Spatially “Connecting the Dots”

Authors Name/s per 1st Affiliation *(Author)*
line 1 (of *Affiliation*): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: e-mail address if desired

Authors Name/s per 2nd Affiliation *(Author)*
line 1 (of *Affiliation*): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: e-mail address if desired

***Abstract*— As analysts attempt to make sense of a collection of documents, such as intelligence analysis reports, they may wish to “connect the dots” between pieces of information that may initially seem unrelated. This process of synthesizing information between information requires users to make connections between pairs of documents, creating a conceptual story. We conducted a user study to analyze the process by which users connect pairs of documents and how they spatially arrange information. Users created conceptual stories that connected the dots using organizational strategies that ranged in complexity. We propose taxonomies for cognitive connections and physical structures used when trying to “connect the dots” between two documents. We compared the user-created stories with a data-mining algorithm that constructs chains of documents using co-occurrence metrics. Using the insight gained into the storytelling process, we offer design considerations for the existing data mining algorithm and corresponding tools to combine the power of data mining and the complex cognitive processing of analysts.**

Keywords—sensemaking; synthesis; data mining; text analytics

I. Introduction

In order to help analysts manage a sometimes overwhelming amount of information and gain insight from that information, researchers have been developing visual analytics tools [13]. These tools vary in what portions of the sensemaking process they target. Sensemaking can be defined as either the entire process of making sense of a collection of information or the synthesizing portion of this process [18]. This first definition can be re-written to say sensemaking is the process of foraging information and making sense of it. In this paper, we will use synthesis to refer to this portion of the sensemaking process. While some tools aim to support the overall sensemaking process, from raw data to coherent hypothesis presentation, many have a strong focus on either foraging for information [8, 12, 19] or synthesizing found information into strong hypotheses [2, 17, 23].

None of the aforementioned tools offer the ability to automatically “connect the dots” between two documents. In order for analysts to conduct this part of an investigation, they would have a few options, none of them optimal. It cannot be predicted what strategies users will employ in order to reason through the information [11]. The analyst could begin sensemaking from either document and move in the direction that they believe will lead them to the other document. Another approach is to turn the two documents into two sensemaking tasks and seeing where they intersect. These approaches could lead to confirmation bias and/or missed connections [14, 16]. Additionally, doing this kind of directed analysis by hand or even with analytical tools can be time intensive.

Data mining algorithms can be used to help automate the process of connecting the dots. Analyst input is still crucial in determining the insight, if any, that is gained from algorithm-generated stories. The connecting the dots problem has appeared in the literature in different guises and for different applications: cellular networks [3], social networks [7], image collections [9], and document collections [6, 15, 20].

In this paper, we study how users construct stories without the aid of a computer in order to improve a specific data mining algorithm, referred to as the storytelling algorithm. This algorithm was originally developed for biologists to connect genes and proteins across refereed research papers. However, it was later expanded for use in intelligence analysis [10], but the algorithm has not been evaluated to see if it mimics the way humans manually construct conceptual stories. We conducted a user study that tasked participants with manually constructing stories using two pairs of start and end points generated from the storytelling algorithm on a 47 document dataset. One story was conceptually complex and the other was more straightforward. We used a think aloud protocol [22] in order to observe “*micro-level*” connections at the document-to-document connection level and “*macro-level*” connections spanning start and end documents. Using this insight into the storytelling process, we offer design considerations for future versions of the storytelling algorithm and corresponding tools to combine the computing power of data mining and the complex cognitive processing of human analysts.

II. Storytelling Algorithm

The storytelling algorithm attempts to formalize and support the ways analysts conduct unstructured discovery, chases leads, and marshalls evidence to support or refute potentially promising chains. The story generation framework is exploratory in nature. Given starting and ending documents, it explores candidate documents applies heuristics to construct a path between the end points. These paths are then presented to the analyst who can revise or adapt them as needed.

A story between documents d_1 and d_n is a sequence of intermediate documents $d_2, d_3, \ldots, d_{n-1}$ such that every neighboring pair of documents satisfies some user defined criteria. Given a story, analysts perform one of two tasks: attempt to strengthen individual connections (*distance threshold*), or organize evidence around the given path (*clique size*). Distance threshold refers to the maximum acceptable distance between two neighboring documents in a story. Clique size threshold refers to the minimum size of the group that every pair of neighboring documents must participate in. Lower distance thresholds and greater clique sizes impose stricter requirements and neighborhood constraints, respectively, leading to longer paths. The story is composed of overlapping cliques, resulting in a story being refered to as a "clique chain." Note that a clique chain could represent multiple stories.For document modeling, storytelling uses a bag-of-words (vector) representation where the terms are weighted by tf-idf with cosine normalization. Additional details of the above mentioned algorithm are described in [10].

III. Study Description

We conducted a user study to analyze how users "connect the dots" between documents. In order to not bias the participants toward any particular analytical strategy, their analysis was done without computer aid. Even if a simple text editor was used, participants could search for keywords, which could restrict chains constructed in the "connect the dots" process to purely entity-to-entity links without any cognitive understanding of their story.

We sought to answer the following research questions:

- How do humans connect two documents when trying to connect the dots? (R1)
- How do humans connect the dots between documents into a story? (R2)
- How do human-created stories compare to algorithm-created stories? (R3)

We recruited ten participants (P1 – P10) from [omitted for review]. All participants were computer science undergraduate or graduate students. Although we did not use real-world analysts for this study, the dataset used is solvable without domain knowledge or experience in intelligence analysis.

A horizontal workspace was constructed on a large office table by covering it with a sheet of white paper (approximately 5'x3') in order to mimic the affordances of large, high-resolution displays [1]. Participants were provided with pens, pencils, highlighters, and tape, giving them the freedom to write on or highlight documents and annotate the workspace as they deemed appropriate. Individually cut documents were given to the participants, allowing flexible spatial positioning.

Participants were tasked with connecting the dots between two pairs of documents. The document collection used was a subset of the fictional "Atlantic Storm" text dataset, which has a known ground truth regarding the plot.

Before the subset was constructed, the storytelling algorithm was run on two pairs of documents that represented two subplots of the dataset. Documents were included if they were directly or moderately related to the plot. Additionally, approximately ten documents that were unrelated to either subplot were included in order to see how users discarded irrelevant information, if they discarded it at all. The storytelling algorithm was re-run on the document pairs after the document set (originally 111 documents) was scaled down to 47 documents, including the starts and ends of the stories. This ensured that the storytelling algorithm was not basing its stories on more information than was available to participants.

Two documents were designated as "Start" documents and two documents were designated as "End" documents. The starting points were paired with an ending point. In other words, "Start 1" and "End 1" could be connected through a series of documents or conceptual connections to form a story, and the same could be done with "Start 2" and "End 2." They were given two hours to complete this task and were told that there were no restrictions on the types of connections they made in order to link the starting and ending documents. They were told that constructed document chains did not have to be linear or adhere to a specific shape.

Participants were told that there were no correct or incorrect answers to the constructed stories. No practice exercise was given for the because we wanted to observe many different strategies and felt that showing participants an example of this would bias their analytical process.

After participants felt that they had found a way to sufficiently connect the two pairs of documents or they had run out of time, the proctor conducted a semi-structured interview. Participants were asked to explain how they connected the pairs of documents on two levels. They were first asked the overall conceptual story, if they were able to construct one. Then they were asked to inform the proctor of the smaller links within the story that linked documents, or clusters of documents, together. Participants were also asked if they saw any overlap between their two stories, and what that overlap was, or if they saw the two as disjoint stories.

There were no correct or incorrect answers to the stories constructed, so there were no associated solution scores. However, user-generated solutions did vary in quality based on if they were able to conceptually link the start and end document pairs. Not all participants were able to do this.

A consistent proctor was present at each study session to take observational notes regarding user process, connections, and quotes regarding the former categories. Additionally, all study sessions were video recorded to refer back to specific pieces of dialog as well as view the intermediary spatial structures users constructed. The final workspaces were preserved by taping all documents to their final positions.

IV. Connecting Document Pairs

In order to answer R1, which is a more specific form of using cognitive mechanisms to connect information [24], we analyzed the types of connections participants used to relate documents to each other. Unfortunately, it is difficult, if not impossible, to state the precise number of times each type of connection was made. Participants did not state how a document related to every other document in the dataset, and they did not always state how a document precisely related to another. Some documents related to a group of documents instead of individual ones, and others were placed into partitions of documents based on incremental formalism [21] of relevant entities.

We used open coding [5] to discover the types of cognitive connections users made in order to link two documents together. The connection types identified were entity, conceptual, temporal, speculative, and domain knowledge. These connections are listed in [Table 1], along with which participants made which types of connections. These connection types can be placed into two categories: low-level connections and high-level connections. Low-level connections are basic links between documents on keyword, or entity, matching. This connection type is employed by the storytelling algorithm.

Humans, however, also link documents based on semantic connections. These high-level connections involve participants applying cognitive schemas [4] to synthesize pieces of information between documents. In other words, low-level connections link documents based on data, and high-level connections link based on information. The connection types that fall into the high-level category are conceptual connections (a general type of high-level connection), temporal, speculative, and domain knowledge, which are specific types of high-level connections. The algorithm does not attempt to replicate these types of connections.

High-level connections involved users relating information gleaned from the data with their own cognitive schemas to gain more insight into the data than the low-level connections. The general form of this high-level connection found by users completing the storytelling task was labeled "conceptual connection." Conceptual connections cover a broad range of domains, but they are all related by the use of cognitive schemas to connect information, rather than data. An entity-entity connection does not necessarily require this additional context or ability to derive that an event is occurring.

Not all conceptual connections include an underlying entity connection. General conceptual connections can also involve emergent themes, such as "strategic planning" or "background information." Conceptual connections can be identified by participants describing relationships or events, using synonyms of entities occurring across multiple documents, or describing connections that go beyond co-occurrence. Connections may be represented spatially through proximity or overlap, but this is not always the case.

Temporal connections are links between documents that are linked explicitly because of a chronological relation. Documents are related specifically because of a relation across a period of time. For example, some participants identified that a specific transaction was occurring repeatedly over a period of time. This type of connection is a subset of a conceptual connection because participants applied specific types of schemas, specifically relating to the passage or closeness of time. Temporal connections can be identified by the use of time-related words or prepositions such as "before" or "after." Additionally, these connections can be detected by checking the dates associated with documents when they are arranged in linear shapes.

Speculative connections are connections participants made between documents that were not explicitly supported in the documents themselves but could potentially be implied. These connections had ranging confidence levels. In many instances, participants used logic and deduction to connect documents. Speculative connections were also used to express uncertainty on the participant's current hypothesis. Sometimes, participants had a difficult time identifying what the specific connection was between two documents, but they had a hunch that it exists. In addition to stating these types of hunches, participants used speculative connections to motivate further analysis. Speculative connections can be identified by words such as "I think," "might," and "not sure." As with the general conceptual connection, it may be difficult for a computer to identify this specific type of connection.

Domain knowledge connections are document links based on the participant's own outside knowledge. The documents being linked often did not have co-occurring entities. Domain knowledge connections were not always factually correct due to gaps or incorrect information in the participant's knowledge base. Domain knowledge connections can be identified by the use of words not present in either document.

Table 1. Types of cognitive connections participants made to connect pairs of documents

	1	2	3	4	5	6	7	8	9	10
entity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
conceptual		✓	✓	✓	✓	✓	✓	✓	✓	✓
temporal		✓		✓					✓	✓
speculative			✓	✓	✓	✓	✓	✓	✓	✓
domain knowledge			✓	✓	✓	✓	✓	✓	✓	✓

V. CONSTRUCTING CONCEPTUAL STORIES

In order to answer R2, we analyzed the spatial and cognitive strategies participants employed during their analysis. Participants combined many document to document connections to form stories that linked the starting and ending documents. The types of connections used varied by participant. All types of connections that participants made during their analysis (Table 1) were involved in their final stories, although the degree to which they were used varied. In this section, we discuss the cognitive and spatial strategies employed by participants to construct stories.

A. Storytelling Strategies

Some participants approached their analysis by constructing an understanding of the data without forming hypotheses until they felt they had sufficient information to support those hypotheses. They incrementally decided what was important to them in the dataset. Other participants generated hypotheses almost immediately and constantly amended them as more information was found. The third group of participants adopted a hybrid approach where hypotheses drove the analysis at times, such as deciding whether or not two people were actually the same person using an alias.

Participants differed in their strategy of how to go about connecting the dots between the start and end documents. Some participants worked from the start documents toward

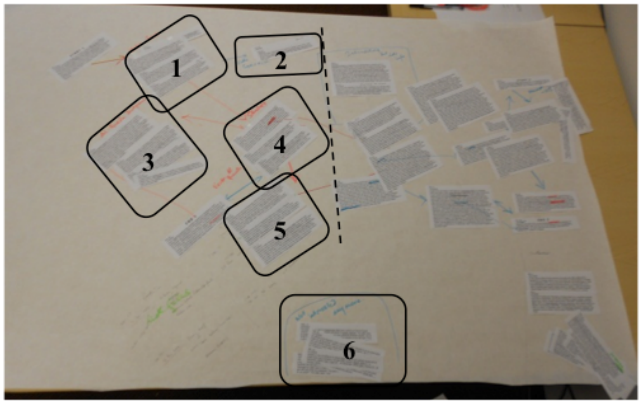

Figure 1. Participant P8's final layout with written cluster labels. The dotted line is the separation between story 1 (left) and story 2 (right.) 1: Vector, 2: Side connection, 3: Al-Queda background, 4: Strategy, 5: Vector – Al-Queda, 6: Not

their corresponding end documents. They tried to structure their analysis toward what they thought would lead to the end point. Participants expressed frustration at the difficult nature of this task. Others, through the course of their analysis, actually worked from end document to start document for one of the stories. This strategy was also present in the informal interview where participants verbally explained their stories. These participants explained the first story from start to end, and then progressed to tell the second moving from end to start. A third group of participants treated each start or end document as a starting point for their analysis and tried to see where the information overlapped.

Across all of the above-mentioned strategies, some participants worked on one story, then the next, and others worked on both simultaneously. In many case, this altered their perception of how much overlap there was between the stories. Participants that worked on the stories simultaneously tended to see many points of overlap, whereas participants that worked on the stories separately tended to see the stories as separate with maybe one point of overlap.

A few participants were not able to construct a conceptual story that linked the start and end documents. P1 only linked the start and end documents based on co-occurring entities and documents placed in chronological order. Other participants, P2 and P5, were unable to bridge a gap in their stories. These failures to connect the dots were not because the participants could not identify relevant information. Instead, these were instances of information within documents being inadvertently overlooked.

Participants were reluctant to discard information that did not initially match with either story. Even when they had established a pile of unrelated or irrelevant documents, no participant went as far as to call it a "junk pile." Instead, they used terminology such as "this doesn't seem related yet" instead of saying "this is not related." In fact, participants sifted through all documents in those piles before they concluded that they were done with their storytelling task. In order to separate irrelevant documents from relevant documents, participants placed the former category spatially far away from the relevant documents. This reluctance to

Figure 2. Participant P9's final layout with documents placed over written notes regarding document content

discard information and exclude documents from the stories resulted in human-created stories that contained more documents and more side plots than the algorithm-created stories.

B. *Intermediate Spatial Representations*

Many participants changed their spatial representation of the data at least once throughout their storytelling process. The three types of spatial representations we saw were clusters, concept maps, and timelines [Table 2].

Seven participants created **clusters** of documents during their analysis. All created clusters were based on relevance. Six out of seven participants who clustered information represented these clusters spatially. Many clusters spanned multiple terms and documents were placed along the spectrum depending on which of the terms were mentioned in the document. Multiple participants labeled clusters with words not found in any of the documents contained in the cluster.

One participant did not use spatial proximity to represent his clusters. P2 had created timelines partitioned by reporting agency (the fictional reports in the dataset came from various government agencies). He then wanted to indicate that documents were related, but he was reluctant to move documents out of their chronological position. P2 drew symbols on the upper right hand corner in order to indicate cluster membership.

Six participants constructed **concept maps**. Three of the participants, P2, P7, and P10, created their concept maps by writing entities as nodes and drawing lines between the nodes. One of these participants, P7, placed supporting evidence from documents on nodes or links. One participant that did not place documents on top of their concept graph, P2, was unable to recall the specific documents that supported his understanding of his stories. P8 linked labeled clusters of documents to form a web that represented his conceptual understanding of the stories and points of overlap [Figure 1]. The remaining two participants that created concept maps, P5 and P9, did so by writing notes on the paper. Documents that supported the written notes were placed on top of the corresponding note [Figure 2].

Five participants created **timelines**. P1 created two timelines that were composed of transitively shared entities. A pile was created for information that she did not yet find to be related to either story. P2 and P3 created timelines separated by reporting agency. P3 ranked his perceived importance of these timelines based on which agencies traditionally deal with international concerns then agencies that deal primary with domestic information. P4 and P6 both created two timelines separated by year. All participants used knowledge of their spatial layout to re-find information since a search feature was not available.

Table 2. Spatial representations used by participants

	1	2	3	4	5	6	7	8	9	10
clusters	✓	✓	✓		✓		✓	✓		✓
concept map		✓			✓		✓	✓	✓	✓
timeline	✓	✓	✓	✓		✓				

C. *Final Spatial Representations*

The final shapes of the document layouts can be found in [Table 3]. The different shapes we observed were linear with branching, web, and disorganized. As seen in Table 3, half of the participants used a linear structure, while three constructed webs and two had a layout that had no discernible shape.

Linear with branching, a layout used by five participants, resembled a narrow tree structure. These layouts were vertically oriented and were primarily a straight line with occasional documents placed next to the main line. Three out of five of the participants who used timelines in their analysis preserved the timeline in their final layout. These participants did not have a solid conceptual understanding of the stories. The rigidness of the timeline structure prevented the participants from imparting additional conceptual information through document position. The remaining two participants that created a linear with branching shape did so by matching entities. P6 followed this layout, but was left with a gap in the document chains.

Web structures, a layout used by three participants, consisted of documents with lines drawn between them [Figure 1]. This structure arose from concept mapping using documents as nodes or edges on the graph. These participants had a conceptual understanding of how the stories unfolded. In fact, these participants moved away from thinking about the plot in terms of documents to thinking in terms of entities.

"Disorganized" layouts, constructed by two participants (one can be seen in [Figure 2], were spatial representations of the data that would be extremely hard for a third party person to walk up to and derive any meaningful organization of the data. The participants who constructed these types of layouts, however, were able to talk through their conceptual stories, referencing specific documents in the mess. They did not appear to have a difficult time navigating the space. They had a good understanding of the conceptual stories, but the organized web structure participants were more coherent and structured in their stories, especially compared to users with linear with branching layouts.

Table 3. Final document layouts: physical shapes

	1	2	3	4	5	6	7	8	9	10
linear with branching	✓	✓		✓		✓				✓
web			✓				✓	✓		
disorganized					✓				✓	

VI. COMPARISON TO ALGORITHM-GENERATED STORIES

In order to answer our final research question, R3, we quantitatively compared the documents participants included in their stories with those chosen by the algorithm. Out of the ten participants, eight noted which documents pertained to which story. The two participants that did not organized their documents as a timeline. They forgot which documents the story information came from. We analyzed the documents used by the eight participants in order to compare the stories from the algorithm and the participant stories.

The storytelling algorithm was run using the same subset of 47 documents given to our experiment participants. For each start and end document pair, the algorithm produced a chain of documents and cliques. Story 1 contained three documents and four clique documents. Story 2 contained two documents and two clique documents. Participants' stories contained almost triple this amount. Table 4 compares the lengths of the algorithm's stories to the participants' stories.

There were an average of 18.9 documents contained in story 1 and 14.4 in story 2. This means there was an average of 13.7 documents deemed not relevant to either story. Participants included documents even if the relevance was low, whereas the algorithm finds the best story containing the least number of documents. Because of this, the algorithm finds *a* story, but people find a *more complete* story with supporting evidence (e.g. background information, subplots). Humans are able to say why and how the start and end documents are related, whereas the algorithm can only tell if they are related and how.

We analyzed all documents included by either the algorithm or the participants. Over 50% of participants included all documents contained in story 2. Over 75% of participants included all documents contained in the algorithm's story 1. Participants also included documents that were not found by the algorithms. Over 63% of participants included documents that were not contained in the algorithm's story 2. Over 50% of participants included documents not found in the algorithm's story 1. Two participants included other documents not contained in either the algorithm's stories or the majority of the other participants' stories. These statistics support the notion that participants creating the same conceptual stories as the algorithm, but going beyond the simple stories and including supporting information.

One interesting case occurred when a document was left out of two participants' stories even though it was included in the algorithm's story was a document that contained a list of individuals staying at various hotels. To the algorithm, this document provided a rich source of entities on which it could match to other documents. However, to these two participants, this was a document with information with little to no context that was set aside.

The algorithm should take into consideration the density of entities mentioned in each document. If the number of entities is extremely close to the number of words in the document, this is most likely a list of information that will have limited contextual meaning for users. However, these documents could contain relevant information that users may initially pass over due to the lack of context and overwhelming number of entities. We recommend presenting these documents alongside less dense documents to provide context as well as the data.

Tools integrating this algorithm should consider presenting the found documents in a web-like structure along with off-shooting documents that likely contain side plots. This allows users to then apply meaning to this layout and recall where documents are located. Options should also be available for users to quickly access information that is found to be background information concerning particular persons. This could be accomplished by identifying cliques that are formed by matching on the person's name.

Table 4. Number of documents included in the algorithm- or user-created stories

	algorithm	user min	user max	mean	median
story 1	7	14	26	18.9	19
story 2	4	9	18	14.4	16

VII. Conclusion

We described a user study that investigated how humans connect the dots between two documents in a fictional intelligence analysis dataset. We analyzed the cognitive connections participants used to connect documents to each other, as well as the spatial representations used to arrange documents and externalize information. In these results, we were able to conclude that conceptual connections (and the more specific types of conceptual connections) yielded greater insight into document relationships. We also pointed out the importance of domain knowledge and how domain knowledge gaps can lead to nonsensical connections. We found that intricate and sometimes messy spatial representations of the data on the whole yielded higher levels of comprehension and conceptual cohesion than primarily linear and orderly layouts. Additionally, the user-created stories were compared with those created by the storytelling algorithm. We found that users create larger stories that cover a wider conceptual range.

It is our hope that a better understanding of this analytical technique (trying to connect the dots between specific documents) will result in the integration of similar data mining algorithms into visual analytics tools in order to combine the computational power of computers and the analytical finesse that humans provide in order to achieve insight into datasets

References

1. Andrews, C., Endert, A. and North, C. Space to think: large high-resolution displays for sensemaking *Proceedings of the 28th international conference on Human factors in computing systems*, ACM, Atlanta, Georgia, USA, 2010.
2. Bier, E., Ishak, E. and Chi, E. Entity Workspace: An Evidence File That Aids Memory, Inference, and Reading

Intelligence and Security Informatics. in Mehrotra, S., Zeng, D., Chen, H., Thuraisingham, B. and Wang, F.-Y. eds., Springer Berlin / Heidelberg, 2006, 466-472.

3. Brassard, J.-P. and Gecsei, J. Path building in cellular partitioning networks *Proceedings of the 7th annual symposium on Computer Architecture*, ACM, La Baule, United States, 1980, 44-50.
4. Cooper, G. and Sweller, J. Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79* (4). 347-362.
5. Corbin, J.M. and Strauss, A. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13* (1). 3-21.
6. Das-Neves, F., Fox, E.A. and Yu, X. Connecting topics in document collections with stepping stones and pathways *Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM, Bremen, Germany, 2005, 91-98.
7. Faloutsos, C., McCurley, K.S. and Tomkins, A. Fast discovery of connection subgraphs *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Seattle, WA, USA, 2004, 118-127.
8. Havre, S., Hetzler, E., Whitney, P. and Nowell, L. ThemeRiver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, *8* (1). 9-20.
9. Heath, K., Gelfand, N., Ovsjanikov, M., Aanjaneya, M. and Guibas, L.J., Image webs: Computing and exploiting connectivity in image collections. in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, (2010), 3432-3439.
10. Hossain, M.S., Andrews, C., Ramakrishnan, N. and North, C. Helping Intelligence Analysts make Connections *AAAI '11 Workshop on Scalable Integration of Analytics and Visualization*, 2011.
11. Johnson-Laird, P.N. *How we reason*. Oxford Univ. Press, New York, NY [u.a.], 2006.
12. Kang, H., Plaisant, C., Lee, B. and Bederson, B.B. NetLens: Iterative Exploration of Content-Actor Network Data. *Information visualization*, *6* (1). 18-31.
13. Keim, D., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. Visual Analytics: Scope and Challenges

Visual Data Mining. in Simoff, S., Böhlen, M. and Mazeika, A. eds., Springer Berlin / Heidelberg, 2008, 76-90.

14. Klein, G., Moon, B. and Hoffman, R.R. Making Sense of Sensemaking 2: A Macrocognitive Model. *Intelligent Systems, IEEE*, *21* (5). 88-92.
15. Kumar, D., Ramakrishnan, N., Helm, R.F. and Potts, M. Algorithms for Storytelling. *Knowledge and Data Engineering, IEEE Transactions on*, *20* (6). 736-751.
16. Mynatt, C.R., Doherty, M.E. and Tweney, R.D. Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, *30* (3). 395-406.
17. Payne, J., Solomon, J., Sankar, R. and McGrew, B., Grand challenge award: Interactive visual analytics palantir: The future of analysis. in *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, (2008), 201-202.
18. Pirolli, P. and Card, S., The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. in *International Conference on Intelligence Analysis*, (2005).
19. PNNL. Pacific Northwest National Laboratory, INSPIRE visual document analysis.
20. Shahaf, D. and Guestrin, C. Connecting the dots between news articles *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Washington, DC, USA, 2010, 623-632.
21. Shipman, F.M. and Marshall, C.C. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)*, *8* (4). 333-352.
22. Someren, M.W.v., Barnard, Y.F. and Sandberg, J.A.C. *The think aloud method : a practical guide to modelling cognitive processes*. Academic Press, London; Boston [etc.], 1994.
23. Wright, W., Schroh, D., Proulx, P., Skaburskis, A. and Cort, B. The Sandbox for analysis: concepts and methods *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, Montreal, Quebec, Canada, 2006, 801-810.
24. Zhang, P., Soergel, D., Klavans, J.L. and Oard, D.W. Extending sense-making models with ideas from cognition and learning theories. *Proceedings of the American Society for Information Science and Technology*, *45* (1). 23-23.