Supporting hypotheses management during asynchronous collaboration for visual analytics for text

Ankit Gupta, Chris Shaw
School of Interactive Arts & Technology
Simon Fraser University, Surrey, BC, Canada

Abstract

During analysis of large collections of text documents in a collaborative environment, analysts divide their work to save time and reduce duplication of effort. Based on their reading and analysis, each analyst forms multiple hypotheses. For any given hypotheses, the analysts present arguments and counter-arguments to accept or reject it. Managing these hypotheses and associated discussion can be challenging. In this paper, we present a tool that helps analysts by integrating hypotheses management into the analysis process. The tool we present is designed to support asynchronous collaboration.

Introduction

Professional analysts such as intelligence analysts and journalists are constantly required to make sense of large document collections consisting of reports, telephone intercepts, speeches, blogs and newspaper articles. They find new information and gain new insights. The process of evolution of insights consists of analysts providing arguments and counter-arguments.

In the research domain of visual analytics, several software products have been developed to support analytical reasoning with interactive visual interfaces. Existing tools such as Jigsaw[15], CZSaw[11] and IN-Spire™[1] focus on individual analysis. Software designed to support individual analysis are designed to provide analytic capability to a single analyst. They do not focus on the collaboration that takes places among multiple analysts working together. Other projects that include collaboration in their design, such as Cambiera[9] and E-Wall[12] focus on collocated and synchronous remote collaboration. In collocated setting, collaborators work at the same time and are located in the same space. They will typically share the device/tool being used for the analysis process. In synchronous remote settings, the collaborators are located remotely and work on a shared visual workspace. These are powerful tools but they do not allow collaborators to work at their own time and place.

Existing software applications for visual analytics do not concentrate on supporting the "collaborative reasoning" aspect of analysis. Software such as sense.us[8] provide a standard comment-reply system, often with ability to link visualizations. In theory, the comment-reply system can be used by collaborators to discuss a hypotheses. However, making sense of a stream of comments is hard. It is difficult to get a sense of whether or not the main argument being discussed is supported or not. In this paper, we present a system that integrates the management and discussion around hypotheses using visual techniques to provide better sense-making of the reasoning process.

Related Work

The visual analytics agenda[16] provides several recommendations/areas of focus for visual analytics tools. In this paper we focus on asynchronous collaboration and support for reasoning.

Collaboration

Researchers have explored asynchronous and synchronous collaboration in information visualization and visual analytics environments. Hajizadeh et al.[6], for example studied brushing techniques for providing awareness in a synchronous remote setting on tabular data. They compared three brushing techniques (brushing and linking, selection and persistent selection) for providing awareness. In their research, they identified awareness as the ability of collaborators to understand the brushing actions taken by their remote collaborators. They studied these techniques using a collaborative visualization of tabular data where two collaborators shared a visualization workspace. The results of the study indicated that persistent selections in which users saw their collaborators previous as well as current selection provided most awareness among the three techniques.

In another work Isenberg et al.[10], studied the use of a collaborative awareness technique called "collaborative brushing and linking" in which the collaborators are aware of each other's selection via brushing and linking. In their exploratory study, they studied a system called Cambiera[9]. Cambiera, a visual analysis tool for text documents is designed for collocated collaboration. They found that Cambiera's implementation of collaborative brushing and linking provided awareness in a collocated collaborative environment. Collaborative brushing and linking, however is not sufficient in asynchronous environments as analysts do not necessarily work simultaneously.

Sense.us[8] is another tool for collaborative visualization. The tool is designed to support asynchronous collaboration during analysis of tabular data. Heer et al. studied the sense.us to provide design recommendations for encouraging social interaction in asynchronous collaborative visualization. They recognize awareness and provenance[1] as important goals of asynchronous collaborative visualization. In another paper, Heer et. al.[7] provide design recommendations for collaborative visual analytics. These design considerations are grouped into seven topical areas: division and allocation of work; common ground and awareness; reference and deixis; identity, trust and reputation; group dynamics; consensus and decision making. While all considerations may not be applicable for a single tool, these considerations suggested

[1]The authors do not use this word. However, their notion of doubly linked discussions is similar to provenance.

in this paper are useful for researchers focusing on different aspects of collaboration in visual analytic environments.

Chen et. al[4] have focused on common ground construction in asynchronous collaborative visual analytics. They present a platform called ManyInsights that allows users to record their insights, and provides views to support common ground construction. These views provide overview of the insights and provide users with content and other insights related to a given insight. Thus, ManyInsights allows users to explore large number of evolving insights. One major difference between our tool and ManyInsights is that we focus on a team of analysts working on the same task, whereas ManyInsights is intented for users not necessarily working as a team. Instead, ManyInsights focuses on helping a user find insights made previously by other analysts on same or related datasets.

Support for Sensemaking

There are several visual analytics tools that support the reasoning process explicitly or implicitly. In this discussion, we do not review systems that provide Artificial Intelligence or Machine Learning enabled support for automated reasoning. We limit our discussion to systems that provide support reasoning by human analysts[14, 13].

Shrinivasan et al.[14] implemented a system called Aruvi, which was designed for individual analysis. In Aruvi, the authors implemented a view called knowledge view. The knowledge view in Aruvi allowed an analysts to create a graph of their annotations.

Sanfillipino et. al. implemented a hypothesis space in IN-Spire^{TM}[1]. The hypothesis space was designed to allow an analyst to form hypotheses. The hypothesis space however provided sensemaking only for a single analyst.

Entity-based Analysis for Text

Entity-based analysis is a kind of text analytics in which the goal of the analyst is to find important entities in a document collection and to find relationship between different named entities like person, location, organization etc. Our system is designed to support entity-based text analytics.

Entity Workspace[3] was one of first VA tools to be developed to support entity-based analysis. Entity Workspace provided automatic extraction of named entities using Natural Language Processing (NLP) algorithms. It also allowed the analysts to find important entities and find entities related to any given entity. The authors showed how entity-extraction allowed the analysts in finding documents that are important for analysis.

Another important tool is Jigsaw[15] which also provided entity-based collaboration. However, Jigsaw added several visualizations for visualization of relationships among entities as well as several NLP algorithms including sentiment analysis. Jigsaw allowed an analyst to look at the importance of an entity (based on its frequency) using a list view in which entity names were displayed with a bar to indicate the frequency of the entity. In addition, the list view also allowed the analyst to visualize related entities. The related entities were found based on the concept of bibliographic coupling. According to this concept, two entities are considered related if they appear in at least one document.

CZSaw[11], in addition to providing entity-based collaboration and entity-based visualizations similar to Jigsaw, provided a history mechanism, an editable script and a dependency graph. The history view allowed the analyst to track his actions and jump back to any previous state. CZSaw also captured user actions and inputs in the form of an editable script called CZScript. It allowed the analyst to revise his analysis by changing the inputs at some previous point in the script. On changing the inputs in the script, CZSaw would perform the analysis with new inputs without requiring the analyst to manually performing the intermediate steps. A dependency graph provided the analyst with a visualization of the several dependencies among different components of the analysis.

Existing tools have been designed either to support collaboration, particularly in collocated and synchronous remote settings or to support the reasoning process. However, none of the existing tools support both. This presents us with an opportunity for developing a system that does both and to study collaboration during hypothesis formation using this system.

Design Principles

To support analytical reasoning process during asynchronous collaboration in a visual analytic environment we identified the following design principles: *Awareness*, *Collaborative Hypothesis Formation & Evaluation* and *Provenance*.

Awareness

Awareness of collaborators' activities is an important element of collaboration. Information about awareness helps an analyst in assessing what has been done and where more effort needs to be put. By awareness, we mean allowing a collaborator in a distributed team of analysts to know what has happened since they last logged. In addition, the analysts should be aware of when the new information has been added and who are the contributors of the new information.

Collaborative Hypothesis Formation & Evaluation

Based on the requirement to enable collaborative sensemaking, we consider hypothesis formation and evaluation as one of the principles while designing our tool. In asynchronous collaborative settings, analysts would work individually to find new information. This new information is used by all team members to validate or reject a hypothesis. Incorporating this principle, requires a shared hypothesis space where analyst can collaboratively make sense of information.

Provenance

As discussed in the previous section, we want to support collaborative hypothesis formation. When analysts look at new information created by other analysts, they might often feel the need to look at the source of the evidence. We assume this based on personal experience while working on VAST challenges and the recommendations given by Thomas and Cook [16]. We designed our tool keeping this requirement as another important principle in our design. When looking at a piece of information, we want to allow the analyst to jump back to the source of the information as well as know about who contributed the new piece of information.lo

Design Evolution

Our goal was to support text analysis and the associated hypotheses management in an asynchronous collaboration

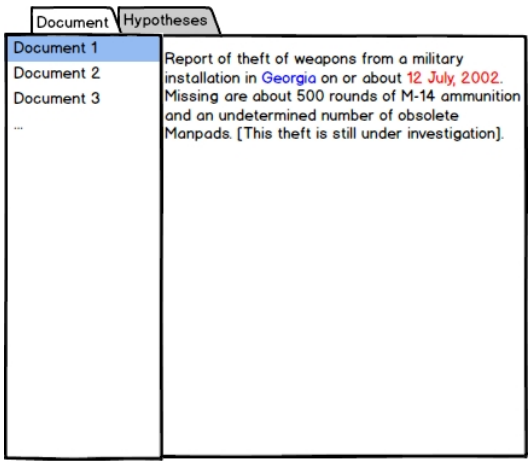

Figure 1. *A wireframe showing the initial design of the document view.*

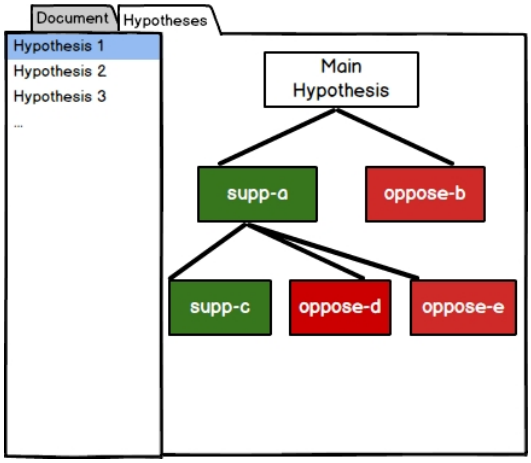

Figure 2. *A wireframe showing the initial design of the hypotheses view.*

setting. For this purpose, we identified two primary visual workspaces as a requirement. First, a document view that allows an analyst to read document, view the entities contained within the documents and add annotations. Second, a visual space for forming hypothesis and evaluating them.

Figure 1 and 2 show the initial wireframe designs of the two views (created using myBalsamiq). Figure 1 shows the document view with a list of documents and a text area for showing text of the selected document and the contained entities highlighted. Figure 2 shows the initial design of the Hypothesis view. It contains a list of hypotheses and a visual space for showing a hypothesis as the root node of a tree and the supporting arguments and counter-arguments as intermediate and leaf nodes. These initial designs were then evaluated using paper-prototypes. The result of this activity resulted in evolution of the two views into their current state, as we discuss them below.

Document View

For the document view, we included entity highlights to allow the analyst to quickly make sense of what the document is about. The initial design of the document view, however, was limited. While evaluating the design using paper-prototype, we quickly found new requirements. We found the need for free-text search and, a way for analysts to avoid duplication of effort. In collaborative environment, analysts will need to be aware of the documents that have already been analyzed by other analysts. In addition, we also found a need to be aware of important documents. In our case, we use the number of annotations in a document to be a measure of its importance. More complex algorithms can be used to find documents that are more important than others. However, that is out of the scope of this paper. Another requirement we found was that when reading an already read document, we wanted to be aware of who added what annotations, similar to what a word processor like Microsoft Word or Google docs provides.

Based on these new requirements, we designed the document view as shown in figure 3.

It consists of a list of documents and a search box to filter the list by their title (figure 3a). In the document list, the title of the documents are prefixed with a small colored rectangle (see figure 3e). In the figure the documents **CIA03.txt** and **ArmyCID01.txt** have red rectangles as a prefix. The color of the rectangle indicates the number of annotations that were added to the document by any analyst such that more saturation indicates more annotations. This helps the analyst quickly identify the documents that have contain a large number of annotations and those that do not. In this case **CIA03.txt** contain more notes than **ArmyCID01.txt**. An analyst can use this information to decide that he needs to annotate other un-analyzed documents or he/she may decide to take a look at the documents that are highly annotated to find out the important pieces of information in those documents.

Figure 3b shows the content area of the document view. The content area highlights the entities within the text. The entities are extracted using Named Entity Recognition algorithm[5]. Each color represents a different type of named entity such as location, organization etc. We use opacity of the color to represent the importance of an entity within a document collection. An darker highlight (i.e. higher opacity) indicates that the entity has been mentioned by more documents as compared to an entity with lower opacity. This helps the analyst in searching the document collection based on entities that are important. In the content view an analyst can select a piece of text and add a note to it. When the analyst adds a note, the annotated text gets underlined with the color corresponding to the analyst. Here one can see that the word "manpads" was annotated by Joe (joe@example.com). In addition, the analyst can find all documents containing a highlighted entity by just clicking on that entity. This helps the analyst in filtering the documents that are of interest at a given time. While this feature does not help collaboration, it is an important part of the process as it helps the analyst in quickly filtering the documents and what new piece of information can he/she get about a particular entity of interest.

Figure 3c shows the analysts collaborating on the project and the notes that are added to the current document. There is a single annotation "Manpads theft indicates a possible air attack.". The color of the annotation corresponds to the color assigned to the analyst who created the annotation.

The notes also show the date and time of creation of a note. This is an important piece of information for providing awareness about time. As analysts analyze more and more documents, their understanding about the document collection grows. This means

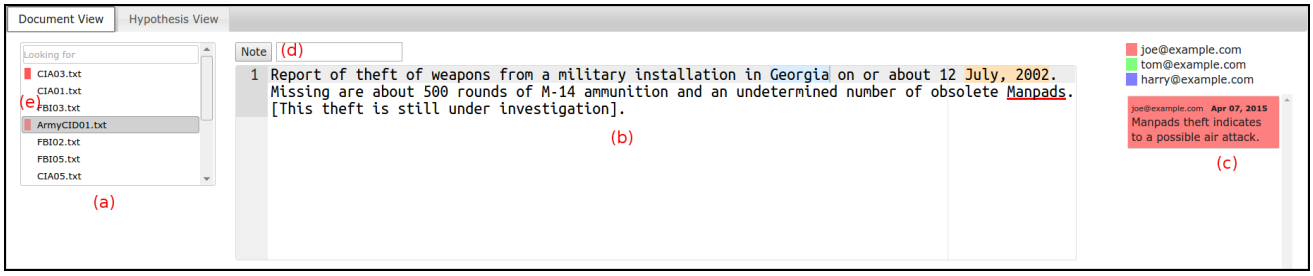

Figure 3. *The Document View. (a)Document List; (b) Content View; (c) Legend and Notes*

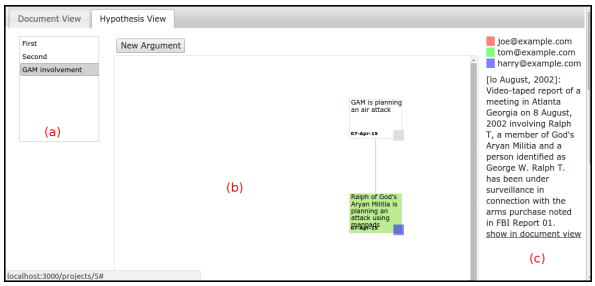

Figure 4. *Hypothesis view with two argument maps (only one visible in the image). (a) The list of arguments and (b) the argument map visualization*

that the annotations that were made during the initial stages of analysis might no longer be relevant. An analyst can look at the date of the annotation created by another analyst and decide on its importance by looking at its creation date.

Hypothesis View

While designing the hypothesis view, the visual structure to use for the hypothesis and its arguments, which we call hypothesis tree, was important. One alternative to the current design was to use node-link diagram like the knowledge-view in ARUVI[14]. In the knowledge view of ARUVI, an analyst can organize annotations by assigning them colors and grouping them. While the knowledge-view structure is very powerful for individual analysis and sense-making, we find it unsuitable for collaborative reasoning during asynchronous collaboration. It is important that a simple visual structure is used to improve understanding of an hypothesis and its arguments. A random structure like that of node-link diagram can cause be hard to read. More importantly, as we find a way to replace the comment-reply thread with visual structure, our design is closer to the comment-reply structure than the node-link diagram. In an asynchronous environment, keeping track of changes can be hard. Therefore, we do not allow users to edit nodes of the hypothesis view. Instead, collaborators can add new nodes to support or oppose an idea or insight presented in a given node. The tree structure is controlled by the collaborators. A collaborator choses what node to support or oppose and interacts with the visualization accordingly.

In the initial design of figure 2, the nodes of the hypothesis tree linked to the annotations created by analysts. Clicking these nodes will result in the browser navigating to the corresponding document. While evaluating the paper-prototype we realized that instead, it was also important to view not just the annotation but the text that was annotated and the corresponding document, while viewing the argument. Going back and forth between the hypothesis view and document view makes it hard for the analyst to remember the structure of the argument and the role an annotation plays within the whole hypothesis tree. In addition, the nodes of the hypothesis tree required the same temporal information as an annotation in the document view.

The timestamps for a node and the corresponding source document are important as they provide awareness and provenance. During asynchronous collaboration, an analyst looking at a concept map will want to know who added a particular annotation. By looking at the creation date and time, the analyst can visualize how an argument progressed. Finally, the ability to go back to the source of a note helps the analyst in determining whether or not to trust a piece of information.

Figure 4 shows the final implementation of the hypothesis view. Figure 4a shows the list of hypothesis created by any of the collaborating analysts. Figure 4b shows a small argument map. In the argument map, analysts find evidence that supports or rejects a given hypothesis and add it to the map. The background color of any node in the argument indicates whether the node is supporting or opposing the hypothesis. Green nodes are in support of the hypothesis and the red nodes are opposing the hypothesis. In addition to the text of the source note, a node in the argument map also contains information about its creation time as well as its author. In figure 4b, the creation time of the note can be seen on the bottom left corner of the nodes. The bottom right corner of the node, contains a small square icon. The color of the icon indicates the analyst who added the given node. The analyst for a given color can be seen in the analyst legend (see figure 4c) just like the document view. When the user clicks on this icon, the text of the source document gets displayed in the right side bar of the hypothesis view. An analyst can then click on the link “show in document view” to see the contents of the document with the entities highlighted and all the contained notes, as described in the document view. In the figure 4b an analyst Harry(harry@example.com) has added evidence supporting the hypothesis.

Usage Scenario using VAST 2010 dataset

As a formal evaluation of the system needs to be done, we show instead an example scenario to explain how we expect the tool to be used. In this section, we demonstrate how analysts can use our tool for collaborative reasoning while exploring the VAST 2010 dataset[2] about arms dealing. This dataset contains several documents about events related to arms and weapons smuggling around the world. The goal of the analysis is to find connections between different locations and people who are involved in the illegal arms dealings.

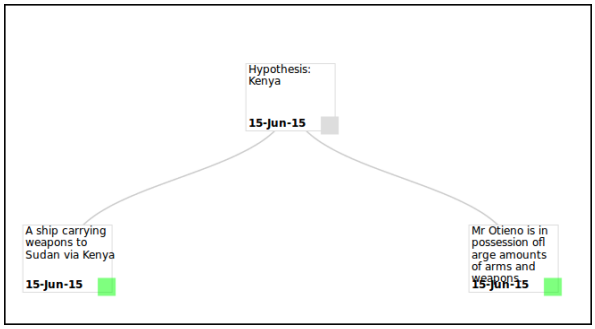

Figure 5. *Summary for Kenya*

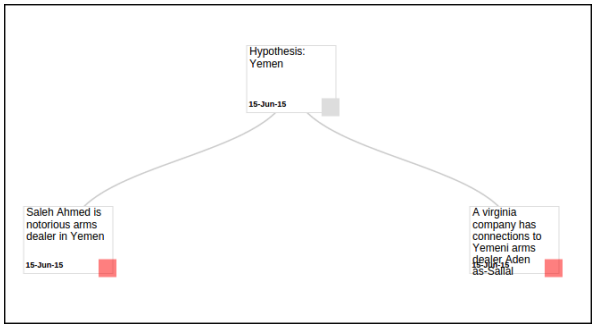

Figure 6. *Summary for Yemen*

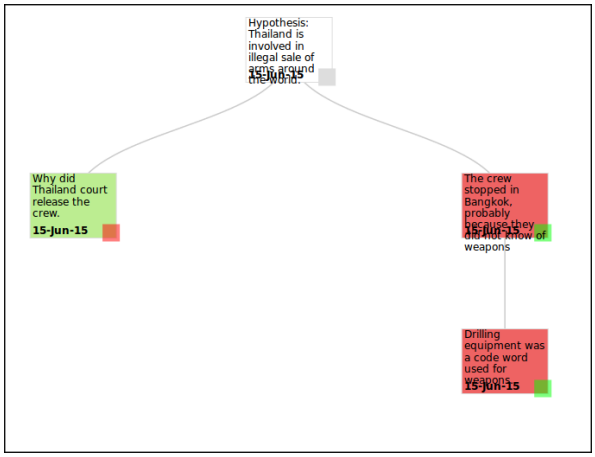

Figure 7. *Alice and Bob argue about involvement of Thailand*

.

Consider a team of two analysts, Alice and Bob. We will show how they use our tool for finding connections between different entities. Since the complete analysis of the document collection is out of scope of this paper, we will show part of their analysis.

As Alice and Bob know that they are investigating documents about arms and weapons dealings, they decide to divide their work. Alice and Bob analyse documents containing "arms" and "weapons" respectively. Alice finds the documents that mention the word "arms" using the tools full-text search. While reading the documents, she finds a report about a plane carrying 35 tonnes of weapons from North Korea, that stopped in Bangkok for refueling. The article contained a concern about the plane stopping in Bangkok despite safer options elsewhere. She makes a hypothesis that either the officials or someone else in Bangkok is involved in illegal smuggling of arms. She continues analyzing other "arms" documents and adds notes about Kenya which she or Bob can later find by just searching for "Kenya". While Alice was reading "arms" documents, Bob read "weapons" documents. Figures 5 and 6 show the information Alice and Bob collected for Kenya and Yemen respectively[2]. They were able to share their finding using the hypothesis view. Note, in this case the analysts did not use the visual structure for argumentation, but for knowledge sharing.

Bob then notices that Alice has created a hypothesis about involvement of Thailand in illegal arms dealings. Bob Alice then find new evidence and collaboratively construct the argument. Figure 7 shows the hypothesis view containing an argument map about involvement of Thailand in illegal arms trade. Alice provides acquittal of the crew as an evidence in support of the argument. Bob does not agree with Alice's assessment. He finds that the crew believed that they were carrying oil drilling equipment and that might be the reason that they landed in Bangkok, despite other safer options. To further support his argument, Bob does a full-text search to find if there is any mention of drilling equipments and find a document where "drilling equipment" was used as a code-word for arms and weapons. He annotates the document and adds it as further evidence to oppose the argument that "Thailand is involved in illegal arms trade". In manner, the analysts can continue their argumentation by providing supporting or opposing evidence from documents or freely add new notes to include information external to the document collection.

Discussion

In this paper, we present a tool to support asynchronous collaboration in analytical reasoning in a visual analytic environment. The novelty of the tool lies in the hypothesis view. The hypothesis view allows collaborators to use insights from multiple documents to argue about a hypothesis and also be aware of the evolution of the hypothesis.

Before we begin any discussion, it is important to distinguish our implementation from tools like Google Docs and Microsoft Word. These tools provide annotation capabilities that might look same as that present in our tool. However, our annotation support is different. In case of tools like MS Word and Google Docs, an author makes an annotation and others reply to that annotation, thereby creating a comment-reply thread with every annotation. Our approach differs at least in two areas. First, our annotations are not a comment-reply chain. Instead, they are simple annotations that analysts can use to store their insights based on a particular section of a document. We allow users to view and organize annotations made on multiple documents in the hypothesis view. Second, the annotation systems of these tools are not focused towards sharing insights. They are focused more towards making remarks about a section of text or adding todos. In our case, the annotations are not local to a document region. Instead, they become part of the reasoning process using hypothesis view.

One limitation of the tool lies in the visual encoding for the collaborators. Use of color limits the number of easily distin-

[2]We only show information about two of several countries to keep the discussion simple

guishable users. This limits the scalability of the tool. Another limitation that limits the scalability of the tool is the current design of the argument maps. A normal desktop or laptop screen does not provide enough screen space for large argument maps without the need to scroll. This design needs to be improved to better use screen space. While limiting scalability of argument size, the argument map is still useful in supporting the reasoning among collaborators.

Heer et. al.[7] mention identity, trust and reputation as an important consideration while designing collaborative VA tools. We need to improve the design of this tool to incorporate these considerations as well. As mentioned before, the annotations cannot be edited. We made this decision to ensure that an argument does not get invalid because a user changes an annotation. This limitation can result in argument map growing with nodes to make correction. This needs to be addressed better.

Conclusion

In this paper, we presented a system for supporting hypothesis formation as part of the collaborative reasoning process during text analytics. The system we presented consists of a document view to read and annotate documents with insights and a hypothesis view to organize insights to argue about a hypothesis. The novelty of our tool lies in the hypothesis view which allows collaborators an easy way to organize insights from multiple documents in a visual analytic environment for text. In future, we plan to address some of the limitations mentioned above and conduct a formal evaluation of the system.

References

[1] IN-SPIRE^{TM}. `http://in-spire.pnnl.gov/`.

[2] IEEE VAST 2010 Challenge, 2010.

[3] Eric A Bier, Edward W Ishak, and Ed Chi. Entity Workspace : an evidence file that aids memory, inference, and reading. In Sharad Mehrotra, Daniel D. Zeng, Hsinchun Chen, Bhavani Thuraisingham, and Fei-Yue Wang, editors, *Intelligence and Security Informatics*, pages 466–472. Springer Berlin Heidelberg, 2006.

[4] Yang Chen, J. Alsakran, S. Barlowe, Jing Yang, and Ye Zhao. Supporting effective common ground construction in asynchronous collaborative visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 101–110, Oct 2011.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[6] Amir Hossein Hajizadeh, Melanie Tory, and Rock Leung. Supporting awareness through collaborative brushing and linking of tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2189–2197, 2013.

[7] Jeffrey Heer and Maneesh Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, February 2008.

[8] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. Voyagers and Voyeurs : Supporting Asynchronous Collaborative Visualization. (April):87–97, 2007.

[9] Petra Isenberg and Danyel Fisher. Cambiera: Collaborative Tabletop Visual Analytics. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 581–582, New York, NY, USA, 2011. ACM.

[10] Petra Isenberg, Danyel Fisher, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. An exploratory study of co-located collaborative visual analytics around a tabletop display. *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 179–186, October 2010.

[11] N Kadivar, V Chen, D Dunsmuir, E Lee, C Qian, J Dill, C Shaw, and R Woodbury. Capturing and supporting the analysis process. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 131–138, 2009.

[12] Paul E Keel. EWall: A visual analytics environment for collaborative sense-making. *Information Visualization*, 6(October 2006):48–63, 2007.

[13] A. Sanfilippo, B. Baddeley, A. J. Cowell, M. L. Gregory, R. Hohimer, and S. Tratz. Building a Human Information Discourse Interface to Uncover Scenario Content. *Military Intelligence*, pages 1–6, 1999.

[14] Yedendra Babu Shrinivasan and Jarke J. van Wijk. Supporting the analytical reasoning process in information visualization. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 1237, New York, New York, USA, 2008. ACM Press.

[15] John Stasko, Carsten Görg, and Robert Spence. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[16] J J Thomas and K A Cook, editors. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society, 2005.