How Analysts Cognitively “Connect the Dots”

Lauren Bradel, Jessica Zeitz Self, Alex Endert, M. Shahriar Hossain, Chris North, Naren Ramakrishnan
Department of Computer Science
Virginia Tech, Blacksburg, VA

Abstract— **As analysts attempt to make sense of a collection of documents, such as intelligence analysis reports, they need to “connect the dots” between pieces of information that may initially seem unrelated. We conducted a user study to analyze the cognitive process by which users connect pairs of documents and how they spatialize connections. Users created conceptual stories that connected the dots using a range of organizational strategies and spatial representations. Insights from our study can drive the design of data mining algorithms and visual analytic tools to support analysts’ complex cognitive processes.**

Keywords—sensemaking; synthesis; data mining; text analytics

I. Introduction

To help analysts comprehend overwhelming amounts of information, researchers have been developing visual analytics tools [7]. These tools vary in what portions of the sensemaking process they target [6]. 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 [5] or synthesizing found information into strong hypotheses [8]. However, existing tools do not offer significant support for connecting seemingly unrelated documents. It cannot be anticipated what strategies users will employ in order to reason through the information [4]. Data mining algorithms [3] can be used to help automate the process of connecting the dots, yet analyst input is still crucial in determining what insights are gained while forming stories [2].

In this paper, we study how users connect the dots and construct stories without the aid of a computer in order to inform the design of storytelling algorithms. The framework of storytelling algorithms we use is from [3], originally developed for knowledge discovery in collections of short texts such as biomedical abstracts. In addition to finding short paths between desired end points, the algorithm also aims to marshall supporting documents to form local neighborhoods around each link in the path. Although successful, the algorithm has not been evaluated to see if it mimics the way humans manually construct conceptual stories.

We conducted a user study that tasks participants with manually constructing stories using two pairs of start and end points on a 47 document dataset. One story is intended to be conceptually complex and the other was more straightforward. We observed “*micro-level*” connections at the document-to-document connection level and “*macro-level*” connections spanning several documents. Using insights gleaned from the user study, we aim to research design considerations for future versions of the storytelling algorithm and corresponding visual analytics tools to combine the computing power of data mining and the complex cognitive processing of human analysts. In particulary, we hope to inform the design of interactive approaches in which mining algorithms can observe human analytic processes to gain additional hints about potential semantic connections identified by the user.

II. Study Description

We sought to answer the following research questions: “How do humans connect two documents when trying to connect the dots?” (R1) and “how do humans connect the dots between many documents into a whole story?” (R2). We recruited ten participants (P1 – P10) who were computer science undergraduate or graduate students. Although we did not use real-world analysts, the dataset used is solvable without experience in intelligence analysis.

A horizontal workspace was constructed on a large table by covering it with a sheet of white paper (approximately 5’x3’). 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 within two hours with no restrictions on how to connect the documents. The document collection used was a subset of the fictional “Atlantic Storm” text dataset. After completing their analysis, the participants explained their overall stories as well as the smaller connections within each story. There were no correct or incorrect answers to the stories constructed, and thus no associated solution scores. However, user-generated solutions varied in conceptual quality and cohesion.

III. Connecting Document Pairs

To answer R1, we analyzed the types of connections participants used to relate documents. We used open coding to discover the types of cognitive connections users made to link documents pairs. The connection types identified were *entity, conceptual, temporal, speculative,* and *domain knowledge* [**Table 1**], and can be organized into: low-level and high-level connections. **Entity** connections (e.g. “*Hmm…same place, Peshawar, Peshawar.*”) are low-level, and represent simple links between documents based on entity co-occurence. This connection type is employed by the storytelling algorithm and is easy to recognize. The remaining connection types are high-level, and involve participants applying cognitive schemas to

Supported in part by NSF grant CCF-0937133.

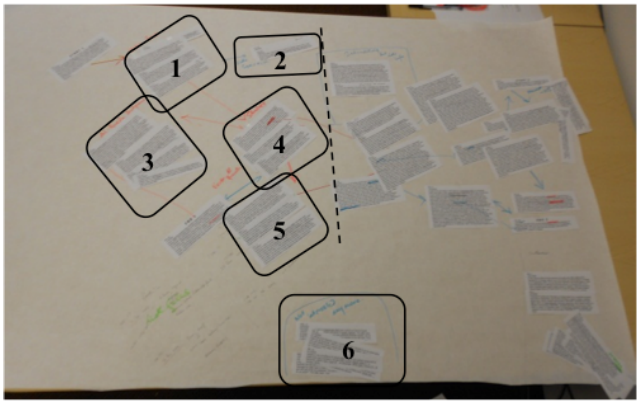

Fig. 2. P8's final web-shaped layout (numbered regions denote nodes used in story 1) using concept maps and clusters of

synthesize information between documents. These micro-level connections combined to construct the ways by which participants connectedc the starting and ending documents.

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 high-level connection found by users completing the storytelling task was labeled "conceptual connection." **Conceptual** connections (e.g. "*These documents are about trading diamonds on the black market.*") cover a broad range of domains, but they are all related by the use of cognitive schemas to connect information, rather than data. 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. Users typically represent conceptual connections spatially through proximity or overlap, but this is not always the case.

Temporal connections (e.g. "*This happened after they got the money.*") between documents 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 made by users can be identified by their use of time-related words or prepositions such as "before" or "after", or by the dates associated with documents when users spatially arrange them in linear shapes.

Speculative connections (e.g. "*I think money from this diamond trade is being used to fund a scholarship.*") 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. Sometimes, participants had difficulty 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 recognized by users' 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 type of connection through the user's interactions.

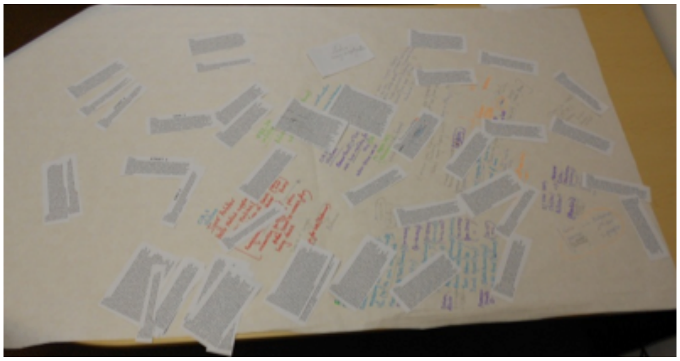

Fig. 1. P9's final disorganized layout using concept maps. Evidentiary documents are placed on top of the written concepts.

Domain knowledge connections (e.g. "*Tanzanite is a precious stone. I bet trading tanzanite stones is related to trading diamonds.*") are 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 user's addition of supporting information not present in either document.

TABLE 1. TYPES OF COGNITIVE CONNECTIONS PARTICIPANTS MADE TO CONNECT DOCUMENTS, MARKED AS LOW-LEVEL (L) OR HIGH-LEVEL (H)

Participant #	1	2	3	4	5	6	7	8	9	10
Entity (l)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Conceptual (h)		✓	✓	✓	✓	✓	✓	✓	✓	✓
Temporal (h)		✓		✓					✓	✓
Speculative (h)			✓	✓	✓	✓	✓	✓	✓	✓
Domain kn.(h)			✓	✓	✓	✓	✓	✓	✓	✓

IV. CONSTRUCTING STORIES

To answer R2, we analyzed the spatial and cognitive strategies participants employed during their analysis to construct stories. Participants combined many inter-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.

A. 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**]. All participants used knowledge of their spatial layout to re-find information since a search feature was not available.

Seven participants created **clusters** of documents based on relevance during their analysis. Six out of seven participants who clustered information represented these clusters spatially. The remaining participant tagged documents with symbols to cluster them while maintaining their temporal layout. Multiple participants labeled clusters with words not found in any of the documents contained in the cluster, as seen in [1].

Six participants constructed **concept maps**. Three of the participants created them by writing entities as nodes and drawing lines between the nodes [**Fig. 2**]. One of these participants placed supporting evidence from documents on nodes or links. One participant that did not place documents on top of their concept graph was unable to recall the specific documents that supported his understanding of his stories. 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 [**Fig. 1**].

Five participants created **timelines**. P1 created two timelines that were composed of transitively shared entities. 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 vs. domestic concerns. P4 and P6 both created two timelines separated by year.

TABLE 2. SPATIAL REPRESENTATIONS USED BY PARTICIPANTS

Participant #	1	2	3	4	5	6	7	8	9	10
Clusters	✓	✓	✓		✓		✓	✓		✓
Concept maps		✓			✓		✓	✓	✓	✓
Timelines	✓	✓	✓	✓		✓				

B. 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*. The linear with branching layout contained each start and end document as end points of the structure, while the web and disorganized layouts typically had no clear starting or ending point, requiring the users to conceptually connect the dots instead of following a path between the two target documents.

Linear with branching, a layout used by five participants, resembled a narrow tree structure. This layout was primarily formed by low-level entity connections and slightly higher-level temporal connections. 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 making entity connections. It should be noted that this is the primary method by which the storytelling algorithm presents its solution to users.

Web structures, a layout used by three participants, consisted of documents with lines drawn between them [**Fig. 2**]. 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.

TABLE 3. FINAL DOCUMENT LAYOUTS: PHYSICAL SHAPES

Participant #	1	2	3	4	5	6	7	8	9	10
Linear branching	✓	✓		✓		✓				✓
Web			✓				✓	✓		
Disorganized					✓				✓	

"**Disorganized**" layouts, constructed by two participants [**Fig. 1**], were spatial representations of the data that would be extremely difficult for third-party persons to understand but was easily navigated by the participant. 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 layouts.

V. CONCLUSION

Our user studies have revealed the importance of conceptual connections (and the more specific types of conceptual connections) in enabling users to gain insight into document relationships. We have also highlighted the importance of domain knowledge, and how domain knowledge gaps can lead to nonsensical connections. Finally, intricate webs and sometimes messy spatial representations of the data on the whole have yielded higher levels of comprehension and conceptual cohesion than primarily linear layouts. These results motivate new research into the design of storytelling algorithms that (a) can connect the dots using complex webs beyond simple linear paths, (b) combine multiple models including clusters, concept maps and timelines, (c) recognize cues for high-level connections from users, and (d) support highly interactive approaches for combining advantages of both cognition and computation.

REFERENCES

1. Endert, A., Fox, S., Maiti, D., Leman, S. and North, C. "The semantics of clustering: analysis of user-generated spatializations of text documents," Proceedings of the International Working Conference on Advanced Visual Interfaces, 2012, 555-562.
2. 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.
3. Hossain, M.S., Gresock, J., Edmonds, Y., Helm, R., Potts, M. and Ramakrishnan, N. "Connecting the dots between PubMed abstracts," PloS one, 7 (1).
4. Johnson-Laird, P.N. "How we reason," Oxford Univ. Press, New York, NY, 2006.
5. Kang, Y.-a., Gorg, C. and Stasko, J. "Evaluating visual analytics systems for investigative analysis: deriving design principles from a case study," IEEE Visual Analytics Science and Technology, 2009, 139-146.
6. 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.
7. Thomas, J. and Cook, K. "Illuminating the path: the research and development agenda for visual analytics," 2005.
8. 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, 2006, 801-810.