community often comes in the form of written reports. Analysts read reports to induce a narrative explanation of the facts reported in the documents. The problem is that there are too many documents to read in a reasonable amount of time. Analysts need tools to help them discover information and organize their findings.

To better understand the features tools need, researchers have sought to understand the process of analysis. Pirolli and Card performed a cognitive task analysis of intelligence analysts to create a model of sensemaking [1] that describes the steps analysts take as they work. Analysts begin by searching data sources for information relevant to their problem. As they search, they save pieces of information that caught their attention in a *shoebox*, or evidence file. Analysts generate schemas to organize the information they have saved, which forms the story or evidence in support for the analysts' hypotheses. Finally, the analysis is presented to those who requested it be done. While this process is described in a linear fashion, Pirolli and Card observe these steps occur in several loops. Analysts repeat steps in the analysis as their needs require.

Kang and Stasko emphasize the non-linear nature of sensemaking when they suggest that Wheaton's Model might provide a better description of analytical activity [2]. Wheaton's Model describes intelligence analysis as a process of four tightly coupled functions: production, analysis, collection, and modelling. All functions start at the beginning of the analysis process, and the emphasis placed on each function changes over time. The process is dynamic and non-linear. An analyst continues to refine and change knowledge of the data until the moment the work is delivered.

Analysts, generate thoughts, ideas, and questions, to follow up on as they read and analyze data. In this paper, we refer to such artifacts as *notes*. Notes may be annotations on the data itself [3], visualizations, diagrams or informal text and symbols written on a separate notepad [4], or VA tool-specific note objects [5] [6] [7]. Analysts need to record and store such notes in a shoebox in the foraging loop or write down hypotheses as a note during the sensemaking loop [4]. Research has also shown that analysts create annotations and notes throughout the iterative analysis process. Taking notes is necessary because it is impossible for analysts to keep all of their analytical findings in their memory. Heuer [8] warns that analysts are more prone to cognitive biases if they primarily do analysis in their heads. Therefore, offloading notes externally can lessen their cognitive load and ensures that such captured artifacts are available, reviewable, and complete. Analysts use notes to support their sensemaking goals and ultimately come to a conclusion in their analysis. We could argue that, in some cases, such record keeping of notes can be as important, or more so, than the data itself.

Many visual analytic tools now include capabilities to capture such notes, visualizations, hypotheses, evidence, annotations, reminders and to-dos to support data analysis. Some, such as Sandbox for Analysis [9] and Entity Workspace [10] build links between the note and references to the data. Some tools enable note-taking functionality as an add-on to a web browser such as ScratchPad [11]. Click2Annotate automatically generates an annotation based on a

set of templates and a selected subset of data in a scatterplot visualization and can be viewed in a separate annotations window [12].

However, most such VA tools represent notes as passive text and as another object to manage during analysis. Furthermore, existing tools expect note taking activities and the organization of notes to be done in a separate workspace within the tool. Search mechanisms for notes and the data are also often separated. Sensemaking requires an integrated knowledge base of "sensemaking results" and the data being analyzed to draw a conclusion about the data. Analysis "is a progressive process in which newly synthesized knowledge becomes the foundation for future discovery" says [13]. However, in many of the current VA tools described above, the analyst must manage the connection between two workspaces in such tools, one for note taking activities and one for data exploration.

Therefore, we propose that VA tools should be designed to provide a greater integration of notes and the data within the system. Our system performs this integration by simply representing notes in the same way as the data documents. This has benefits of a simpler software design, since user created notes in a VA tool can be fed back into the knowledge database (now consisting of both notes and data). Such notes then seamlessly appear in future data queries (because notes are returned in the data queries as well as normal data documents). This integration of notes and data could also provide a more cohesive knowledge base of both previous analytical findings and the data, and potentially introduce new ways for analysts to create and use notes in VA tools.

This paper is organized as follows: section 2 describes related work in note-taking and the analysis process, the representation of notes in VA tools, and related VA tools that support note taking. Section 3 describes the design of CZNotes. Section 4 describes a case study in which participants used CZNotes to solve a VAST 2011 mini challenge [14]. Section 5 provides a discussion and conclusions.

2. RELATED WORK

In Pirolli and Card's sensemaking model, note-taking can occur at any of the activities in the model, especially during the schematizing and hypotheses phases where threads of analysis and associating evidence for and against each hypothesis might be recorded and later reviewed. Mahyar et al's study [4] of teams of analysts found many stages of analysis similar to those documented by other models of analysis [15] [1]. In addition, Mahyar's [4] framework for collaborative analysis identified that note taking occurred throughout the analysis process. The researchers found multiple uses for such notes for facilitating problem solving, for example, "by recording the direction and sequence of the steps taken", creating reminders, and eventually the final report.

Marshall observed this natural co-location of annotations and data when analysts worked with pen and paper. She identified three types of note-taking styles from her Work Practices Study with professional analysts - annotative notes, interpretive notes, and reminding notes [3]. The annotative notes Marshall described included notes in the margins of a text document and the highlighting of the text. Interpretive notes are those that refer to one or more source "to record conclusions they have reached or material they have integrated from several sources". Reminding notes are those that document a process or list a set of tasks to do, possibly representing procedural knowledge of the analytical process.

Thus, notes have multiple roles to play in the analysis process. A good notes facility should support all of these note activities.

2.1 Integrating Notes and Data

Analysis requires the continual build-up of understanding of the data. IN-SPIRE is a text analysis tool which can analyze and group textual data documents into clusters using its Galaxy view or ThemeView tools [16]. The user can place documents into groups. These groups can each represent a different hypothesis. IN-SPIRE enables the user to mark a document as supporting or refuting a hypothesis (group) and visualizes this information in a table for further analysis.

Click2Annotate [12] and Touch2Annotate [17] provide an automatic generation of notes in natural language style in the domain of multi-dimensional data. The analyst supplements these automatic annotations with their own notes in a separate window.

Jigsaw provides a Tablet view that allows the user to add both data entities and notes into the same view. Linkages can be drawn among the items and it also supports the creation of a timeline of events to support "story building" [18].

Analytical findings are referred to as synthesized knowledge by [13] and are usually represented by notes of some sort. Gotz et al.'s HARVEST system allows the user to both annotate existing information and refine existing annotations. In their tool, the synthesis space is placed next to the exploration space. These spaces are linked such that if the user clicks on the hypothesized suspect, the corresponding evidence linked to the suspect is also highlighted in the exploration space. They state "… it is important to allow users to quickly check the correspondence of the synthesized knowledge and its evidence at any given point of the investigation."

Another research effort supporting a tighter integration of information foraging and sensemaking (using notes) is nSpace [19]. nSpace is a combination of two tools: TRIST [20], an Information Retrieval tool and Sandbox [9], a sensemaking and evidence marshalling tool. Although both of these components are within the same system, there is still a spatial disconnection between the information retrieval workspace and the analysis workspace.

Entity Workspace [10] also tries to achieve support for the foraging and sensemaking loops by incorporating into a single application the CorpusView document reader (foraging), and the Entity Workspace evidence panel (sense-making).

Although HARVEST, Entity Workspace and nSpace attempt to integrate the synthesis space and the exploration space into the same tool, the two functions are still separated from one another as separate panels within a window and their underlying representation of notes (synthesized objects) and data are separate. We attempt to go a step further with the design and development of CZNotes and merge the "Synthesis Space" with the "Exploration Space" into a single space.

Several VA tools capture the states of the analysis process for provenance purposes and some allow the user to annotate these analysis or visualization states. For example, both Sense.us [21] and CommentSpace [22] provide a collaborative setting to perform visual analytical tasks and also write comments associated with visualization states.

The ARUVI system for scatterplot analysis, note taking and sensemaking in the Knowledge View, and analysis state history capture [7]. ARUVI provides scatterplot visualizations in combination with what they call a "knowledge view" diagramming workspace for recording notes similar to Sandbox. In addition, ARUVI records analysis states. ARUVI also allows users to link a note to a visualization state.

InsightFinder is a web based tool to support post-search activities to find specific content within a resulting web page [23]. Notes taken during web searching are used as a source of recommendation info for new page views. The goal is to make a connection between a user's previous notes and the new content they are viewing.

In a similar tool called ScratchPad, notes can be created in a "notes sidebar" of a web browser. These notes are compared with the current web page in view. Content similarities among the notes and a new webpage triggers highlighting of the relevant "notes sidebar" items [11].

Endert et al. report on ForceSpire [24] [25], which presents a scatterplot of documents similar to IN-SPIRE. The user can edit the layout of documents by moving them around the scatterplot, and by adjusting the weighting of dimensions. Similar to CZNotes, users may add annotations to documents that are then processed by ForceSpire's force-directed layout system.

3. DESIGN OF CZNOTES

We have designed and integrated notes into CZSaw [26], an existing VA tool that focuses on analyzing unstructured text. CZNotes takes a different approach to manging annotations by simply including CZNotes as search results in the CZSaw data queries. We hypothesize that when the user sees the CZNotes in the search results, this could potentially remind analysts of their previously created analytical steps, potentially helping with the analytical process.

Another note property that can be inherited from the data document representation in most unstructured text analysis tools is the idea of entity extraction. In the domain of unstructured text data analysis, tools such as TRIST [20], Jigsaw [27], CZSaw [26] and Entity Workspace [28] have employed named entity recognition tools to extract some structure from text data. Interactions with entities rather than plain text makes the analysis more structured and enables the user to do more with the data such as building entity graphs and working with entity objects.

CZSaw, an entity based text analysis tool, was inspired by Jigsaw and similarly uses an entity-structured approach [26]. However, these named entity recognition tools are far from perfect, resulting in missed or misidentified entities. Thus VA tools like Jigsaw and CZSaw provide the user with the ability to refine the entity extraction in real-time. For example, the user can remove wrongly identified entities and extract new entities within the data documents.

By representing CZNotes the same way as the data documents within CZSaw, CZNotes inherit such entity extraction capabilities and enables the analyst to identify entities within a CZNote. This entity structure is also a means to create relationships among documents. For example, two entities are related if they are both contained in the same document. Therefore, CZNotes could be used to create relationships among entities if, for example, an analyst extracts two entities within one CZNote, than these two entities are related by being contained in the same CZNote – bibliographic coupling.

3.1 CZSaw Overview

CZSaw helps analysts explore document sets with a combination of data views and process support. To start analysis, documents are imported into CZSaw. The text of the documents is processed using Named Entity Recognition (NER) to find all the people, places, locations, and dates mentioned within the documents. With the extracted entities, a data model is created that relates documents and their contained entities. With the data model built, documents and entities

can be searched using the search panel and added to the desired data view. Each of these steps is recorded in CZSaw's script.

Each entity has a type and various attributes that are pre-defined. A CZSaw document is also an entity of type *Report*. For example, a news article would be represented in CZSaw as a *Report* entity with an attribute containing the text of the article and another attribute listing all the entities contained within it (e.g., people, locations, dates, etc). Data visualizations and data queries all act upon these fundamental *Entity* objects.

Interaction with the user interface creates script statements that are processed to generate the dependency graph. The dependency graph represents the state of the CZSaw computation. Nodes in the Dependency Graph represent collections of entities, which could be a single person's name, a *Report*, a list of places, or any other tuple of entities. Node B depends on node A if B is assigned the result of a function called with A as its parameter, in a manner similar to spreadsheet dependencies. When executed, CZSaw script statements generate these functional dependencies, which remain active during the CZSaw session. CZSaw has functions for searching for an entity, and displaying a set of entities in a visualization. All meaningful interactions with the system are captured in the script. An analyst can use the script to return to a previous state of his analysis, or reuse his analysis with different initial queries. Because of the dependency graph, if the keyword passed to the search function is changed, the whole system is updated. This allows an analyst to repeat analysis steps for different search data without having to redo the work.

With this entity structure, we utilize connections among entities and the documents to support queries and interactive analysis. When documents *contain* the same entity, they are connected through the common entity and have a "*bibliographic coupling*" relation. For example, if documents C and D contain the person entity "Bob", then documents C and D are related by bibliographic coupling.

CZSaw provides 3 data views for exploring the data: the Document View, the Hybrid Graph View, and the Semantic Zoom View. The Document View (DV) displays data document text and a list of entities contained within all documents.

The *Hybrid Graph View (HGV)* visualizes entities and documents in a node-link graph view. Double clicking on a node will execute a query to retrieve related entities and documents and display a connection to it in the graph (Figure 2). Thus, the *HGV* enables the user to discover and explore entity relationships within the data.

The *Semantic Zoom View (SZV)* [29] is focused on visualizing subsets of the documents. The SZV allows the user to apply clustering algorithms to visually group (cluster) documents with similar content. The SZV also allows the user to zoom to different semantic levels of a CZSaw document. For example, one level consists of a listing of all entities within a document. A deeper level contains the full (scrollable) text of the document (Figure 3).

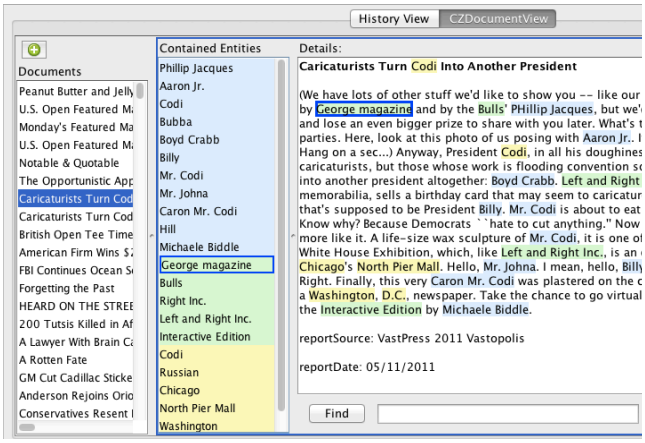

Figure 1 CZSaw's Document View. The document list is in the left panel, the selected document is on the right, which contains entities of various types shown in the center panel and highlighted in context on the right

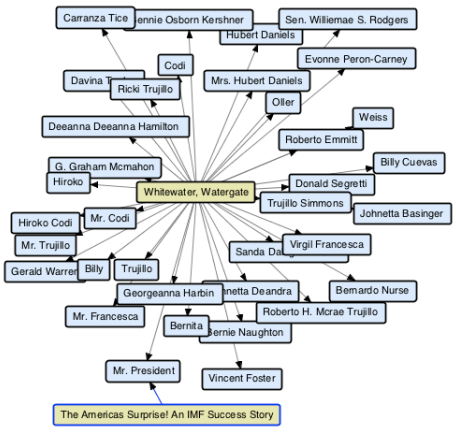

Figure 2 CZSaw's Hybrid Graph View. The user double clicked on the document "Whitewater, Watergate" to retrieve a set of people nodes (blue nodes) contained in that document.

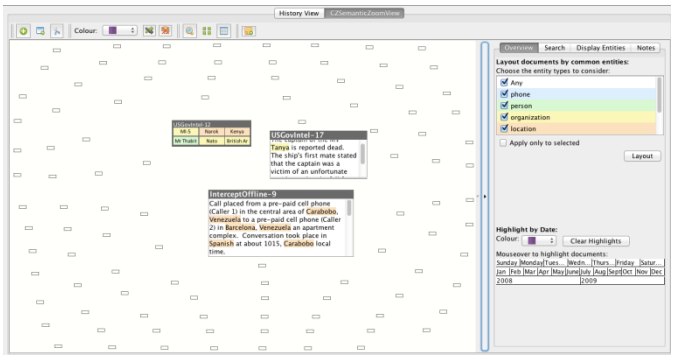

Figure 3. CZSaw's Semantic Zoom View. Documents are laid out in a space as nodes. These nodes can be zoomed into with multiple semantic levels. For example, one semantic level contains all the entities contained within the document and a deeper level of zoom reveals the document text.

3.2 CZNotes Design Approach

Before we begin the discussion of the design of CZNotes, we need to define what constitutes a "note". In general, a note can be any artifact created by an analyst to support their sensemaking, such as a text note, graphical annotation, highlights, etc. For this initial exploration however, we restrict notes to be text documents only.

In the initial design of CZNotes as a text object within CZSaw, we noticed a resemblance to CZSaw documents – text with entities extracted. We thought it would be potentially very helpful to analysts to also enable such entity identification and refinement capabilities within CZNotes since notes are often written about the entities and the data documents being analyzed. Thus, the simplest method for incorporating CZNotes into CZSaw was to represent notes in the same way as CZSaw documents are represented. CZNotes are represented as entity objects with the entity type "note". This simplifies the implementation. Users can use the same CZSaw tools to interact with CZNotes as they could with any other documents and entities:

- Visualize and interact with CZNotes in existing CZSaw Data Views.
- Query CZNotes with existing CZSaw search capabilities.
- Inherit the Entity Refinement capabilities of CZSaw documents.
- Incorporate CZNote interactions (i.e., creation and editing a note) into the CZSaw Script.
- Integrate CZNotes into the dependency model.

Although much can be accomplished using existing functions of CZSaw with CZNotes, we also needed to add some unique CZNote interactions to support note-taking tasks:

- Create and Edit a CZNote.
- Support for entity recommendations.

By integrating CZNotes with CZSaw and representing notes as data, we intended that several potential benefits would accrue to help the user's analysis.

1. **Note Reminders in Query Results**. By integrating CZNotes with the data, we can search notes and the data simultaneously. Combining the query of data and notes and returning notes in the query results along with data results can remind the user of their previously created notes.
2. **Support for iterative Analysis**. Notes can be a trigger to further analysis. Users can use notes and the relationships among their contained entities to query for other related entities and documents.
3. **Support the creation of new relationships among entities.** Entities and documents are related based on bibliographic coupling and co-citation relationships. In CZSaw, for example, entities contained in the same document are marked as related. Therefore, an analyst can link two entities together by adding them to a CZNote.
4. **Bring new data into CZSaw.** Analysts can write any type of text content in a CZNote. The analyst potentially could refer to information outside of the current data set in the CZNote and identify new entities the note. Thus, such new information and new entities could be integrated with the current data set as official CZSaw objects.

Our goal in representing notes as data is to benefit analysts in their analyses. For example, feeding notes back into the knowledge set (data and synthesized notes) can assist the analyst in the iterative analysis process. Visualizing notes and data in the same data views can benefit the analyst by enabling the visualization of data and notes in the same space to form a more cohesive picture of the facts (data) and synthesized knowledge (CZNotes). We are also interested to see if analysts may in fact develop new usage scenarios of CZNotes as data. We outline the following possible scenario to illustrate the use of notes, from the VAST 2011 Grand Challenge.

3.3 Example Analysis Scenario

An analyst initially starts her analysis to uncover a terrorist plot by using the Semantic Zoom View to determine the major themes of the data documents (news articles, communication transcripts, etc). She notices two prominent people named "Nicolai" and "Boonmee" within a cluster label. She creates a new CZNote reminding herself to find more information about these two people, creating CZNotes containing hypotheses regarding the relationship of Nicolai and Boonmee. In one instance, she created a new CZNote about Nicolai and Boonmee staying in Nigeria for a few days. The analyst then wraps up for her weekend. At the start of her following workweek a colleague informs her of a virus outbreak and the analyst focuses her attention analyzing the data set for the cause of the outbreak. She investigates the location entity "Nigeria" as a possible source of outbreak by adding the entity "Nigeria" into the Hybrid Graph View. Double clicking on the Nigeria entity node triggers a query to find related entities. The results of her query returned Nicolai and also all the previous CZNotes that she had written about Nicolai in Nigeria and this helps her to forge the connection between the hypotheses she had written about Nicolai's terrorist activities and the virus outbreak. The script recording the creation and editing of CZNotes was then shared and loaded into another collaborating analyst's CZSaw machine for follow up analysis and review. The collaborating analyst re-executes the script and loads all the CZNotes into the Semantic Zoom View and runs the clustering tool to determine the theme of the shared CZNotes and continues from there.

3.4 Design Discussion

It might seem that by simply adding a new entity label for *Note* accomplishes nothing new programmatically, but we justify this design choice in two ways. First, building upon the findings of [3] [8] [1] and [2], notes are pervasive throughout the sensemaking process. Notes convey annotations, sequencing of events, interpretations, hypotheses, and analysis process. In the cases of pencil and paper, notes are co-located with the annotated item.

This corresponds to the Gestalt principle of *Proximity* [30], which is the perceptual principle that items close together are perceived as related (absent other cues). The point of this discussion is that a note-taking facility that presents notes in a separate window with separate set of operations will operate against this perceptual principle. Depicting CZNotes close to their annotated item will enhance the perceived relationship between the note and the related item, and thereby decrease the perceptual and cognitive burden of relating the two.

Second, since the CZNotes system participates in the main computational engine of CZSaw, it enables new forms of activities with notes that are greater than simply a human-readable annotation. Since CZNotes are first class citizens in the CZSaw system, they can be used within the system as a means of imposing organization on the data that can be computed and updated as the analyst's thinking evolves. As the case study reported in section 4 illustrates, our experimental analysts developed ways of using notes that depended on the computational power of CZSaw.

3.5 Entities and CZNotes

When data is imported into CZSaw, Named Entity Recognition (NER) tags text strings within each data document to identify possible people, places, time, etc. These entity objects are the core component of CZSaw's data model. An entity has an entity type and a set of attributes. For example, a CZSaw document is an entity with entity type "**report**" and an attribute is the *title* of the document. Another attribute would be a list of entities contained within the document. Each entity (eg. Place) is indexed using the Lucene text indexing package. When a user creates a new CZNote, the note is added to this index as an entity with type "note" that has attributes similar to those of "report" entities such as title, text content, and the entities it contains. These entity objects are stored in an index, which can be searched during runtime.

Owing to imperfections in NER, CZSaw provides facilities to refine entities – merging duplicates, removing incorrect entities. These operations can similarly be applied to the text and entities within a CZNote. Allowing this capability within a CZNote opens up possibilities in terms of how the user can interact with the data set. For example, CZNotes could be used as a vehicle to bring external data into the existing data set since the user can write his/her own content in a CZNote and then subsequently extract entities from this content.

CZNotes can also reference CZSaw documents and other CZNotes as they are also represented as entities. CZNotes can thus be used as a mechanism to build a deep level of interconnectivity among CZSaw documents, other CZNotes, and entities. When a user types a CZNote, the words are is scanned for matching items from the index, which can be added to the CZNote's entity list. The interaction is much like spellcheck incrementally checking for wrong words.

Although the initial set of relationships among the entities is important, analysts also need the capability to add their own relationships among documents and entities based on a combination of their own knowledge and external information. Analysts need a way to explicitly create such relationships in order to allow analysts to continue progressive work with them. The creation of bibliographic coupling and co-citations relationships is possible through the use of CZNotes.

For example, an analyst writes in a CZNote that "John" and "Rob" were at the same location at the same time after reading two separate documents. The initial data set has no document that describes such a connection between "John" and "Rob". The analyst then creates a CZNote stating that John and Rob were both in Vancouver at the same time and extracts the people and location as entities (Figure 4). Now future queries for people related to John would reveal Rob as a connection, which was not present in the initial data set. Similarly, a query for people associated with Vancouver would return John and Rob.

In this way, we are using bibliographic coupling within notes to create new relationships. This synthesized knowledge of the note is added to the integrated pool of knowledge, i.e. the data and notes, and can support future analysis.

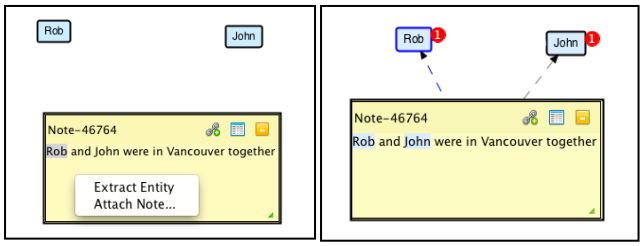

Figure 4 The images above show the creation of a relationship between Rob and John through a user created CZNote by identifying the two entities in the same note.

When the analyst completes the creation of a CZNote, a line of CZSaw script is issued which has the parameters *title, note text*, and the *list of referenced entities*. This allows the analyst to replay the script and reload all the CZNotes from a previous session.

When a CZNote is added, the script updates the dependency graph, and all items (visualizations) that depend on the set of entities are updated. Thus, Figure 4 right shows a red bubble with the number 1 indicating that the entities *Rob* and *John* each have a CZNote attached. All views are similarly updated, as appropriate.

We hypothesize that with the integration of CZNotes and the data, we can remind the user of states of the analysis. For example, in the scenario posed above, the analyst found new information that John and Rob were in the same place at the same time and created a CZNote about this. This note creates a link between John and Rob by a co-citation relationship. Relationships with CZNotes are depicted with a dashed line. When the analyst queries for data related to John in the Hybrid Graph View, the CZNote linking John and Rob in Vancouver is returned. Thus, this approach reminds the analyst about their previous findings without adding additional functions to the system. Furthermore, the dependency model described earlier will also update previous user query results to include any new CZNotes that match the query.

To view all of the CZNotes, a user may simply use the document view (Figure 1), and select all of the CZNotes for inclusion in the document list of the left pane. CZNotes are distinguished by color. When a user selects a CZNote, the entity list in the middle and the document pane on the right are populated appropriately.

4. CASE STUDY

In this exploratory study we investigate how representing CZNotes as CZSaw data documents can affect the users in their analytical processes. The questions we were interested in answering were:

- Would returning CZNotes in query results remind analysts of previously created notes?
- Would CZNotes be used to 'continue' an ongoing analysis? Would CZNotes be used as a means to create relationships among the data through co-citation?
- Would CZNotes be used as a way to introduce new data into the CZSaw system?

We took the mixed methods longitudinal approach suggested by [31]. They suggest a multi-dimensional approach is required with case studies over long periods of time, along with surveys, interviews, ethnographic observation of participants and automated user activity logging.

We employed a mixed methods longitudinal study to observe the use of CZNotes within CZSaw. These multiple means of data collection provided us with a more comprehensive understanding of the use of CZNotes and CZSaw.

The participants' task in our user study is to solve the IEEE VAST 2011 Mini-Challenge 3 using the data and task descriptions provided in the challenge [14]. The study consisted of single team collaboration among 4 participants over the course of 3 weeks to solve the Mini Challenge. The data set consists of 4000 fictional news articles created as a combination of machine generated documents and manually edited content. Participants were provided a written description of the task in which they were asked to use CZSaw and CZNotes to analyze the news articles and identify and report on any threats to a fictional place called "Vastopolis" based on their analysis. The articles publishing dates ranged over a period of 3 months. We encouraged participants to use CZSaw and CZNotes, but did not disallow the use of other analysis tools if they were felt to be necessary. Participants did initiate the use of a wiki as a central collaboration space as well as email and instant messaging to communicate. Participants were strongly encouraged, but not required, to first create a CZNote with the information that was going to be shared on the wiki. The participants were allowed to work on the challenge at any time and place and with any other participants. Participants spend at least 2 hours per week doing analysis in addition to the time spent in the discussion sessions.

The participants were all visual analytics researchers: one Masters student, two PhD students, and one Post Doctoral Fellow. All were also members of the CZSaw research group, although none of the participants participated in the design or implementation of the CZNotes facility.

We chose team members for this study primarily because the goals of the study were to explore the design space of possible uses of CZNotes. No other group was available to us at this early stage of CZNotes development that would have sufficient knowledge of CZSaw to be able to use it strategically, and for such a significant length of time.

Moreover, CZSaw team member also had significant practice in using CZSaw to solve VAST Challenges, so the element of practiced prior experience was essential.

Scripts

The script files users generated from their use of CZSaw were collected for analysis. These script files enabled the users and the researcher to reload past analysis sessions. This script file was useful to allow the users to reload past analysis sessions and continue their analysis.

Discussion Sessions

Participants met with the investigator seven times during the 3-week period, twice per week for the first two weeks and three times in the last week, in discussion group sessions. Generally, the first half of each session was made available for the participants to collaborate on the analysis task itself and share their findings, describe their processes and plan next steps, and provide their feedback on their usage of CZNotes. In the latter half of each session, the investigator asked more focused questions to gather feedback about specific CZNotes features, usability issues, etc. These focus sessions were video and audio recorded.

Journaling

Participants were asked to keep a journal of their activities throughout the study regarding their activities each time they worked on solving the challenge. The journal was guided by a set of questions asking the participants what the goals of their analysis session were and their feedback on CZNotes usage. Participants were allowed to submit their journal entries by whatever means they were most comfortable with. All participants submitted electronic journals, either through a web survey the investigator prepared or as a text file that the participant maintained.

4.1 Results

Over the course of the 3-week study, participants created 103 *CZNotes*, edited *CZNotes* 138 times and appended entity references to *CZNotes* 93 times. Participants P1, P3, and P4 were the primary contributors to these *CZNote* actions. P2 created 1 *CZNote* during the study, perhaps due to P2's very busy schedule during the time frame of the study.

Participant ID	*CZNotes* Created	*CZNotes* Edited	Entities Added to *CZNotes*
P1	59	131	88
P2	1	0	0
P3	15	7	5
P4	28	0	0

Table 1. The number of CZNotes created, and edited per participant.

4.1.1 CZNotes and the Foraging Loop

In the beginning of the study, participants explored the data without a specific thread to follow, except for the contest's main goal. Participants used CZNotes similar to Pirolli and Card's description of a shoebox by creating "quick and dirty" CZNotes and adding any interesting data to them as references. Participants initially created CZNotes with a minimal amount of text, essentially using CZNotes as a container to tag the data, e.g., analogous to pasting sticky notes onto a paper document.

Participants created CZNotes primarily by attaching new CZNotes to documents and entities rather than creating a blank CZNote without any references (Figure 5). This may suggest that when foraging in the data participants wanted to create CZNotes related to a source. It seems that no new blank notes (green bars in Figure 5) were created in later analysis stages. Participants reported that CZNotes were primarily used as a "shoebox" during the early part of analysis.

The log indicates that CZNotes were primarily attached to documents when reading through documents in the Document View as shown in Table 2.

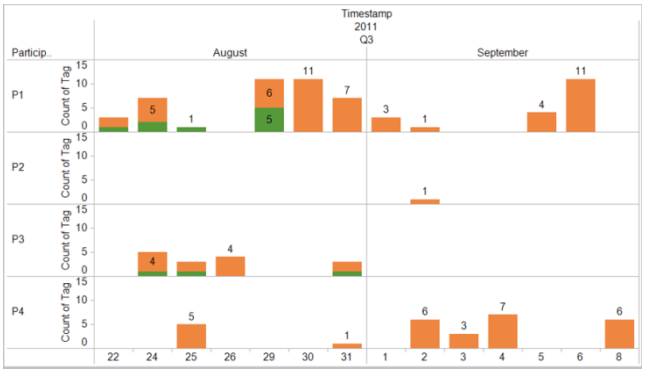

Figure 5 Total number of notes created over the course of study. Green bars: Creation of a new CZNote unlinked to any object. Orange bars: A CZNote was attached to a CZSaw object (a document or to another entity)

By replaying one of P4's scripts we observed the process for attaching CZNotes. P4 opened a Document View and added all the new article reports to it. He then scanned the reports and attached a CZNote to any interesting entities. He did not add these newly created CZNotes to a data view, suggesting he used the notes primarily for record keeping purposes.

Attach *CZNote* in:	Count
Document View	75
Hybrid Graph View	3
Semantic Zoom View	13

Table 2. A count of how many times CZNotes were attached to data in each of data views during the course of the study.

Creating too many CZNotes at this stage of analysis apparently was not a major concern. P4 stated he was "not too worried about useless notes because it's pretty easy to identify them and just not pay attention to them." He used CZNotes as a memory aid, knowing that his notes would be available later. P4 described using CZNotes as a form of external memory. The fact that participants were creating many CZNotes and linking them to the data would improve the chances that CZNotes would be returned in query results.

4.1.2 CZNotes Support for Continued Analysis

When continuing analysis from a previous session, participants P1, P3, and P4 stated that they did retrieve existing CZNotes created in previous sessions. One way for a participant to retrieve CZNotes was to add all CZNotes into a data view.

CZNotes were frequently used as a shortcut to retrieve data referenced in the CZNote. P3, for example, said he would sometimes use notes as a big pile of links, as a shortcut, to access all the documents related to a CZNote. There were multiple ways to do this but the following steps capture the general usage scenario:

1. Open relevant CZNotes in a Data View.
2. Retrieve the referenced entities or documents from the note and add them to the data views.
3. Continuing reading the documents (e.g. in the Document View).

A participant described that by using CZNotes as shortcuts, he avoided the need to re-do previous queries to retrieve data, i.e. CZNotes were used for retrieving data instead.

4.1.3 Building a Story with CZNotes

One way participants used *CZNotes* was to create notes as a way to store small chunks of information during the analysis. At some point however, they wanted to gather up all these disparate pieces of information and aggregate them. P4 said this was analogous to gathering all the documents one would need for a writing project. P3 said this was similar to organizing a set of sticky notes to compile something coherent. Ultimately, the users wanted to build a story from the data. Participants suggested that combining relevant *CZNotes* could help to build up a narrative or a story of what was happening in the data. They reviewed their *CZNotes* to create a summary to post on the wiki. P1 took this compilation approach by creating one large *CZNote* containing the summary of the whole story. This *CZNote* referenced other notes that it summarized.

4.1.4 Live Links

The connection between the *CZNotes* and the data is the key capability enabling the retrieval of data through the notes and the support of continued analysis using previously created *CZNotes*. All participants discussed the desirability of this live link between *CZNotes* and the data. P3 said it reminded him of hypertext links. P1 said this linkage between *CZNote* and data cannot be done with a wiki. These *live links* were not only useful when working with a participant's own *CZNotes* but also to retrieve data from *CZNotes* of another team member. P3 collaborated with P1 to access the data linked to P1's set of shared *CZNotes*. P3 stated that without these linkages from the *CZNotes* to the data, *CZNotes* would simply be a "big old box of words". Furthermore, P1 said you might as well create the notes outside of *CZSaw* if there were no such *live links*.

P3 described the *live links* feature as an important functionality to keep even during the final production process. He said, "When you're doing final production you need the ability to work with an image, a screenshot. You need to be able to do styling, basic text formatting. So, all your standard word processing stuff. But it would be best if those features were in conjunction with all the power of *CZNote* provides—the ability to get links and stuff like that."

Interestingly, *live links* were used also as a filtering tool by P1. P1 began collecting the (irrelevant) sports score articles into one *CZNote* by adding them as references to the note. When working with the *Semantic Zoom View* to analyze the entire data set, he wanted to filter out irrelevant articles such as sports scores. While the *SZV* had a filter function, there was no easy method to select all the irrelevant articles. Therefore, he opened the *CZNote* with the irrelevant articles in a *Hybrid Graph View* and selected all the irrelevant articles contained within. This selection of the articles also triggered a selection in the *SZV*. With the irrelevant articles now selected in the *SZV*, he could use the *SZV*'s "filter" functionality to remove them from the view and continue with his analysis.

4.2 CZNote Reminders

When iteratively performing data queries, it might be beneficial to be reminded of previous analytical findings by using CZNotes. The log file data showed that 36% of the 162 queries (made by all participants using the search panel returned CZNotes, 15% (24 out of 162) contained a mixture of CZNotes and other Entity Types, and 21% (34 out of 162) contained only CZNotes. This shows that CZNotes were, in fact, being "fed back" into subsequent query results to potentially support iterative analysis.

We asked participants if they noticed CZNotes being returned as part of the results of a query. Half of the participants ignored the CZNotes that were retrieved this way, while the other half took steps to filter out CZNotes when they were unwanted. Since CZNotes are a distinct color, they can easily be ignored.

P4 mentioned that seeing his CZNotes in a search result reminded him that he had gone down this path before and not to pursue it any further. P4 commented he would want notes to provide context.

4.3 Collaboration and Sharing *CZNotes*

Synchronous collaboration support was not a primary goal in *CZNote*'s initial design, since CZSaw supports asynchronous collaboration by the sharing of scripts. Participants used various techniques to share *CZNotes* with one another by (1) sharing, via a wiki, the script file containing the note creation commands and/or (2) posting notes' content directly on the wiki.

It was clear from their comments and actions that participants wished to share analytical artifacts despite the current barriers to doing so in *CZSaw* and *CZNotes*. *CZSaw* does not provide a centrally shared online repository of data and *CZNotes*. However, all user interactions are recorded in a script file and this file can be given to other team members. This shared script file can be re-executed within any *CZSaw* instance and will import any *CZNotes* created by collaborators. Unfortunately, since there is currently no automated way of merging two scripts into one, collaborative activities require manual merging and managing of various scripts from collaborators.

4.3.1 Wiki versus CZNotes

At the beginning of the study, participants initiated the use a *CZSaw* research group wiki to share their analytical findings during the course of their analysis. They foresaw the need to share their findings in a central and easily accessible location and understood the limitations of *CZSaw* and *CZNotes* collaboration support. We encouraged, but did not require, that participants first write any notes in a *CZNote* before posting them onto the wiki.

Participants also anticipated that wiki postings could be used for different purposes compared to *CZNotes*. One difference was that *CZNotes* are initial findings and collections of synthesized knowledge obtained during data exploration, whereas the wiki posts might be a distillation of multiple *CZNotes*. Participants' feedback indicated wiki postings were in fact a distillation of *CZNotes*. Most participants would create a summary of their *CZNotes*, rather than posting all their *CZNotes*, and share the summary on the wiki. P1 said he created a "super-note to summarize what was found and described the documents in order by dates. That note's content is basically copied to the wiki page exactly as it appears in the actual note."

4.3.2 Pursuing Different Threads

All the participants agreed to use a divide and conquer approach to pursue different threads of analysis as they emerged. They used a portion of the focus group session time to discuss possible threads and assign them to team members. Participants created one page per thread on the wiki to share their findings with others. The wiki offered a way for participants to setup different pages for each of these threads. They discussed that if one participant found information about another participant's thread; they could post on that thread's wiki page. We noticed a similar behaviour in using *CZNotes* individually as a categorizing tool (see "*CZNotes* as a Categorizing Tool" in section 4.4).

4.3.3 Live Links and their Need

While the Wiki had the advantage of being a central share of information and easily accessible, it did not provide the "live link" capabilities of *CZNotes*. Participants felt *CZNotes* had the advantage of linking to data and other notes as described in section 4.1.4. P1 said, "... the wiki's not live. It won't live connect you to the entities in *CZSaw*, but like I've just started a new project last night and so I have notes in two different places now. So I would need to manually put those scripts together if I wanted to keep all the notes. So I'll probably transfer the notes from both of these two sets of projects to the wiki and be able to see them in there. Yeah you lose stuff because you can't live link back but ...". P3 said that " The ability to add references to notes were the most used feature in my analysis. I used notes to reference

documents of importance." The ability to use *CZNotes* as a link to the data was important enough for the participant's to spend time to export their notes into a script file and store it on the wiki to share with the rest of the team.

4.3.4 Sharing CZNote Scripts

Participants encouraged one another to upload *CZNote* script files to the wiki along with the notes posted on the wiki. The desire to share the script file containing the *CZNote* creation commands was so that collaborating participants could use the *CZNotes* live in their own running instance of *CZSaw*.

To utilize the live links of *CZNotes*, P1 shared his script containing only *CZNote* creation commands with P3 (by uploading the script to the wiki) for P3 to download into his *CZSaw*. P1 did this by hand editing the CZSaw script. P3 found this quite valuable, but in future work this would be automated.

4.4 *CZNotes* as a Categorizing Tool

Participants also described using *CZNotes* as a "shoebox", i.e., as a container for categorizing interesting data found during analysis. Using *CZNote*'s entity referencing capabilities, participants were able, for example, to collect all "noise" documents into a single *CZNote*, as P1 did with sports scores, as mentioned in section 4.1.4

Another viewpoint is to think of *CZNotes* as "tagging" items put into them, rather than as a container. Participants used the term tagging when describing the action of adding a reference to an entity or document in a *CZNote*.

The primary way participants used *CZNotes* to categorize information during this study was by thread of analysis. P4 described this working style of having one *CZNote* per thread of analysis and wanted to continually add more documents and information to each of them as he finds more relevant information. Similarly, P1 said, "… I was making a note for each of the suspicious activities or interesting things and then I made a super note to summarize what I found and described the documents in order by dates."

The Semantic Zoom View also had a grouping feature that is specific to the view. However, P1 describes that he preferred to create such categorizations of the data with *CZNotes* and their references rather than with the *SZV* for efficiency considerations.

4.5 Summary

The data collected in this preliminary study suggest that integrating *CZNotes* with *CZSaw* data was not only plausible but demonstrated several interesting uses of *CZNotes*.

Following is a summary of the different *CZNotes* use cases identified by the performance of the participants in this study.

1. CZNotes were used as a categorizing tool. Relevant data and other CZNotes regarding a particular thread of analysis were collected within a single CZNote.
2. Similarly CZNotes were used as a filtering tool in circumstances where forming a comprehensive query would be labor intensive (eg. sports scores).
3. CZNotes served as a shortcut to retrieve data referenced in a note. This was done primarily at the beginning of a subsequent analysis session.
4. CZNotes were used as a bridge to link existing data and synthesized knowledge into a single knowledge element. CZNotes supported the creation of new relationships among the original data, as well as relationships between the data and other CZNotes.
5. CZNotes enabled users to link entities they thought were the same, i.e., provided a way to merge or link entities.
6. For one participant, seeing existing CZNotes in his search results served as a reminder that the search was done before and triggered him to pursue a different analysis path.

5. CONCLUSIONS

To recap, the purpose of this study was to examine the possible benefits of integrating analyst's' notes into the data set used by an entity-based visual analytics tool, CZSaw. We have taken a novel approach that integrates notes and data by representing notes the same way as the data in CZSaw.

To recap our questions from section 4,

Would CZNotes Remind?

CZNotes could be seen in query results and, for one participant, finding notes in these results unexpectedly reminded him that his current query was linked to a previous dead end analysis path. This prompted him to pursue another analysis path. Participants hypothesized that feeding CZNotes back into the data set could potentially be more useful if the duration of the analysis was longer (i.e., more than 3 weeks).

Would CZNotes enable continuation?

The three participants who created CZNotes reused them repeatedly in subsequent sessions.

Would CZNotes enable new relationships to be created?

Most emphatically yes. CZNotes enabled tagged foraging (4.1.1), enabled the components of story building (4.1.3), enabled the pursuit of different threads (4.3.2), and enabled categorization (4.4).

Would CZNotes enable the introduction of new data?

The CZNotes capability was there, but it remained unused in this experiment. The drawback of the study is that the problem is fictional, so drawing in new information from elsewhere is impossible.

5.1 Limitations

The obvious limitation of this case study is that the participants were CZSaw team members, and thus had a vested interest in making the tool work. Our goals in reporting this work are therefore limited to describing the expressive capabilities of CZNotes when applied to a realistic problem over a significant length of time. We do not report preference information or comparisons to other tools, since this is a design study that is meant to understand the potential capabilities of this method of integrating analysts' notes into the analysis process and into the data.

REFERENCES

[1] P. Pirolli and S. K. A. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of International Conference on Intelligence Analysis*, 2005.

[2] Y. Kang and J. Stasko, "Characterizing the Intelligence Analysis Process: Informing Visual Analytics Design through a Longitudinal Field Study," *IEEE Symposium on Visual Analytics Science and Technology,* pp. 19-28, 2011.

[3] C. Marshall, "WORK PRACTICES STUDY : ANALYSTS AND NOTETAKING," 1990. [Online]. Available: http://www.csdl.tamu.edu/~marshall/oswrstudy.pdf.

[4] N. Mahyar, A. Sarvghad and M. Tory, "A closer look at note taking in the co-located collaborative visual analytics process.," *IEEE VAST,* pp. 171-178, 2010.

[5] C. Gorg, Z. Liu, N. Parekh, K. Singhal and J. Stasko, "Visual analytics with Jigsaw," *IEEE Symposium on Visual Analytics Science and Technology,* pp. 201-202, 2007.

[6] P. R. Quinlan, C. Reed and A. Thompson, "INSPIRE: An Integrated Agent Based System for Hypothesis Generation within Cancer Datasets," *Web Intelligence and Intelligent Agent Technology,* vol. 3, pp. 587-590, 2008.

[7] Y. B. Shrinivasan and J. J. Van Wijk, "Supporting the analytical reasoning process in information visualization," *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08,* pp. 1237-1246, 2008.

[8] R. J. J. Heuer, Psychology of Intelligence Analysis, Government Printing Office, 1999.

[9] W. Wright, D. Schroh, P. Proulx, A. Skaburskis and B. Cort, "The Sandbox for analysis: concepts and methods," *Proceedings of the SIGCHI conference on Human Factors in computing systems,* pp. 801-810, 2006.

[10] E. A. Bier, E. Ishak and E. Chi, "Entity Workspace: An Evidence File That Aids Memory, Inference, and Reading," *IEEE Symposium on Visual Analytics Science and Technology,* p. 466–472, 2006.

[11] D. Gotz, "The ScratchPad: sensemaking support for the web," *Proc. of the Inter. WWW Conf. Posters,* pp. 1329-1330, 2007.

[12] Y. Chen, S. Barlowe and J. Yang, "Click2Annotate: Automated Insight Externalization with rich semantics," *IEEE Symposium on Visual Analytics Science and Technology,* pp. 155-162, 2010.

[13] D. Gotz, M. Zhou and V. Aggarwal, "Interactive Visual Synthesis of Analytic Knowledge," *IEEE Symposium On Visual Analytics And Technology,* pp. 51-58, 2006.

[14] "IEEE VAST Challenge 2011.," 2011. [Online]. Available: http://hcil.cs.umd.edu/localphp/hcil/vast11/. [Accessed 15 September 2012].

[15] P. Isenberg, A. Tang and S. Carpendale, "An exploratory study of visual information analysis," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 1217-1226, 2008.

[16] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner and J. Thomas, "IN-SPIRE," in *InfoVis 2004 Contest Entry*, 2004.

[17] Y. Chen, J. Yang, S. Barlowe and D. H. Jeong, "Touch2Annotate: generating better annotations with less human effort on multi-touch interfaces.," in *CHI extended abstracts'10*, 2010.

[18] Z. Liu, C. Gorg, J. Kihm, H. Lee, J. Choo, H. Park and J. Stasko, "Data ingestion and evidence marshalling in Jigsaw (VAST 2010 Mini Challenge 1 award: Good support for data ingest)," in *IEEE Symposium on Visual Analytics Science and Technology*, 2010.

[19] C. M. Canfield and D. Sheffield, "Interactive data analysis with nSpace2®," *IEEE Conference on Visual Analytics Science and Technology (VAST),* pp. 327-328, 2011.

[20] D. Jonker, W. Wright, D. Schroh, P. Proulx and B. Cort, "Information Triage with TRIST," in *International Conference on Intelligence Analysis*, 2005.

[21] J. Heer, F. B. Viégas and M. Wattenberg, "Voyagers and voyeurs," *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07,* pp. 1029-1038, 2007.

[22] W. Willett, J. Heer, J. Hellerstein and M. Agrawala, "CommentSpace: structured support for collaborative visual analysis," *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11,* pp. 3131-3140, 2011.

[23] W. Cheng and D. Gotz, "Context-based page unit recommendation for web-based sensemaking tasks," *Proceedings of the 13th international conference on Intelligent user interfaces,* pp. 107-116, 2009.

[24] A. Endert, P. Fiaux and C. North, "Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering," *IEEE Transactions on Visualization and Computer Graphics,* pp. 2879-2888, 2012.

[25] A. Endert, P. Fiaux and C. North, "Semantic interaction for visual text analytics.," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 473-482, 2012.

[26] 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,* pp. 131-138, 2009.

[27] J. Stasko, C. Görg and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization,* 7(2), pp. 118-132, 2008.

[28] E. A. Bier, S. K. Card and J. W. Bodnar, "Principles and tools for collaborative entity-based intelligence analysis," *Intelligence and Security Informatics,* pp. 178-191, 2010.

[29] D. Dunsmuir, E. Lee, C. D. Shaw, M. Stone, R. Woodbury and J. Dill, "A Focus + Context Technique for Visualizing a Document Collection," *45th Hawaii International Conference on System Sciences,* pp. 1835-1844, 2012.

[30] R. J. Sternberg, Cognitive Psychology. 2nd. Ed., Harcourt Brace College Publishers, 1996.

[31] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: Multi-dimensional In-depth Long-term Case," *Proceedings of the 2006 AVI workshop on BEyond time and errors novel evaluation methods for information visualization - BELIV'06,* 2006.