The Semantics of Clustering: Analysis of User-Generated Spatializations of Text Documents

Alex Endert[1], Seth Fox[1], Dipayan Maiti[2], Scotland Leman[2], Chris North[1]
[1]Department of Computer Science, [2]Department of Statistics
Virginia Tech
{aendert, north}@vt.edu

ABSTRACT

Analyzing complex textual datasets consists of identifying connections and relationships within the data based on users' intuition and domain expertise. In a spatial workspace, users can do so implicitly by spatially arranging documents into clusters to convey similarity or relationships. Algorithms exist that spatialize and cluster such information mathematically based on similarity metrics. However, analysts often find inconsistencies in these generated clusters based on their expertise. Therefore, to support sensemaking, layouts must be co-created by the user and the model. In this paper, we present the results of a study observing individual users performing a sensemaking task in a spatial workspace. We examine the users' interactions during their analytic process, and also the clusters the users manually created. We found that specific interactions can act as valuable indicators of important structure within a dataset. Further, we analyze and characterize the structure of the user-generated clusters to identify useful metrics to guide future algorithms. Through a deeper understanding of how users spatially cluster information, we can inform the design of interactive algorithms to generate more meaningful spatializations for text analysis tasks, to better respond to user interactions during the analytics process, and ultimately to allow analysts to more rapidly gain insight.

Categories and Subject Descriptors

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors

Keywords

Text analytics, visualization, visual analytics, clustering.

1. INTRODUCTION

Analysts are often tasked with analyzing, understanding, and making sense of collections of text documents. One of the fundamental activities in performing such tasks successfully is to establish complex connections, relationships, and similarities within the data – a task at which humans inherently excel. One helpful approach to supporting analysts is to provide them with a spatial workspace in which they can spatially organize documents into clusters and other visual structures. The familiarity and flexibility afforded by a spatial workspace allows users to establish implicit relationships within the dataset [1]. Thus, users have the ability to create spatial relationships (e.g., by moving documents and creating clusters) without the requirement of explicitly formalizing the relationships.

Large, high-resolution displays further enhance a user's ability to perform sensemaking tasks spatially. The increased physical size and resolution of these displays (such as the one used in this study, shown in Figure 1) present users with a fundamentally different space. That is, when technological constraints such as limited display size and resolution are reduced, users are able to utilize a broader range of human abilities for their task [2]. For sensemaking tasks in particular, the positions of information in the workspace become meaningful to the users, and the overall layout of the information serves as a memory aid during their investigation [3].

Similarly, mathematical algorithms exist that computationally generate such spatially clustered layouts. Through a pre-determined distance function, these algorithms attempt to compute relationships within the data, such as inter-document similarity. Thus, the basis of these algorithms is to extract structure from the dataset, compute similarities, and present users with a two-dimensional view of the information showing an overview of the primary themes and general structure of the dataset. However, we contend that sensemaking is a far more complex process, and cannot be described solely by structure contained directly within the given dataset, and that user input is required.

Sensemaking is the process of generating understanding and insight about a collection of information [4]. Drawing from human intuition and previously developed domain expertise, users can spot connections, see patterns, create stories, and ultimately generate insight. Thus, we are interested in the *semantics* of how users perform such sensemaking tasks spatially. That is, what inferences can we make from their process – both in terms of their interactions in the workspace, as well as the clusters and spatial structures they create? Further, how can these findings help guide statistical algorithms responsible for computationally clustering or

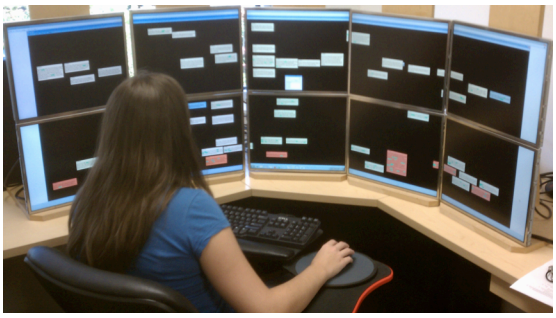

Figure 1. The large, high-resolution display system used in this study (total of 13.1 megapixels).

AVI '12, May 21-25, 2012, Capri Island, Italy

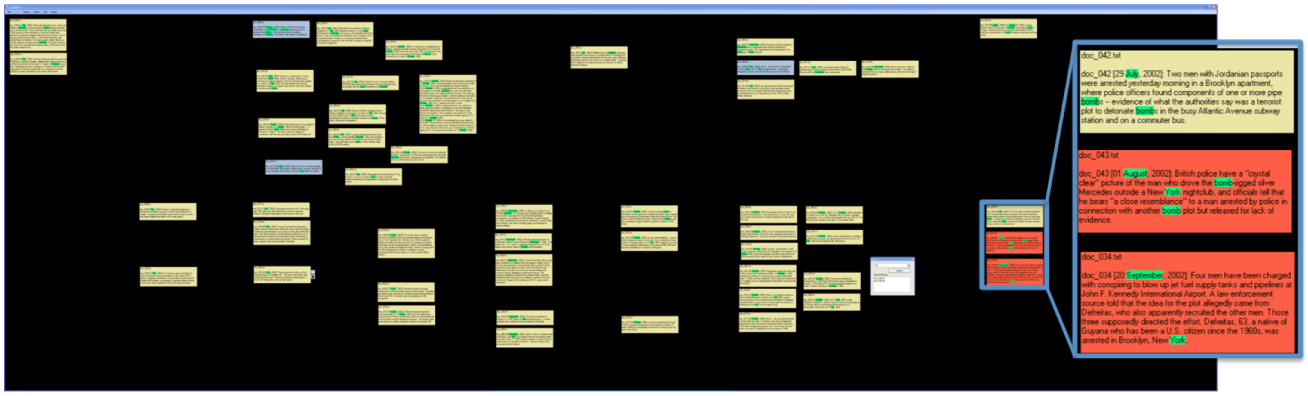

Figure 2. LightSPIRE, a large-display spatial workspace used in this study for organizing text documents. We observed users spatially analyzing a textual dataset, using common interactions such as searching, highlighting, and positioning documents.

spatializing information?

A goal of visual analytics [5] is to combine the analyst's domain knowledge essential for sensemaking with the computational support of clustering algorithms through interactive visualizations. For example, one approach to this is *semantic interaction* [6] that exploits the interactions users perform as part of their exploratory spatial analysis, and couples them with specific updates to the clustering algorithm, shielding the user from having to directly modify these parameters. Thus, throughout the iterative analytic process, the system incrementally learns about the user's domain knowledge, and a spatial layout is co-created by the user and the system. This is an iterative process in which users guide the algorithms to produce layouts consistent with their mental models. In order to design such interactive algorithms, it is critical to understand 1) the basis upon which users cluster information, and 2) what analytical reasoning can be extracted from the interactions occurring during their processes.

In this paper, we present the results of a user study exploring these two challenges. We asked users to perform a spatial sensemaking task on a large, high-resolution display (shown in Figure 1) using a system, LightSPIRE (shown in Figure 2), that provides basic text analysis functionality (i.e., searching, highlighting, annotating, and document positioning). The tool provides only manual layout capabilities, with no algorithmic layout support. Users can manually spatially organize the documents however they desired to help them complete their analysis. We then analyzed the user-generated clusters in each user's spatial layout to better understand the semantics of their clusters, as well as analyzed their process in terms of the interactions.

Based on the findings of the user study, the contributions of this work are as follows. First, we discuss how the criteria on which users clustered information for sensemaking was not necessarily based only on structure found within the data. Users created clusters (and other spatial constructs) based on a combination of the structure within the data (e.g., the entities in the text), as well as their intuition and higher-level concepts. Second, we analyze the user interaction during their analytic process to show how certain interactions indicate important discriminating features in the data. Third, we discuss how these findings can influence the design of statistical models created for interactive visual analytics.

2. Related Work

2.1 Using Space for Sensemaking

Visualizations exist that aid users in sensemaking by allowing them to manually organize information spatially. The cognitive benefits of allowing users to generate spatial layouts of information have been studied. For instance, Marshall and Rogers [1] found that users prefer to create implicit relationships between information by positioning related information closer together. They found that the ease, flexibility, and informality associated with creating these relationships spatially were important to users.

From Andrews et al., we learn that users externalize semantic information about a dataset into the layout and organization of documents [3]. The spatial layouts created represent specific meaning about each individual user's analysis. Therefore, user's findings from their analysis task were present in their spatial layouts. This study extends on this work by quantifying these relationships in terms of the clusters generated, as well as the interactions utilized during the processes.

Pirolli and Card present a model for sensemaking for intelligence analysis task [4]. This model illustrates the series of cognitive stages users proceed through when performing a sensemaking task. Most notably, we learn from their work that much of the success of sensemaking is based on the ability for humans to combine their domain expertise gained from previous experiences and the information from the dataset they are currently investigating. It is through this combination that they are successful in identifying complex relationships within the data and ultimately gaining insight.

2.2 Spatialization and Clustering Algorithms

Several algorithms exist with a similar purpose of mathematically generating two-dimensional layouts from which users can interpret important information about a dataset. In general, algorithms group or organize data based on similarity, which is a function of the features of the dataset. Dimensionality reduction algorithms can provide a 2-d spatial visualization of the clustered data. For example, algorithms like self-organizing maps [7] or generative topographic mapping [8] provide a direct method of visualizing text data spatially, but do not provide explicit cluster membership information. A survey of clustering algorithms can be found in [9] and is outside the scope of this paper. The primary criteria upon which these models generate layouts are structure extracted from the dataset, such as term frequencies, temporal attributes [10], etc. from textual datasets [11].

However, we contend that in order for these algorithms to aid users during the entire process of sensemaking, their design needs to change from focusing primarily on the structure of the data to combining this structure with semantics derived from the user.

2.3 User Interaction in Visualization

One of the challenges for information visualization is to gain a deeper understanding of how users interact within visualization,

and more importantly how these interactions are integrated into their analytic process [12]. Yi et al. have addressed this lack of understanding by presenting an extensive categorization of user interactions available in popular exploratory visualization tools [13]. However, interaction in visualization has been shown to be inherently complicated to categorize [14].

Dou et al. have shown that through logging user interactions in a visualization of financial data, low-level analytical processes can be reconstructed [15, 16]. Most importantly, these results indicate that a detectable connection exists between the low-level user interaction and the analytic process of that user.

Our work described in this paper addresses a related topic area – analyzing the relationships between the spatial layouts users create while exploring a dataset, and investigating how the user interactions within that process can be correlated to the solution users generated. We discuss how these findings can extend to help enhance the effectiveness of clustering algorithms.

3. Method

The purpose of this study is to analyze users' spatial clustering of information to aid with their sensemaking task. Participants in the study analyzed a textual dataset to understand and uncover a fictional terrorist activity using a simple spatial document organizational tool called LightSPIRE. We chose this task and dataset, as it is representative of intelligence analysis tasks that are largely focused on sensemaking.

This study explores two primary questions:

1. *Analysis of Spatial Layout*. What structure exists within the user-generated clusters? That is, given the clusters created by the users, what structure can be algorithmically detected?
2. *Analysis of Process*. What can we learn from users' interactions during the analytic process that can help guide algorithms? What indicators of analytical reasoning can be derived from these interactions?

3.1 Equipment

Users of the study were given a spatial document organization tool, LightSPIRE, for their task. LightSPIRE (shown in Figure 2) provides a workspace where documents can be manually organized using basic, familiar interactions. The primary interaction afforded by LightSPIRE is movement of the documents. Users also have the choice to view documents using two levels of detail, full-text and filename only (documents could not be deleted). A search function allows users to query the dataset for text strings. Search hits are shown within the documents as permanent, green highlights. The documents that contain the current search query are shown in a darker red until another search is performed or the search is cleared. Users also have the ability to highlight text (in yellow) as they are reading the documents. LightSPIRE captures and logs all of these interactions for post-study analysis.

The workstation used for this study is a large, high-resolution display (LHRD), shown in Figure 1. The workstation is constructed using ten 17" LCD monitors arranged in a 5x2 grid (total resolution: 6400 x 2048, or 13.1 megapixels), curved around the user to provide optimal access to all areas of the workspace [17]. The display is driven using a single workstation running Windows XP, thus allowing familiar mouse and keyboard interaction with the workspace. Using LightSPIRE on this LHRD, users gain the ability to display the entire dataset in full-text if desired, as well as create an environment where spatial location of the information conveys meaning to the user [3]. Users were also given access to a whiteboard and a notepad for notes, although no users made use of these.

3.2 Dataset

The intelligence analysis training dataset used for this study consisted of 50 textual documents containing a hidden fictitious terrorist plot. The dataset includes a known ground truth, and includes a scoring rubric to assess the findings of each user. It also includes a list of "important" documents (22 out of 50) that are relevant to supporting the solution. Thus, we are able to draw conclusions on effectiveness of the solutions based on the scoring rubric, and analyze the interactions and spatial layouts based on which documents are important to the solution.

3.3 Procedure

Users were given practice with the workspace and LightSPIRE prior to beginning their analysis. During this time, all the functionality of LightSPIRE was shown to them, and they were able to ask any questions. Then, they were given instructions to analyze the dataset to uncover any suspicious activity, gathering as much information to support (and refute) their hypothesis as possible. No information was given regarding the important and unimportant documents. They were informed of the one-hour time limit for their analysis, after which they would be asked a series of questions about their solution. During this post-task questionnaire, the workspace would remain visible, but they would not be allowed to interact with it (other than looking at it and reading). This semi-structured interview provides users the ability to explain their solution in as much detail as possible, then goes on to ask details about relationships between people, places, and events to determine how well users could uncover these complex relationships during their investigation. Finally, we asked users to sketch (on a blank piece of paper) a drawing to identify and label their clusters, and help us better understand the meaning of the layout they created. The entire duration of the study lasted approximately two hours.

3.4 Data Collected

LightSPIRE was designed to log all of the user interactions, including search terms, cursor movement and activities, document movement and positioning, and document opening and closing. From these logs, we can analyze the users' process at the interaction level. In addition, screenshots of the entire workspace were taken at 10-second intervals. The screenshots allow us to analyze the clusters and spatial layouts generated by the users. A description of the spatial layout was made through the sketch produced by each user at the end of the study, where clusters and other spatial constructs were clearly labeled. The entire study was video recorded primarily to capture the conversation between the user and the investigator, as well as capture any gestures made towards the workstation during the post-study questionnaire.

3.5 Participants

This study consisted of observing 15 users. The users were all male, undergraduate computer science students. While these participants had no prior training in intelligence analysis, the domain expertise required to correctly solve the dataset is basic intuition and reasoning. The participants were offered an opportunity to receive one of three monetary prizes of $50, $35, and $25 for the top three most accurate and complete solutions (based on the scoring rubric provided with the dataset) to provide motivation for their task.

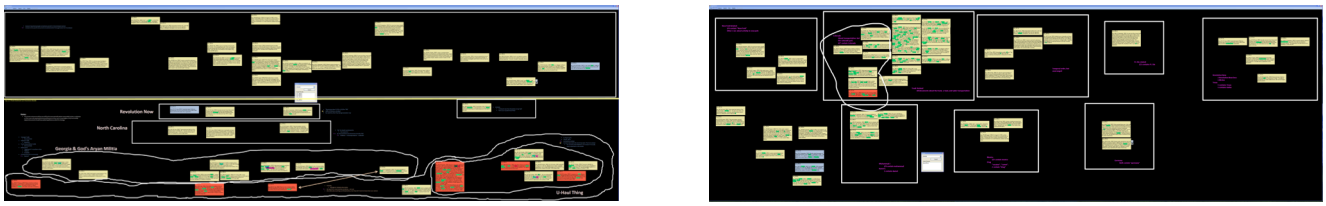

Figure 3 Annotated screenshots of two final layout states. The annotations (white frames and purple text) were added by the investigators based on the cluster boundaries and labels provided by the post-task interviews. (Left) shows an example of the *Hybrid Clustering* spatial layout, where the user organized the documents temporally from left to right, while the separation along the y-axis was used to organize topics of interest. (Right) is an example of the *Topical Clustering* layout, where the user chose to organize the documents in clusters based on topics important to the solution.

4. Results

The results of this study are presented as follows. First, we analyze the final spatial layout the users created. We analyze the spatial layout produced by each user to gain a better understanding of the structure of the user-generate clusters. Second, we analyze the user interactions during the users' processes of creating these layouts.

4.1 Analysis of Spatial Layout

The initial layout of the 50 documents was identical for each user. Each of the documents were minimized (showing only the filename), and arranged based on their filename (i.e., doc_01, doc_02, etc.) in the top left corner of the workspace. Each document was only present once in the workspace.

4.1.1 Primary Spatial Layout

What are users' overall layout strategies? The analysis of overall spatial layout reveals three distinct patterns of how users chose to spatially organize their information.

Topical Clustering. Nine out of the fifteen users in this study chose to organize their workspace based primarily on creating clusters of topically related documents. Figure 2 shows a representative example of a workspace organized by clustering. Users organized information into clusters to synthesize their hypotheses. For example, at times users labeled their clusters "Aryan activities" to represent a cluster that was focused around the documents within the dataset that relate to that information. However, users also created clusters labeled "junk" or "related but not big picture", indicating that clusters can also represent forms of insight about the dataset.

Temporal Clustering. Five of the fifteen users organized their workspace based on the temporal information in the documents. These users arranged their information from left to right based on the dates included with each document (see Figure 3). For this group, the users chose to place no relevant information on the y-axis of the workspace. When asked, one user replied that "[he] used the vertical dimension of the display to make room to fit documents if the dates overlapped". For example, one user outlined an area of the workspace and labeled it "August". We classify each one of such areas as a "cluster" for the purposes of this work.

Hybrid Clustering. One user generated a particularly interesting layout (shown in Figure 3). He started his investigation by organizing the documents based on a timeline on roughly the top half of the workspace. Then, he began investigation the relationships and interesting events within the dataset. As he found interesting terms or events, he pulled these documents out of the timeline and clustered them in the lower portion of the display. However, the documents retained their relative temporal positioning, as he took caution to only move the documents vertically, so as not to disturb the temporal left-to-right organization. As a result, we noticed this user balanced a tradeoff of maintaining temporal awareness of the documents, as well as gaining an understanding of the important events and topics within the dataset by establishing "rows" of related items.

4.1.2 Cluster Structure

We analyze the raw clusters created by each user during the task, and later identified in their post-task interview. The 15 users created a total of 86 clusters. The number of documents contained in each cluster ranged from 1 to 25 documents, with a mean of 7.3 documents per cluster.

How do documents within a cluster relate to each other?

Intra-cluster Co-occurrence First, we analyze if one or more terms occurs in all the documents within a cluster. 26 of the 86 clusters (30%) had at least one term in common among all the documents in the given cluster. 10 out of the 15 users made these clusters containing common terms. 13 of these 26 clusters had a month as one of the common terms (these clusters belonged primarily to the *Temporal Clustering* users). As can be expected, for clusters of smaller sizes, there were more shared common terms. Only 5 of these 26 clusters contained more than two documents. For these 5 clusters with more than two documents, the number of common terms never exceeded four. Hence, for the remaining 70% of clusters the structure of the clusters is not based on any co-occurring terms in all the documents.

Transitivity An alternate but simplistic explanation of cluster structure is that pairs of documents within a cluster are related via terms that are common between them (i.e., clusters represented as connected graphs where nodes are documents and edges represent shared entities between the two documents). Therefore, any two documents within the cluster can be connected transitively via one or more other documents. We refer to such a cluster as a *transitive* cluster. For example, one user created a cluster with three documents in which one pair of documents did not have any words in common (shown in Figure 6). However, a pair of documents shared the term "Arrested*"* and another pair shared the terms "Cartels*"* and "Drug" and a transitive relationship between the documents in the cluster can be given by:

doc_39(Arrested)→ doc_15(Arrested, Cartel)→doc_28(Cartel, Drug)

Hence, while these three documents produce a connected graph, they do not share a common term between all three. The abstraction of a large corpus of text documents as a similarity network (the notion of similarity being induced by terms that are shared between document pairs) has been used by [18, 19] in a "Storytelling algorithm" to connect seemingly unrelated documents via a path referred to by the authors as a *story*. While the ordering of documents in the transitive relationship between two documents might bear some semantic meaning to the users,

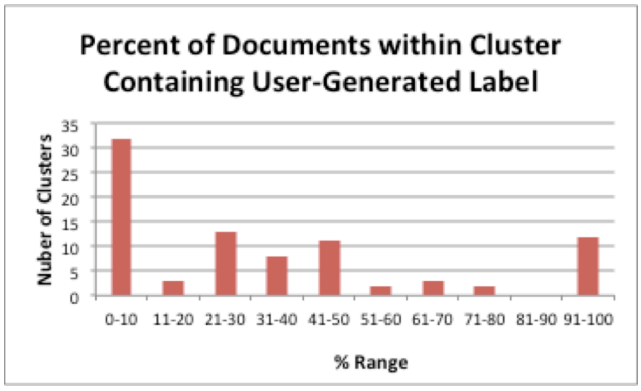

Figure 5 The distribution of the percentage of documents within each cluster that contain the cluster label keywords. Of the 86 user-generated clusters, 28 clusters did not have their label keywords present in any of their documents. 67 clusters do not contain the label keywords in more than 50% of the documents.

we do not account for ordering in our analysis (i.e., our graphs are undirected).

Based on this, 71 of the 86 clusters (83%) are transitive, excluding temporal information. We chose to exclude the temporal information of the documents for these connections, as the month names occur frequently throughout the dataset, creating large connected groups based on solely this information.

Transitive Terms We analyze the terms that cause links between documents within the cluster to determine which terms cause the transitivity. We call these terms *transitive terms.* Our goal is to understand the distributional properties of the transitive terms, and how often they occur within the cluster compared to occurring in the remaining dataset. The first statistic we look at is the proportion of documents in which the transitive term occurs. We observe that the proportion of documents with a transitive term within the user-generated cluster is 20% higher, on average, than the proportion of documents outside the cluster that contain the term ($t(2442) = -46.50$, $p < .0001$). Transitive terms have very low rates of occurrence outside of their clusters, and in some cases the only occurrences are within the single cluster.

4.1.3 User-Generated Cluster Labels

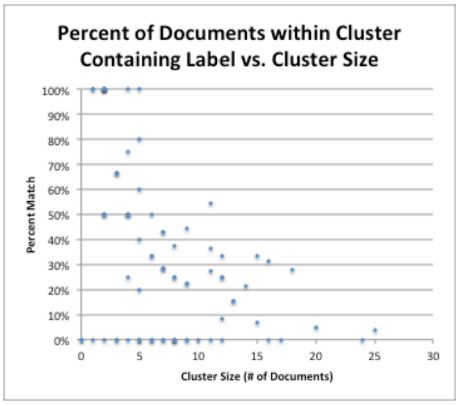

Figure 4 The size of a cluster compared to the percentage of documents within the cluster that contain the user-generated label. Notice that only clusters of 5 or fewer documents match 100%.

How do documents within a cluster relate to the cluster label? To determine if the labels can provide an indicator as to which terms within the cluster are important, we compare the user-generated cluster labels to the content of the documents within the cluster. For example, for a cluster named "Germany and Trucks", we extract the entities "Germany" and "Trucks". Then, we analyze the percentage of documents within the cluster that contain the word "Germany" and the percentage containing "Trucks" (case insensitive and stemming). We report on the highest percentage of these, as we are not concerned with choosing an entity from the label that best represents the cluster of documents (addressed by work such as [20]). Rather, we present the results of how well the best-matched entity within a label matches the entities of the documents within a cluster.

The percentages of documents within each cluster that contain the best-matching entity from the label are shown in Figure 4. These results show that 12 of the 86 clusters (14%) can be characterized based on a single entity extracted from the user-generated label (i.e., 100% of the documents in the cluster contain the given entity). 10 of them are clusters of two or fewer documents (shown in Figure 4). Whereas 67 of the 86 clusters (78%) do not contain the given entity in more than 50% of the documents within the cluster.

Additionally, the users who chose temporal clustering to organize their workspace still showed inconsistencies between their temporal clusters and the documents actually contained in those clusters. For example, one user was very careful to maintain a relative ordering between documents based on the date included in each document. However, analyzing his layout, this ordering did not hold true for the majority of the layout. Another user chose to cluster the information through a broader temporal criteria (i.e., he clustered based on the months the documents occurred). However, 3 of his 5 clusters contained documents from months other than the month with which he labeled the cluster.

From these results, we confirm our hypothesis that users form clusters not solely based on entities or keywords within the data. Cluster labels such as "important people", "unknown", "events that have happened", "random unrelated events", "miscellaneous", "terrorist activity timeline", "big events in southern cities", etc. indicate they are based on higher-level or process-oriented concepts. Further, we found that users struggled to answer what is the meaning of their clusters. This could be because clusters were created based on implicit and informal relationships perceived by the users (as described in [1]). Thus, asking users to formalize these relationships proved challenging.

4.2 Analysis of Process

How can interactions provide effective discrimination of relevant structure? We analyze each user's analytic process in terms of the user interactions performed in LightSPIRE. Our goal is to gain a better understanding of how each interaction is used during the sensemaking process, and how models might exploit these interactions as a means for unobtrusively capturing information from the user about important discriminating features of the data.

4.2.1 Search

Search is a frequent operation in text analytics. Performing a search in LightSPIRE returns results visually within the layout. That is, documents containing the search result change color to red until the search is cleared. Even after the search is cleared (or another search is performed), the text matching the search query within the documents stay highlighted in a neon green. We divide the use of search into two categories: *constructive* and *awareness*.

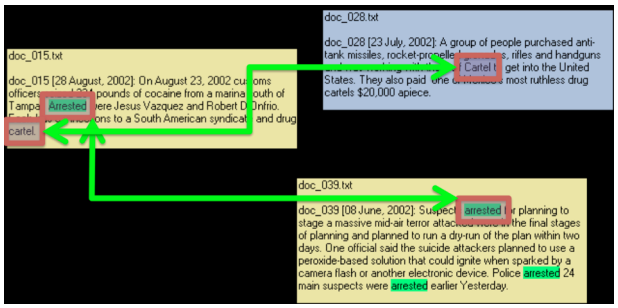

Figure 7 Example of a cluster that can be described by transitive relationships (shown by arrows). While a single term is not present in all three documents, we can form transitive connections between the documents via the terms "arrested" and "cartel".

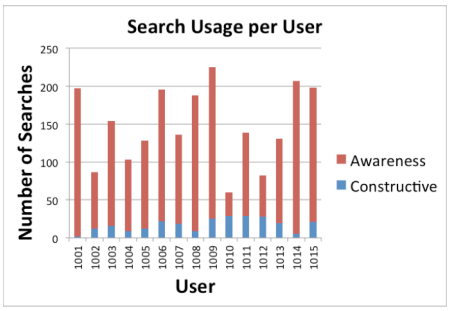

Figure 6 Users performed searches during their investigation for two reasons: constructing clusters (constructive), or to recall where the search term appears in the spatial layout (awareness).

Constructive search indicates that the results of the search were used to create a cluster, whereas awareness search was performed to highlight where in the layout a term occurs.

Users performed a total of 2263 searches (broken down by user in Figure 7), 207 of which were constructive (9%), and 2056 of which were awareness (91%). A total of 326 unique terms were used in the search. Thus, many were repeated, as evidenced by the high number of awareness searches performed. Of these search terms, 222 contained a one word, 100 contained two, and only 4 were three words in length.

A constructive search consisted of performing a search, then creating a cluster based on the documents in which the search term appeared. For example, one user found the term "u-haul" interesting while reading a document. He proceeded to search on this term, found that it appears in other documents, and dragged each of these documents to a location to "construct" a cluster.

This usage pattern for search might initially indicate that clusters are formed as a result of search terms, and therefore can be classified by a collection of entities. However, the structure of the clusters often changed during the investigation as the user gained more insight into the dataset. Clusters changed from their initial creation based on an entity (e.g., the "u-haul" cluster, containing only documents containing that entity), to a collection of documents whose connection or similarity is not based on that particular entity (e.g., the "transportation of suspicious material" cluster). This is evidence of incremental formalism [21].

Search can provide a good indicator as to what documents are important. We analyzed all the search hits (i.e., a document containing the search term is considered a search hit), and with the list of important documents provided with the dataset, found that the average number of times an important document was hit was higher than the non-important documents (Figure 8). The average number of times an important document was hit by each user is 14.2 times, compared to 6.9 times for non-important documents ($t(28) = 4.47$, $p < .0001$).

4.2.2 Highlighting

Analysts frequently highlight information while reading. LightSPIRE allows for two types of highlighting. When users perform a search, the text within each document that contains the search term is highlighted green. Also, users can perform a standard yellow highlight of a phrase within a document using their cursor.

The design decision to create persistent highlights from search terms stemmed from the user feedback from a previous study [3], where the users mentioned that creating highlights within documents served as a means for not only marking important information within the documents, but also created non-uniform visual representations of these documents. That is, the highlights served as a way to transform the documents into visual glyphs, as the pattern of highlights within a document was meaningful to the user. 9 of the 15 users made use of standard highlighting, while 6 used only the highlights from search.

We found these two types of highlighting were used to indicate relevance at two different scales. Search terms were more concise indications of terms or entities that the user found interesting and relevant. This is evidenced by the analysis of search term length, showing that users searched mostly to find single words. In contrast, the standard form of highlighting was used to indicate broader portions of documents as important (e.g., sections or phrases). Users performed a total of 220 highlights, containing an average of 5 words per highlight (sometimes spanning entire sentences). One user even chose to perform a standard highlight spanning an entire document that he referred to numerous times, and wanted to "find [the document] more easily".

We analyze the standard highlights with respect to the cluster labels to determine if the labels match to the highlights. Only 9 of the 86 cluster labels contain entities that were highlighted by the users in the documents. This shows that while highlighting can indicate content relevant to the user, cluster structure is more complex.

4.2.3 Document Movement

Being a spatial workspace, one of the most predominant user interactions is the movement of documents to position (and re-position) documents throughout the analysis. As expected, users positioned documents within their workspace as a means of

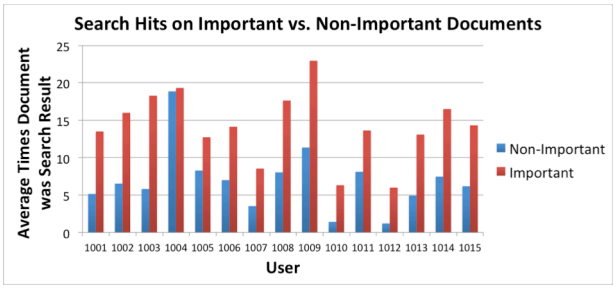

Figure 8 Comparison showing how often important documents were search result hits compared to non-important documents.

externalizing insights about the datasets [3]. However, in this study we are more interested in what information we can quantify about this interaction regarding the user's analytic reasoning.

The analysis of movement was performed based on the number of times a document was moved, and the average distance each document traveled per move (in number of pixels). Important documents were moved an average of 7.1 times, compared to 5.7 times on average for non-important documents ($t(28) = -1.63$, $p < .05$). While important documents were moved more frequently, their moves were more local, indicated by the average path length (in pixels) the document traveled each time it was moved. An important document traveled an average distance of 654 pixels per move, compared to an average of 792 pixels for non-important documents ($t(28) = -1.65$, $p < .05$). Thus, important documents were moved 25% more times, but 17% less distance per move.

Documents displayed in full detail versus the smaller, minimized views reveals a metric for discriminating between important and non-important documents. Given the added resolution and size of the display used, 12 out of 15 users chose to maintain all documents in full detail. The three who minimized some documents only did so for un-important documents.

While these metrics were statistically significant, the most notable difference between important and non-important documents in terms of movement were seen through the observations and post-task interviews. We observed that the important documents served as spatial landmarks for the users. That is, these documents anchored a concept to a specific location in the workspace, from which the remaining layout crystalized. The typical behavior observed for moving important documents was to perform one large movement to position the document in the workspace, with many future short movements to refine the information within the cluster. In contrast, users quickly deemed non-important documents as irrelevant, placing them in such a cluster (e.g., "junk"). Other times, users did not refine the positioning of these documents within a cluster, but rather repositioned them into new clusters, often distant from the previous positions.

5. Discussion

The results of this study reveal new opportunities in the area of statistical models designed for co-creating spatial layouts. We initiate a challenge to statistics and data mining researchers to design models to support the interactive sensemaking process. First, designers can use the structure we analyzed from the layouts users created to design algorithms that better mimic users' clusters. For example, transitivity is a good metric in that it successfully extracted structure from the user-generated clusters. Therefore it could provide a good metric for use in spatial layout algorithms. In contrast, algorithms based on strict term co-occurrence between documents are not likely to coincide well with user's mental models. Algorithms can be designed to support the three layout strategies observed. To support incremental formalism, models can evolve from term co-occurrence to more complex metrics over time, such as transitivity.

Second, the user interactions present in the spatial sensemaking process can be used to guide models during the analytic process for co-creation of the spatial layout. For example, algorithms can observe and incrementally respond to the process of users clustering data. When users perform sensemaking, they gain understanding of the data at a higher level. Models must be able to co-create clusters based on these higher-level concepts. These concepts are based not solely on term co-occurrence, transitivity, or other metrics, but incorporate the user's reasoning. The user interactions can serve as cues to help models understand these higher-level concepts.

For example, models can expand the "data" upon which these models calculate their similarity measures – broadening the scope of the distance metric. These models should incrementally adapt based on the interactions of the user throughout the analytic process. To do so, models must be based not solely on the *hard data* (i.e. the structure within the dataset), but also the user's reasoning derived from interaction (i.e., *soft data*). Soft data is defined as a captured and interpreted representation of a user's semantic knowledge regarding a dataset [6].

As evidenced by the results of this study, the user-generated layouts are often based on information that is *outside the scope of the hard data*. For instance, the user-generated cluster labels do not always map directly to a set of entities within the dataset, implying a need to add this information to the model. Cluster structure was not obvious until users identified and labeled the clusters, but was an important part of their sensemaking process. Knowing which of their three spatial strategies the user has chosen would help models understand the meaning of the clusters. Some soft data, such as search terms, can help distinguish between what hard data is relevant and not. Search terms can help indicate both what documents are important (based on being a more frequent search result), as well as which terms (or entities) to weight more heavily (indicated directly from the search terms). Document movement can be an indicator of not only similarity, but the pattern of movements can indicate the importance of the document. Users' cognitive similarity metrics are not limited to term co-occurrence or transitive relationships. This may indicate that users develop similarity based on higher-level concepts. Sometimes highlighted phrases were an indication of a user's reasoning, based on cluster labels.

In contrast to the results from Dou et al. and Chang et al. [15, 16] who successfully recovered reasoning from user's interactions, our measures indicate that doing so systematically yields lower probabilities. However, we have confirmed that analysts encode meaning into spatializations through complex spatial structures, using a rich set of cues. We can detect hints of meaning through these rich cues, such as the spatial layout and the interactions. All of them provided some benefit, but no single one gave an absolute indication of reasoning. Thus, a probabilistic approach that integrates all of them is the most likely path for success. A tactful combination of the soft data can be exploited by clustering algorithms to help guide and enhance their outcome, incrementally during the course of interaction.

For example, an algorithm can exploit document movement in a spatial metaphor to learn and incrementally update similarity measures within a dataset. Observation-level Interaction [22] uses this form of soft data to couple the movement of data within a spatialization with updating parameters of popular clustering algorithms. In these models, users are given the ability to interact within the visualization, rather than directly with visual controls of parameters of the statistical model. While doing so, it is the responsibility of the model to update the parameters that correspond to the manipulation within the visualization. This is similar to the concept of metric learning, where models adjust the weighting of dimensions according to the user's input [23]. As another example, semantic interaction [6] exploits document movement, highlighting, annotating, and search to update the model and co-create a spatial layout. The system interprets the interaction and updates the layout incrementally.

6. Conclusion

In this paper we present the results of a study observing users analyzing a textual dataset spatially. We analyze the final layouts created by the users, and the captured user interactions performed while generating the clusters. We found how specific criteria within this process (including both the generated clusters and the interactions used) can indicate important and discriminating structure within the dataset.

Through analyzing the clusters created by the users, we found that only 15% of the 86 clusters contain at least one co-occurring term in all the documents within the cluster. Instead, we found that users tend to create clusters using transitive relationships between documents within a cluster. The challenge then, is determining which terms to use to create these relationships. Many of the clusters users created are based on higher-level or process-level concepts during sensemaking. Thus, these concepts rarely relate directly to keywords, making simple term co-occurrence metrics less useful.

The interactions performed by the users (i.e., document movement, highlighting, searching) in spatially analyzing the dataset can provide indicators towards what structure within the dataset is important (or discriminating) to the user. For instance, important documents were returned as search results more frequently than non-important documents. Further, users' highlights sometimes indicated terms or phrases within a document that are important to the cluster definition.

This collection of interaction data, referred to as *soft data,* can be vital to unobtrusively gain an understanding of what aspects of a dataset a user finds important. As such, by incorporating both *hard data* (extracted directly from the dataset) and *soft data*, models can calculate more useful similarity metrics for users, and ultimately generate layouts from which users can gain insight.

7. Acknowledgements

This research was supported by NSF FODAVA grants CCF-0937071, CCF-0937133, and the DHS VACCINE center.

8. References

[1] Marshall, C. C. and Rogers, R. A. Two years before the mist: experiences with Aquanet. *Proceedings of the ACM conference on Hypertext* (Milan, Italy, 1992). ACM, 53-62.

[2] Andrews, C., Endert, A., Yost, B. and North, C. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10, 4 (2011), 341-355.

[3] Andrews, C., Endert, A. and North, C. Space to Think: Large, High-Resolution Displays for Sensemaking. *CHI* (2010), 55-64.

[4] Pirolli, P. and Card, S. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Though Cognitive Task Analysis *Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, Virginia*(2005), 6.

[5] Thomas, J. J. and Cook, K. A. *Illuminating the path*. IEEE Computer Society, 2005.

[6] Endert, A., Fiaux, P. and North, C. Semantic Interaction for Visual Text Analytics. *CHI* (2012).

[7] Skupin, A. A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications*, 22(2002), 50-58.

[8] Kaban, A. A Scalable Generative Topographic Mapping for Sparse Data Sequences. *International Conference on Information Technology: Coding and Computing (ITCC'05),* (2005).

[9] Xu, R. and Wunsch, D., II Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16, 3 (2005), 645-678.

[10] Alonso, O., Gertz, M. and Baeza-Yates, R. Clustering and exploring search results using timeline constructions. *Proceedings of the 18th ACM conference on Information and knowledge management* (Hong Kong, China, 2009). ACM, 97-106.

[11] Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V. Visualizing the non-visual: spatial analysis and interaction with information for text documents. *Readings in information visualization: using vision to think* (1999). Morgan Kaufmann Publishers Inc., 442-450.

[12] Pike, W. A., Stasko, J., Chang, R. and O'Connell, T. A. The science of interaction. *Information Visualization*, 8, 4, 263-274.

[13] Yi, J. S., Kang, Y. a., Stasko, J. and Jacko, J. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13, 6 (2007), 1224-1231.

[14] Card, S. K., Mackinlay, J. D. and Shneiderman, B. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[15] Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R. and Chang, R. Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications*, 29(2009), 52-61.

[16] Chang, R., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D. and Sudjianto, A. WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions. *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology* (2007). IEEE Computer Society, 155-162.

[17] Shupp, L., Andrews, C., Dickey-Kurdziolek, M., Yost, B. and North, C. Shaping the Display of the Future: The Effects of Display Size and Curvature on User Performance and Insights. *Human–Computer Interaction*, 24, 1 (2009), 230 - 272.

[18] Hossain, M. S., Andrews, C., Ramakrishnan, N. and North, C. Helping Intelligence Analysis Make Connections. *Workshop on Scalable Integration of Analytics and Visualization* (San Fransisco, 2011).

[19] Hossain, M. S., Gresock, J., Edmonds, Y., Helm, R., Potts, M. and Ramakrishnan, N. Connecting the Dots between PubMed Abstracts. *PLOS One*, 7, 1 (2012), e29509.

[20] Rose, S., Engel, D., Cramer, N. and Cowley, W. Automatic Keyword Extraction from Individual Documents. *Text Mining* (2010). John Wiley & Sons, Ltd, 1-20.

[21] Shipman, F. and Marshall, C. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Comput. Supported Coop. Work*, 8, 4 (1999), 333-352.

[22] Endert, A., Han, C., Maiti, D., House, L., Leman, S. C. and North, C. Observation-level Interaction with Statistical Models for Visual Analytics. *IEEE VAST* (2011), 121-130.

[23] Xing, E. P., Ng, A. Y., Jordan, M. I. and Russell, S. Distance Metric Learning, with Application to Clustering with Side-information. *Advances in Neural Information Processing Systems 15* (2002). MIT Press.