Solving Intelligence Analysis Problems using Biclusters

Patrick O. Fiaux

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfilment of the requirements for the degree of

Master of Science
In
Computer Science & Applications

Naren Ramakrishnan, Co-chair
Christopher L. North, Co-chair
Manuel A. Pérez-Quiñones

January 27 2012
Blacksburg, VA

Keywords: Biclustering, Visual Analytics

Solving Intelligence Analysis Problems using Biclusters

Patrick O. Fiaux

ABSTRACT

Analysts must filter through an ever-growing amount of data to obtain information relevant to their investigations. Looking at every piece of information individually is in many cases not feasible; there is hence a growing need for new filtering tools and techniques to improve the analyst process with large datasets. We present MineVis – an analytics system that integrates biclustering algorithms and visual analytics tools in one seamless environment. The combination of biclusters and visual data glyphs in a visual analytics spatial environment enables a novel type of filtering. This design allows for rapid exploration and navigation across connected documents. Through a user study we conclude that our system has the potential to help analysts filter data by allowing them to i) form hypotheses before reading documents and subsequently ii) validating them by reading a reduced and focused set of documents.

ACKNOWLEDGMENTS

I want to thank my advisors Chris North and Naren Ramakrishnan for all their support, patience and guidance on my various projects. I also thank Manuel Pérez-Quiñones for introducing me to the MVC paradigm in his user interface software course, as it became invaluable knowledge in designing and building MineVis. I thank Alex Endert for relinquishing his office for me to run my study on this system and always being there to answer questions.

This work was supported by US NSF FODAVA grant CCF-0937133 and the Institute for Critical Technology and Applied Science (ICTAS), Virginia Tech.

All photos by author unless otherwise indicated, 2012.

Table of Contents

Table of Figures

Table of Tables

INTRODUCTION

As facets of our world continue to become digital, the amount of information being created and gathered increases. The task of analysing this growing amount of information in a limited amount of time becomes more difficult. Analysts seek the help of visual analytic technology, combining the computational benefits of statistical models and data mining with the innate cognitive abilities of humans to support sensemaking [1, 2], as shown in Figure 1. The challenge then, is to tactfully design a system that integrates these two areas, creating a usable workspace for analysts to perform sensemaking. In this thesis, we present a prototype mixed-initiative system, MineVis, where data mining and sensemaking are tightly integrated. We also discuss the findings of a user study investigating the effectiveness of MineVis for text analytics.

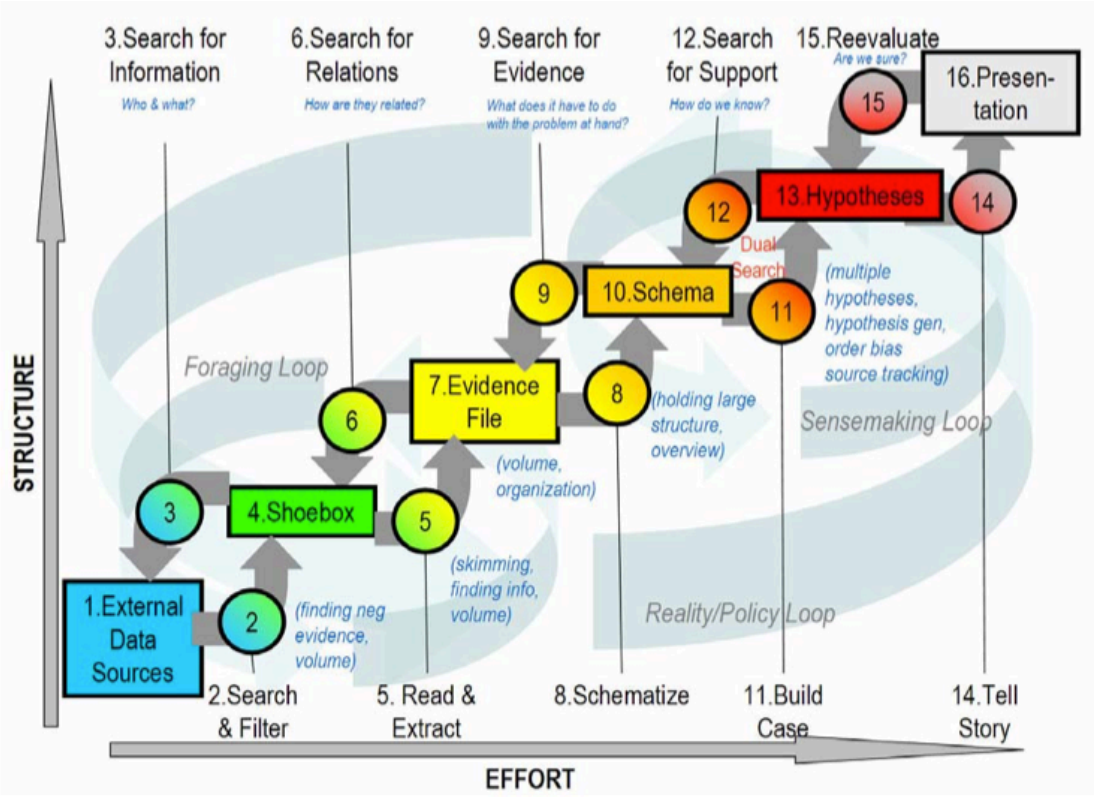

Figure 1: The sensemaking loop from [Pirolli and Card 2005]. Used under Fair Use, 2012.

MineVis enables users to explore textual datasets through biclusters in a spatial workspace. Intuitively, biclusters can be viewed as visual glyphs in a matrix, indicating relationships between entities; see Figure 2 for an example. Biclusters are explained in more detail later. Thus, biclusters represent relational information between entities in the dataset.

Users can organize the biclusters in combination with the raw text documents in a spatial workspace (see Figure 8 and Figure 10). Previous work has shown that the flexibility afforded by space provides a rich medium for representing insights [3, 4, 5]. Manually organizing information into clusters or groups does not require users to formalize explicit relationships between information until these relationships are apparent to the user, a concept referred to as "incremental formalism" [4]. Endert et al. have shown that while detectable structure may exist within these clusters, the fundamental "structure" upon which users generate these clusters is often at a meta-level, where the semantics of the text is critical [5]. Thus, we are interested in how the highly structured relational information of biclusters can be used in conjunction with the documents in a flexible spatial workspace.

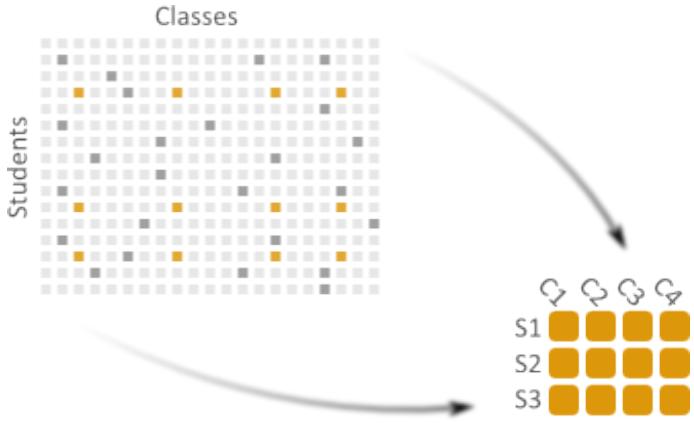

Figure 2: Example bicluster extracted from a student to classes relationship. Dark cells represent relationships, orange cells represent relationships part of this specific bicluster.

The results of our study suggest that MineVis can provide helpful support for the sensemaking process. Biclusters proved to be a rich visual encoding for information both during the analysis, as well as during the reporting of the findings after the study. They were widely used during the analysis, ranging from a

quick overview of relationships between entities (or groups of entities), to being the focal point of entire hypotheses in the dataset.

1 Related Work

MineVis draws heavily from both data mining and visual analytics in its design and construction. The first subsection below looks at how MineVis builds upon prior research but also brings in new concepts. The second subsection discusses some of the algorithms that are integrated into MineVis and defines key concepts of compositional data mining we used in system design.

1.1 Information Visualization Systems

Information Visualization plays a major role in the design of MineVis. This section reviews other works in this space and how MineVis is inspired from them or differentiates itself.

MineVis is not the first web service-based analytics system. The Scalable Reasoning System (SRS) by Pike et al. already demonstrated many of the advantages of web service-based systems [6]. A main focus of SRS is to support distributed analytics across many devices in real time. MineVis also takes advantages of the web architecture to allow for anywhere access; however it focuses more on performing heavy computations on the server and offering a scalable browser interface to the user and only has limited distributed analytics capabilities.

Many systems have focused on assisting the text analytics process for document-based datasets. Tiara by Wei et al. aids users in exploring a data set through a responsive keyword visualization linked to the source documents [7]. Jigsaw allows users to explore textual datasets by exposing relationships between entities through a collection of views [8].

MineVis has been heavily influenced by special layout visual analytics tools such as Analyst's Workspace (AW) [9] and ForceSPIRE by Endert et al. [10, 11]; both of these systems have a heavy focus on spatially arranging documents across the display space. ForceSPIRE focuses on creating a dynamic workspace based on semantic interaction where the users input is integrated with the source data to update the layout in real time based on the user's actions[10, 11]. AW's goal is to "provide an

environment that unifies the activities of foraging and synthesis into a single investigative thread" [9]. MineVis aims to harness the power of data mining techniques to improve the ability to work with large data sets by incorporating biclusters.

Hybrid matrices and node-link diagrams are similar to the linked biclusters in MineVis. NodeTrix, the work of Henry N. et al. allows exploration of social networks through a hybrid visualization of adjacency matrices and node-link diagrams [12]. This enables clustering and linking clusters to explore a singled relationship, such as co-authorship between authors. NodeTrix generates initial clusters for the user and then allows them to group or ungroup nodes at will as they interact with the layout. OntoTrix by Bach et al. extended this technique to work with ontologies with multiple types of relationships [13]Thus allowing clustering and linking nodes of different types on the same graph. MineVis is different in that all the clusters are generated and linked by compositional mining algorithms and that the user cannot edit them, only links between them. Another important difference is that documents used to generate these clusters are also included and linked allowing for the user to not only explore the relationships and see the big picture but dig into the details and get back to the source of the relationships in a seamless environment.

1.2 Entity Extraction & Co-Occurrence Modeling

Entity extraction can be a difficult task depending on the requirements. MineVis tasks require more than just a list of entities for a dataset; the type of the entity and co-occurrence within a document were also very important. Jigsaw has an extensive entity parsing utility for text documents that we used to extract the entities for out datasets [8]. Jigsaw generates an xml-based file with per document entity lists of several types including people, locations, organizations, dates, and money. MineVis loads the Jigsaw export file into a database by conversion the entities, the types and the co-occurrence into a relational format. This allows MineVis to retrieve document, entity or references quickly with simple queries.

1.3 Compositional Data Mining

"Compositional data mining (CDM) is used to identify relationships among sets of entities across the database schema, where these sets are mined automatically and not defined a priori". [14] CDM has been successful in providing insight to massive datasets in the domain of functional genomics [14, 15, 16]. MineVis exploits concepts from CDM to provide insight into a document based dataset.

The key data structure at the heart of MineVis is the bicluster. Biclusters can be generated using a set of relationships between two domains. Each bicluster "is a subset of rows along with a subset of columns with the property that each row element is related to each column element" [16]. For example consider a relationship between students and classes as shown in Figure 2. We could extract several biclusters from it; one is shown in orange on the image as a subset of 3 rows and 4 columns. This bicluster groups a set of students and the classes they take; inferences can then be made about its content. For instance, perhaps these students are in the same research group due to their commonality in taking a required set of courses.

In MineVis source documents are used to generate entities and can be inherently related to biclusters. As described in Section 1.2, the entities are extracted from documents and sorted into several predefined domains. Using co-occurrence within documents we then build pairs of entities in different domains or relationships; each document can then be represented as a set of relationships involving two or more domains.

Consider an example scenario: a document gives an account of five individuals meeting in one location to plan their trip to travel another location. Then consider we have a second document accounting for how these five individuals leave for their trip from a third location. From these two aspects, we can generate five people and three places as well as a relationship from each person to each place in the second domain even though they were not from the same documents. Using this representation of documents we can relate them to biclusters that share common entities. Since biclusters are sets of

relationships across one of more documents they form indirect connections between documents. Since the relationships in a bicluster are only from two domains, they allow grouping of documents through the set of relations of those two domains. Going back to our example we would have a bicluster of those five people and those three locations which links to two distinct documents. We therefore have the ability to go from a bicluster to a set of documents, or from a document to a set of biclusters by using these relationships as links.

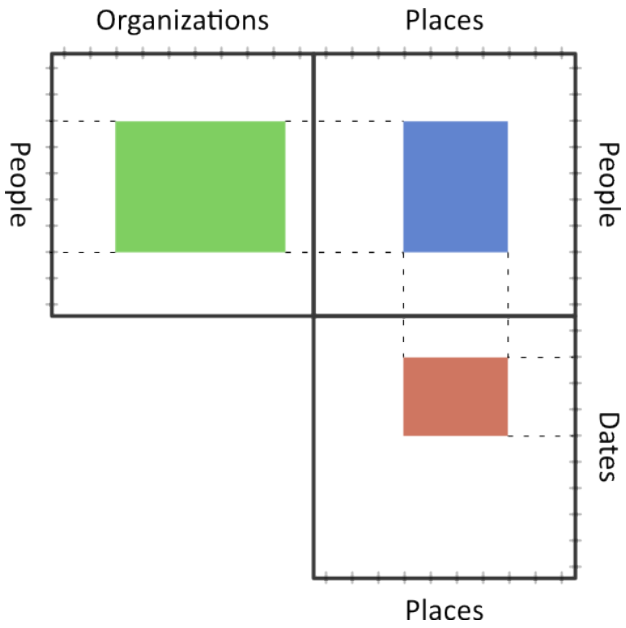

Figure 3: Realizing compositional data mining using biclusters on multiple domains.

Now let us illustrate CDM using the previous scenario. We had a set of documents and relationships between people and places. Imagine that we had more documents and extracted two more domains (dates and organizations), and generated the resulting relationships with the other domains. We can then model the scenario with three biclusters: one between organizations and people, the second between people and places and the third between places and dates.

We can connect the people side of the first bicluster to the people side of the second using a redescription. Then we can connect the places side of the second to the places side of the third bicluster. This creates a link between organizations and dates illustrated in Figure 3; in this example perhaps all

the people belong to the same organization and they are taking a trip to on a specific date. The results of such compositions can be read sequentially from one end to the other, not unlike a story. For instance in the scenario above, we might find that 'research members of the CS department are planning a trip to Austin Texas in May 2012' which might lead us to infer 'their work got accepted to CHI 2012 and they will attend the conference to present it'. We can then look for the documents related to these relationships and use them to look for evidence of our theory. By making use of CDM across multiple domains we become able to generate hypothesis about relationships that can originate from multiple documents without the need to open them.

1.3.1 Algorithms

This section describes the algorithms that can be used to generate data such as biclusters and chains of biclusters described above.

Closed item set mining algorithms are used to generate biclusters. We used two available implementations of these algorithms: Charm [17, 18, 19] and LCM [20, 21, 22, 23].

Chaining of biclusters side by side can be achieved through a combination of methods. MineVis handles the pre-processing tasks of grouping biclusters by common sides. The next step is to generate a look up table for similarity between biclusters and store its results into the database. We used a Cover Tree, an efficient data structure for calculating nearest neighbours [24], to load all the biclusters and then find the nearest neighbours of each distinct bicluster. We used WEKA's Java implementation of the Cover Tree [25] and wrote a customized distance function to measure the similarity of two biclusters. We used the Soergel distance metric as shown in Figure 4, to calculate the distance between two biclusters. It allowed use to compare the ids of the entities in the common sides and generate a normalized measure [26]. MineVis generates a Cover Tree with the biclusters on each domain. For each it generates a look up table to allow constant time access to lookup the nearest neighbours of any bicluster.

$$\frac{(b+c)}{(b+c+d)}$$

Figure 4: Soergel distance: a = in both samples, b = only in one sample, c = only in the other sample, d = in none of the samples.

2 DESIGN & IMPLEMENTATION OF MINEVIS

2.1 System Overview

MineVis takes the user from a raw data set through data mining to a visual analytics interface. The workflow or pipeline consists of four main steps; project configuration, mining, chaining and visualization as shown in Figure 5. The following sections will explains the four steps as well as how they were connected into a seamless analytics environment as well as some design considerations.

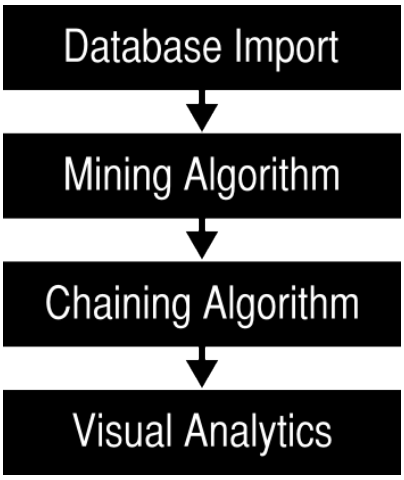

Figure 5: MineVis projects pipeline.

2.2 Design Considerations

We designed the system as a web service-based application. This design choice was driven by two main concerns. First data mining can be quite computation heavy, requiring a lot of time and resources; the client & server architecture of the web enables us to run the expensive computations on the server side and keep the load on the client light.

Doing so has several advantages: it allows the user to access the system with a light client such as laptops, tablets or mobile devices. With a responsive web design the system can also scale the elements on screen to adjust to any display from a mobile phone to a large high-resolution display. And since the process is running remotely the user needs not worry about losing progress due to power loss.

Security concerns can also be addressed through different configurations. The benefit of 24 hour access from anywhere in the world might be out-weighed by the sensitive nature of the data used. In such a

case MineVis can be hosted on an internal server or even on a network physically separated from the web, all the while retaining the ability to use powerful servers for computation.

Another important advantage is modularity. MineVis is built as a modular system in more than one way so that components can be replaced without affecting the overall goal. This is to add flexibility for three different purposes.

The first is scalability. Since the server, and not the user, is in control of the data mining it can make decisions about how to handle incoming jobs. Currently the server hosts the database, the web service and the data mining. However it can be extended, so that the server might be the head node of a cluster or have access to a cluster, cloud or super computer allowing it to offload jobs to different machines or run them simultaneously. This can be done without the need for user input.

The second is customization. The pipeline is a set of modules, and each step can be made of multiple alternative modules. Currently MineVis already supports two mining algorithms, with the possibility for more. Although it only has one chaining algorithm again there is the possibility for more and the same holds true for visualization workspaces. The project configuration can be extended to support multiple databases and the visualization step can also implement multiple views. This allows the system to be tailored for the task or data at hand.

The third is the ability to work with multiple datasets and databases. This will be discussed in more detail in the next section, Project Configuration.

The implementation was done using open source technologies and frameworks both on the server side and user side [27, 28, 29]. This allowed for rapid development of a large complex system while also insuring deployment options on common web servers and easy of update.

2.3 Project Configuration

This step consists of setting up MineVis with a dataset (relational database). MineVis has a central database, which is used for configuration, data mining and result storage. The datasets to be analysed are

stored in separate databases. This allows for the possibility to host them anywhere or for users to use their own database as source data on the system. However this means that each project needs a database to connect to as well as some knowledge of what tables to use. The configuration allows the user to select and configure how tables are related and through which fields. This step enables MineVis to do two key things, first use the database to generate input for mining algorithms, second to load the right labels, names and texts when displaying the mining results.

2.4 Mining

Mining in MineVis is the step where the biclusters are generated for a given project. Multiple minings can be created for each project to accommodate different settings or algorithms. Once a mining is created is allows for selecting an algorithm and then configuring parameters.

Once configured all a user has to do is hit "run"; after that the server takes over and displays a status. Figure 6 shows the status display of a mining under process; the progress bar displays how many relationships have been mined. Each unique pair of entity types is considered a relationship and has to be mined separately to generate complete results. These mining jobs are used to display the status for the user. Input data is generated for each job and then the selected algorithm LCM or Charm runs the algorithm and the results are then parsed into the database.

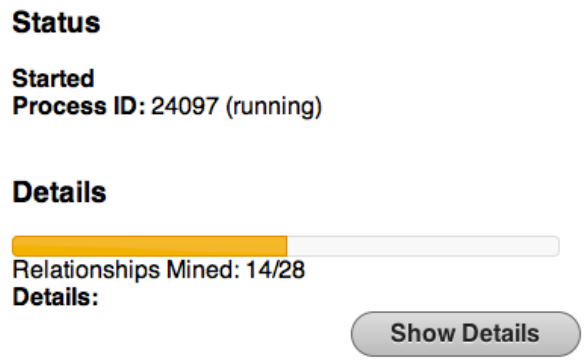

Figure 6: Status of mining algorithm in MineVis.

2.5 Chaining

The third step is to build a set of possible links between the biclusters generated in the Mining step. The configuration process is similar to Mining except for different parameters. The Chaining process itself is also broken down into subjobs, one for each entity type. For each of those types a list of all the biclusters containing it in either its rows or columns is generated. That list in then used to build the input file for the algorithms discussed in Section 1.3.1 to generate the mappings between similar biclusters and the results are stored back into the database.

2.6 Visual Analytics

The last part of the MineVis pipeline is the analytics workspace where the user can make use of all the imported and generated data. The workspace consists of three main areas; the data-browser, the preview pane, and the graph workspace. Each area and its features will be described in the following sections.

2.6.1 Data Browser & Preview

The data-browser allows access to the various data structures generated in the MineVis system as well as the source data. It consists of five different tabs; search, entity frequency list, bicluster list, document list, links list.

The search tab supports two different types of search operations. The first would be a plain text search on the documents of the dataset. The second as shown in Figure 7, is the entity search. The entity search is different from the plain text, as it will return biclusters containing that entity in the results as well. The search bar also features an auto complete function that shows the users a set of possible entities after 3 or more letters are typed. These two search features allow for look up and retrieval of specific information.

The frequency data holds a list of the entities in the dataset ranked by the one appearing in the most biclusters to the least. This tab was typically suggested to the user as a starting point. If they noticed something odd they were able to investigate it further. A single click on a frequency list entry would take the user to the search results for that entry.

The Browse Biclusters tab allowed access to all of the Biclusters manually. They were grouped by domain relationships. This allowed user to browse through a certain category of biclusters like ‘people to money’.

The Browse Documents tab previews a list of all the documents, listed by title and a click would bring the content into the preview window.

The last tab is the Browse Links tab. Its content is similar to that of the bicluster tab as it lists biclusters, and only upon preview can the user see the links between a specific bicluster and others. However the grouping in this tab is done by domain only and not domain relationship. This tab like the previews two was included for completeness sake to allow manual access to the date should the other features break or the users not want to use them.

The preview pane has two main uses: first as its name indicates it allows a previewed of search results. Second it allows adding the currently previewed item into the graph workspace. Thus a user can preview search results and quickly skim them, once they find something of interest they can add it to the graph where they will be able to explore that item and its connections in detail.

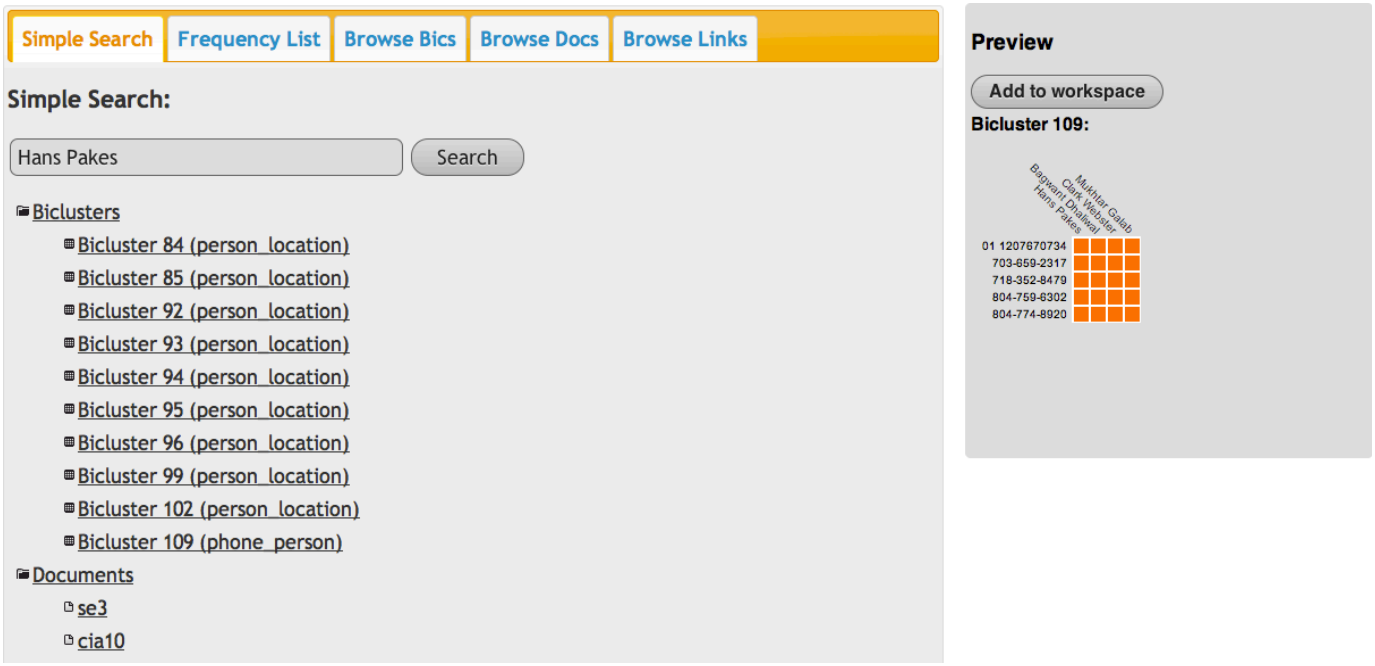

Figure 7: Data-browser showing an entity search and preview window with one of the results loaded.

2.6.2 Graph Workspace

The graph is where the bulk of the investigation happens and also the place where the key MineVis functions are implemented. After creating a new visualization the graph is empty, so the user must go through the data browser and preview at least once to bring data into the graph. Once there is at least one data item, document or bicluster, in the graph it is possible to load more data items without using the data-browser by using the various actions provided via context menus described below.

Some features were available globally across the graph regard less of element type. Biclusters and documents can be dragged for example. The 'Link to…' function from the context menu also works on both, it allows for creation of a user link between elements. User links were blue instead of white to reflect a connection based on analyst insight rather than mining data. Same for the 'Close' menu item, that however works on the links as well. Other features were specific to some elements.

Biclusters have two special actions. The first was 'show documents'. It looks up all the documents containing relationships part of this bicluster and adds them to the workspace; this allows the user to 'dig into the details'. For example in Figure 8 the three documents (in grey) are the result of 'show documents' on the trop right bicluster. The second 'show bicluster links' would load the link data described in the browse link tab for that bicluster directly in the workspace and link it entity by entity to this bicluster. An example of this entity-to-entity link can be seen in Figure 8. It allowed the user to explore indirect relationships between entities.

Documents only have special action, 'show biclusters'. This would load the first fifteen biclusters that contained one or more of the relationships in this document within their cells. Originally this action would load all related biclusters however this caused problems, generic entities such as 'USA' were in many document biclusters, loading up hundreds of biclusters into the workspace.

The last special action is highlight link. This was available for user to make links wider and green to differentiate them from other links on the workspace if they were of particular significance to them.

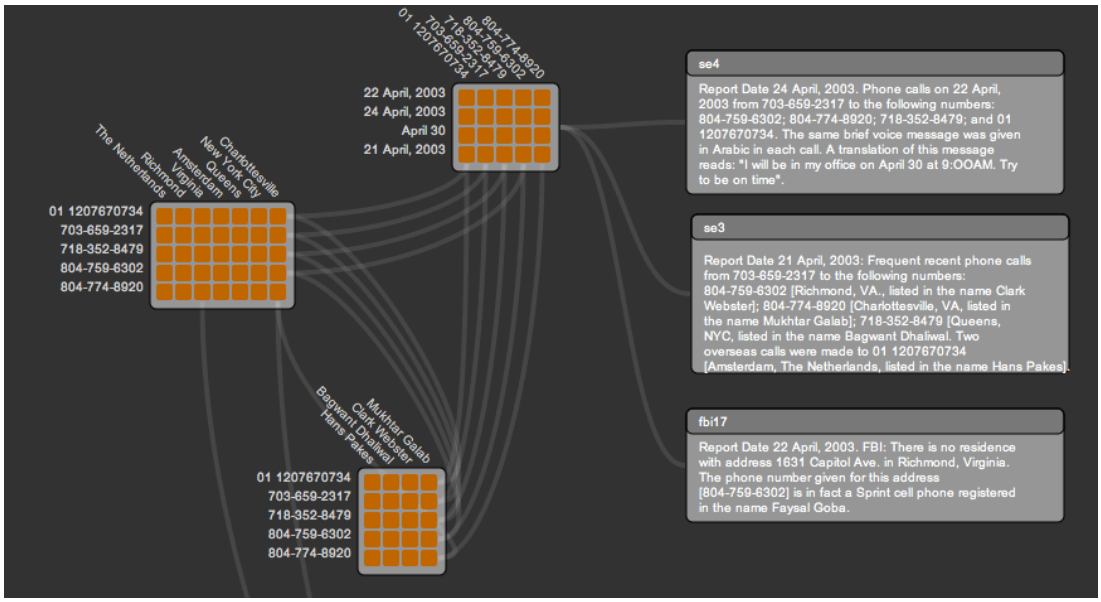

Figure 8: Sample area of graph workspace with biclusters and documents connected.

3 METHODOLOGY

The purpose of this study is to evaluate MineVis for sensemaking of text data. In this study, we ask each user to analyze a set of text documents containing a fictitious terrorist plot. We are directly interested in the following research questions:

Q1: How does the use of biclusters integrate into the story telling and analytics process? Specifically, how are biclusters and documents used in combination during the analytic process?

H1: We hypothesize that users will be able to combine the computed, highly structured relationships between entities (i.e., the biclusters) with the documents in the spatial workspace in order to complete their task. The users may find it helpful to relate across different types of information.

Q2: What insights are grained from biclusters compared to those gained from documents?

H2: Since a bicluster is made of many documents we hypothesize that they will show and highlight the connections between multiple documents. In addition, they can help gain rapid insight compared to reading multiple documents. However we do not expect biclusters to help in gathering evidence as they lack the detail to make compelling arguments.

3.1 Study Setup

We used two Mac Pros to run the system for the users. The first Mac Pro, the server, ran the webserver and MineVis system. The second, the client, was the users machine.

The Client machine was set up as a Large High-Resolution Display outfitted with an array of eight 30’ LCD monitors. This gave the user a workspace of 10’240 pixels by 3’200 pixels to work with. Almost no virtual navigation (scroll bars) was required to navigate the workspace with the exception of the data-browser, which can require scrolling for searches with many results.

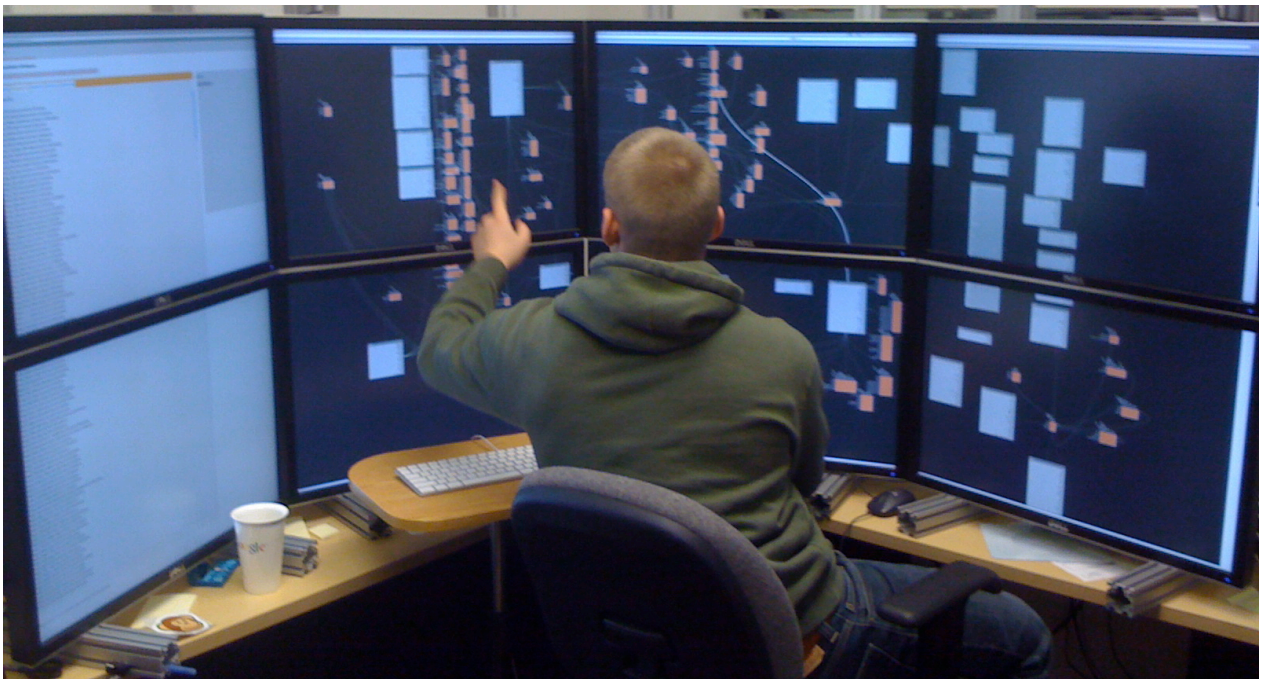

Figure 9: MineVis running on the setup used for the study.

The MineVis workspace components were sized to take advantage of large display space (see Figure 10). The Data-browser and Preview were assigned to the screens on the top and bottom left. The data-browser was on the left and the preview to the right of it at the top of the first screen. The other six screens were reserved for the graph workspace giving the user plenty of space to organize and grow their story graph.

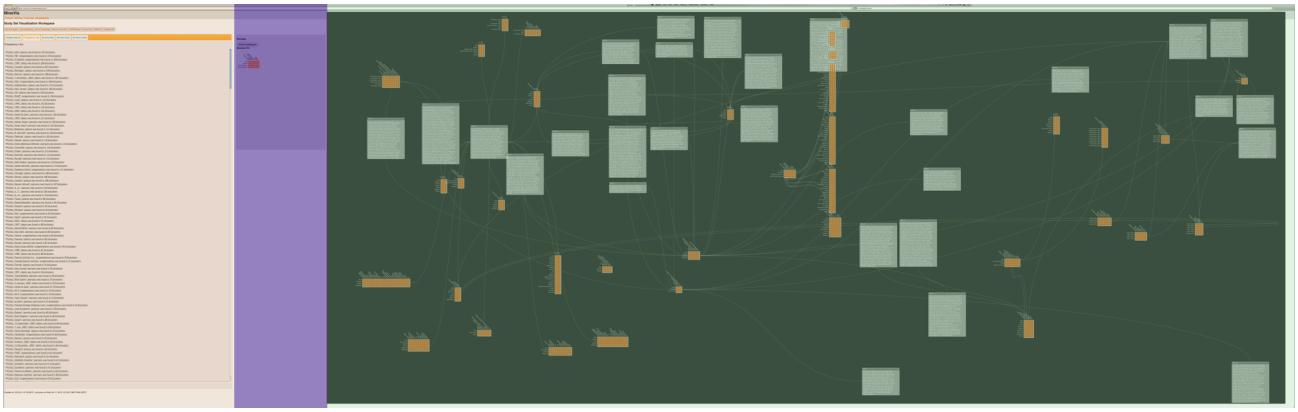

Figure 10: The user workspace with each of the three main components highlighted in colors. Listed from left to right: data browser, preview, and graph workspace.

We used the Atlantic Storm dataset for this study [30]. It consists of 111 documents. The study focused on the analysis and 4th step in the MineVis pipeline; the configuration, mining and chaining algorithms were pre-run for the user. The mining was done with a support of 3 using LCM. This generated biclusters of at least 3 rows and 3 columns. These settings were selected because the individuals planning an attack in the data set rarely meet in large numbers, and hence using a larger support could result in many missing connection. This generated 1001 different biclusters. The Chaining was run with the default settings.

The study procedure consisted of three parts. First we explained the data structures and nature of the dataset and trained them in the use of MineVis on a training dataset. Second in the investigation, the user was asked to assume the role of an analyst and investigate the dataset for any threats of attacks being planned. They were given two hours to investigate and told to be ready to report at the end. The third part was the report and interview session, where we went over their results and recommendations for future actions. Then we interviewed them on their use of MineVis to obtain information on how biclusters affected their analysis.

3.2 User Demographics

The study included five male users between the ages of 25 and 27; each were compensated for participation due to the three hours commitment. Four of the participants were graduate students studying for either a Doctorate or Masters; one had a Bachelors degree but was not currently attending graduate school. Four of them were computer science majors and one was a humanities major. Only two had prior experience with visual analytics tools. Three of the participants wore spectacles and found the text in the graph to be small (for one of them it was actually too small).

4 RESULTS

4.1 Investigation Results

The main goal of the study was to evaluate the use of biclusters within the analytics process, therefore we do not consider that there are right or wrong answers from the users. That said, we give a short description of the plot and talk about what the users found within the allotted time. As aforesaid, the participants were not expected to solve the data set in the two hours of the study. We were however hoping they would have enough time to form some hypothesis about what might be planned and describe the next steps to be taken to validate or further investigate their options.

The dataset consists of documents related to the main plot as well as many unrelated reports. The main plot involves many people and most of which are not aware of the whole plan or what others are doing. The main plot is supplemented by a "distraction" plot, planted by the same individuals in the hope to stray authorities away from their real plan. Many users also found a third plot unrelated to the individuals in the first. The second and third plots might seem compelling however there is no complete solution to them in the documents of the dataset, in a way they are dead ends. Next we look at how the subject fared on their investigations.

Subject 1 found most of the key players in the dataset during the two hours; he was investigating their connection to find their plan. He started with the frequency list on a specific date that seems too frequent. From there he loaded a bicluster and navigated the data set through a combination of show documents and show biclusters within the workspace. Early on he found one man who seemed to be funding an operation and some of the men he was meeting with. The subject did find the distraction plot but he didn't dig into it too much and continued investigating other options. At the end his recommendation were to look into the connections of the man who seemed to fund others: a key figure in the main plot as well as investigate a man who left London, part of the third plot. The subject had

many of the major players in the main plot on his final layout, (see Figure 11) and he also knew they were connected, but only needed to find out how.

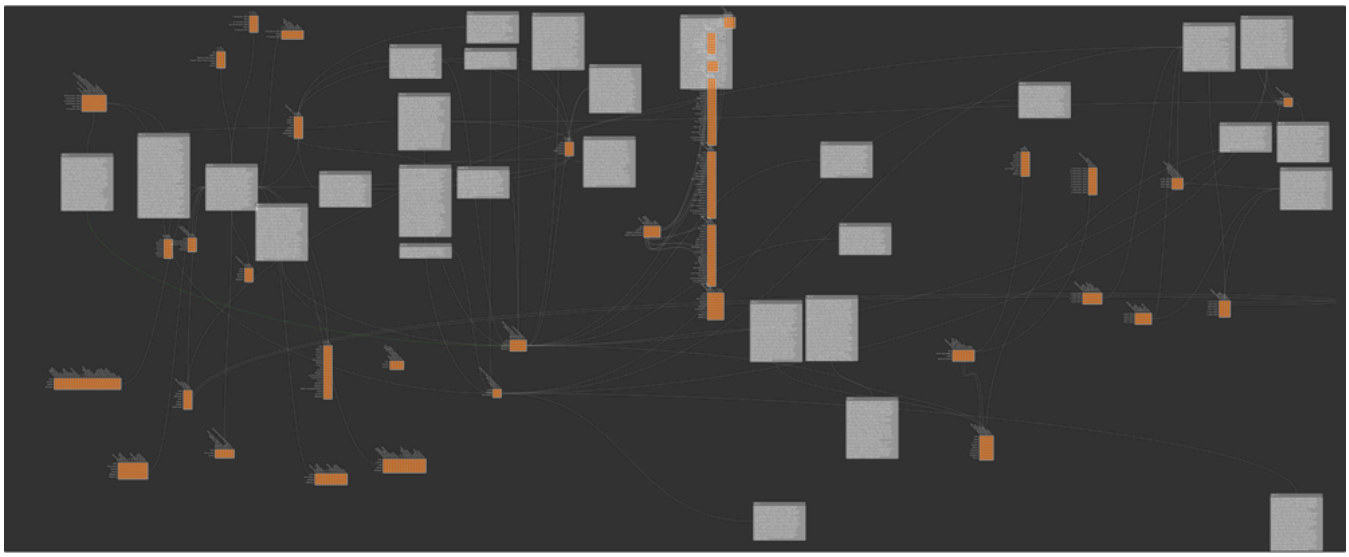

Figure 11: Graph workspace of subject 1 at the end of the study.

Subject 2's main strategy was to follow the money; this in turn led him to most of the main players. At the end the subject was only missing some documents required to connect all the individuals. He started by browsing biclusters with the money type, loaded one, and then displayed its documents to start his investigation. This strategy led him to search for new leads, add biclusters and show documents several times. He was aware of the distraction plot although didn't obtain all the data on it; he investigated the third plot until he hit a dead end. After that he back tracked to the funding and investigated the actions of many of the players. He knew they had supplies but not what, he knew they could ship them and he knew they were planning and meeting but not to what end. His final workspace can be seen in Figure 12; his final recommendations included watching the actors of the distraction and third plot closely, but also investigating the connection between the actors of the main plot.

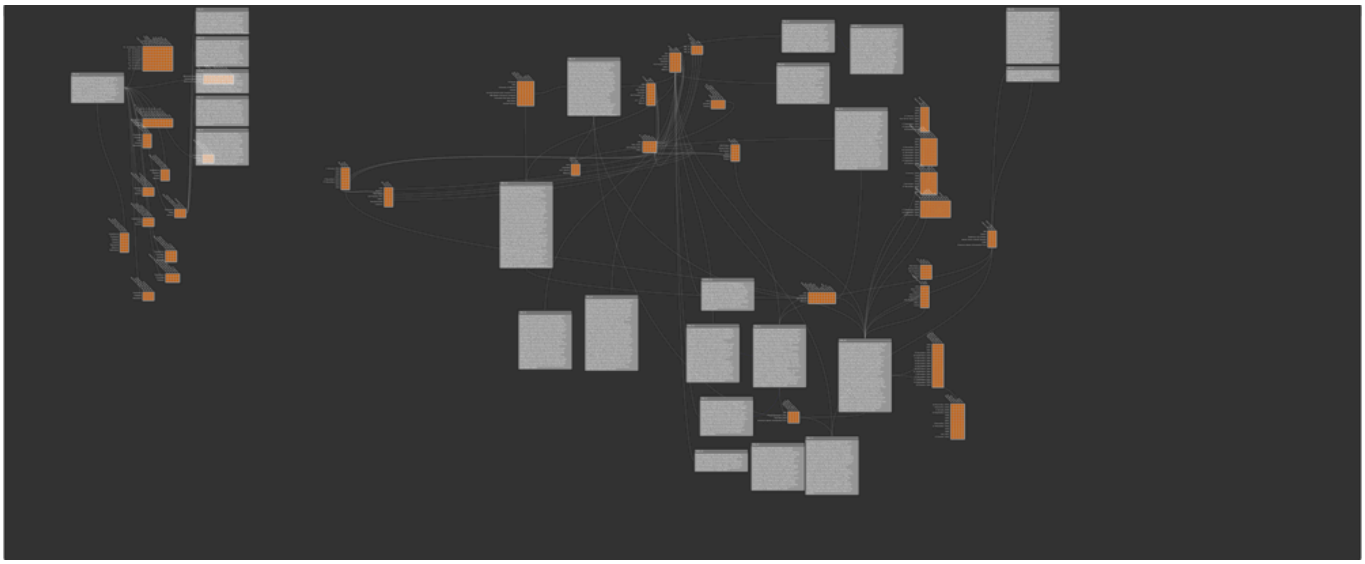

Figure 12: Graph workspace of subject 2 at the end of the study.

Subject 3 started broad and then focused in on leads one at a time. At first he looked up a known terrorist mentioned in FBI reports (see top left of Figure 13). He found that his strategy was too broad and loaded biclusters with Al-Qaeda in them and investigated terrorists with this affiliation (the biclusters on the bottom middle section of the graph); as soon as he found a possible threat he looked into the details. This led him to investigate the third plot until he hit a dead end; these are the documents to the right of the previous group of biclusters. He then back tracked to other Al-Qaeda leads and identified several people as key players of the main plot; see documents in the right third of the layout. From their specialization (doctor, chemist, etc.) he was able to assume that they might be planning a bio weapons attack. He spent the rest of his time trying to connect them and find out how they would attack, you can see several biclusters in the far right of the graph. His final recommendation included the third plot but also a specific list of leads to investigate regarding the main plot. Subject 3 was only missing a few documents in order to piece the whole answer together.

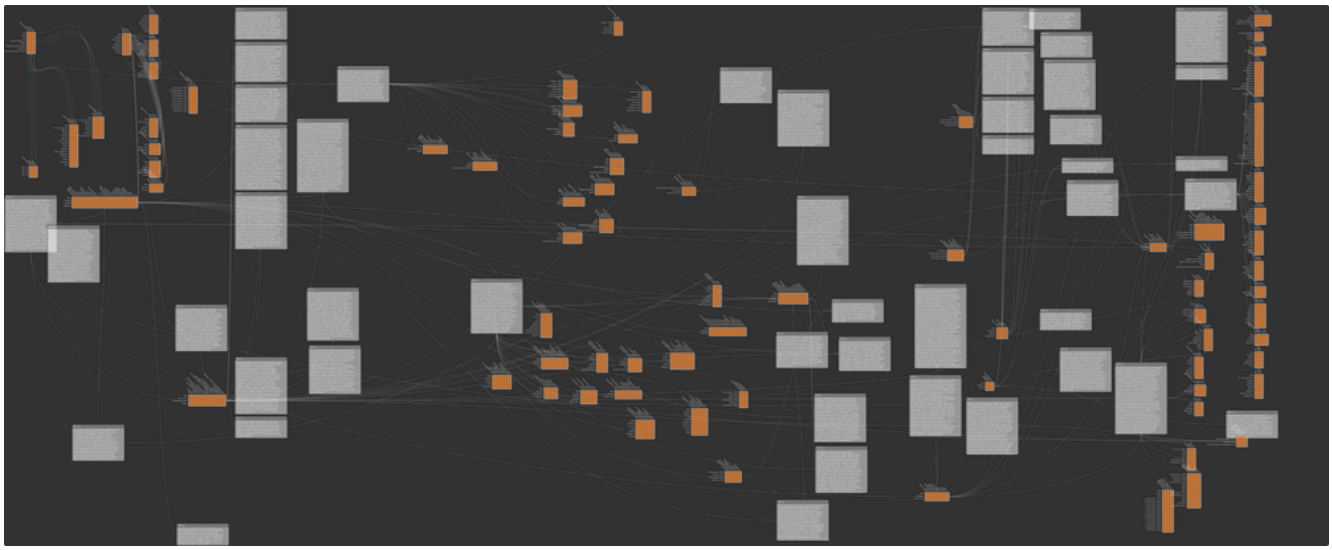

Figure 13: Graph workspace of subject 3 at the end of the study.

Subject 4 found the graph too small & slow; he instead preferred the data browser and used a combination of search and preview to investigate. At first he attempted to use the workspace, but he found the text and also grew frustrated by the amount of data the results of the 'show biclusters' and 'show documents' features returned. This can be seen in Figure 14, where he arranged some data on the left side but then gave up and never used the rest of the space. He found the data-browser to have better font size and to work faster from, and described his approach as brute force going from document to document. He did however still use biclusters in the search results to go from document to document through search terms. He found the distraction plot as well as the third plot. However when looking at documents on the main players he often dismissed them as minor or unimportant, they might be connected but did not seem to pose a threat. His recommendations included close watch of the individuals of the second and third plot as well as further investigation of one of the individuals of the main plot and his activities.

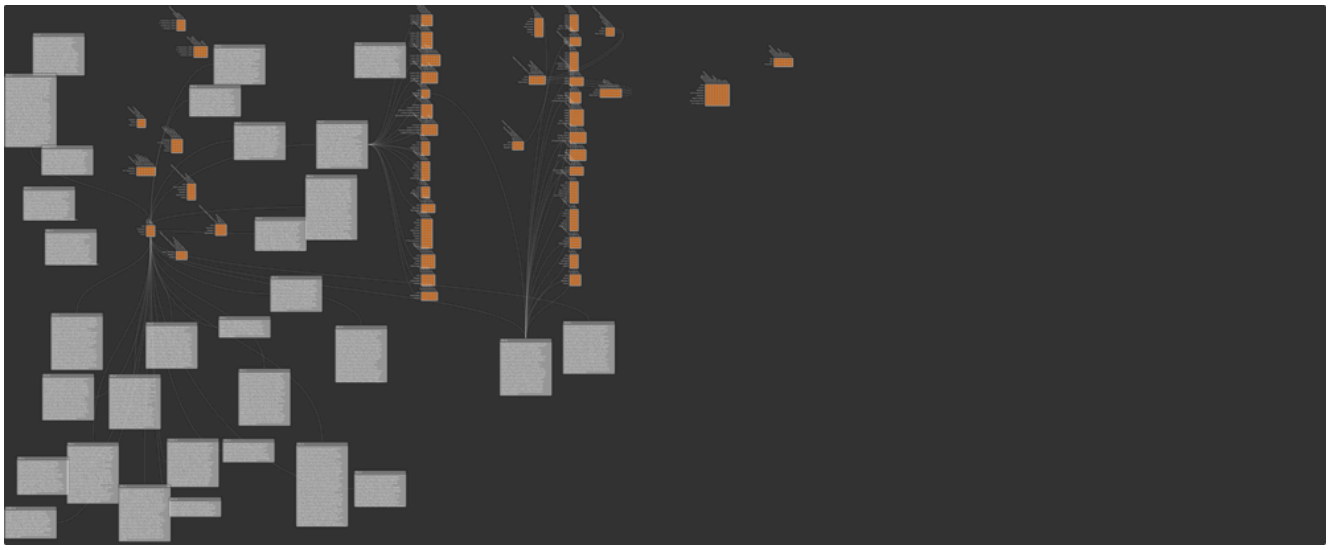

Figure 14: Graph workspace of subject 4 at the end of the study.

Subject 5 quickly developed a strategy to investigate the dataset using a combination of biclusters and documents. He started with people biclusters to obtain an idea of some people to look into and read the documents for these biclusters, then from the interesting ones he went back to biclusters. You can see this looking at the top left of his layout; the four initial documents are laid out vertically. Then he opened related biclusters scanned them, opened linked biclusters for a few of these directly to the right of these documents and quickly scanned them for interesting biclusters as well. He then repeated this process of looking at documents from biclusters with relationships of interest and then going back to biclusters. This can be seen on this workspace in Figure 15, with columns of biclusters followed by columns of documents. He described his workspace as a storyboard for his thoughts from start to finish. He tracked down the third plot to the arrest of it's main actors, considered the distraction plot as a possible plan and recommended investigating the students taking cruises further.

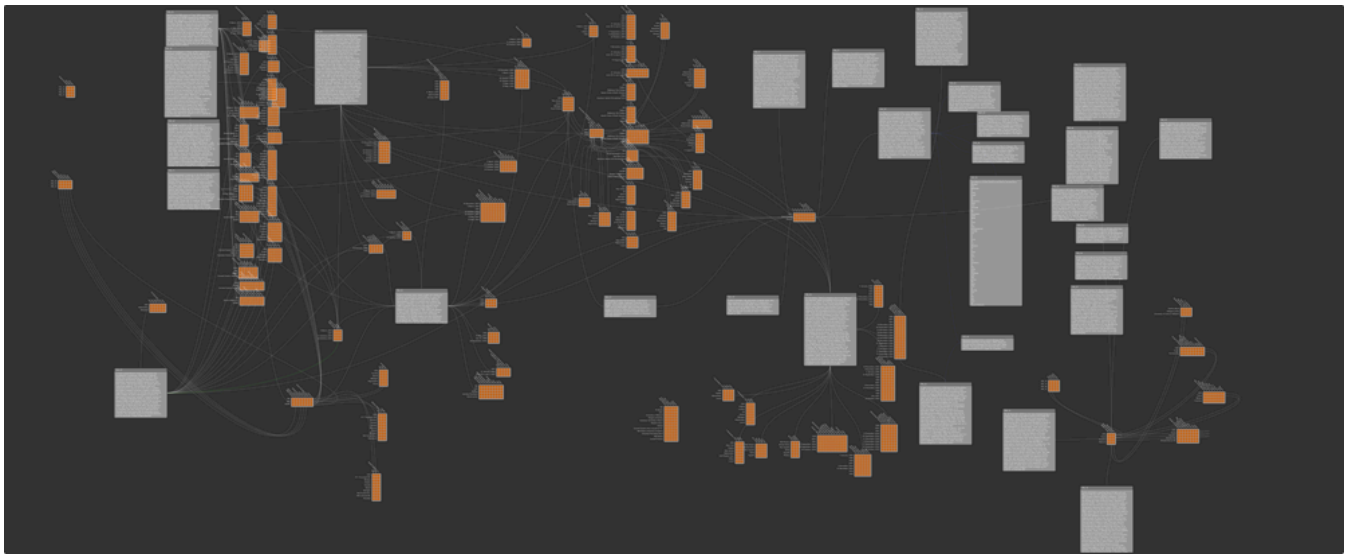

Figure 15: Graph workspace of subject 5 at the end of the study.

4.2 Feature Use

Through their investigation as each user became more familiar with the tool they each developed their own strategies. Here we look at the main features of MineVis and how the different subjects used them. The data browser will be considered a feature although we will explore in more detail how each tab was used.

4.2.1 Data Browser

The Search tab was the most heavily used tab in the data browser. Due to its versatility in returning search results as both biclusters and documents it made the subsequent tabs almost useless. The search was often used in conjunction with the frequency list to search for one of the entities. It was also used to get new leads when the user ran out of leads in the graph workspace, for instance they wanted to investigate a specific person related to a document that might not have showed up in nearby biclusters. We also noticed that of the results they previewed users would skim documents and only add them if pertinent to their train of though. The same is true of biclusters, however it is important to note that it was actually more often biclusters rather than documents that were added to the workspace from the search results. As one user said “I only manually added documents when I knew exactly what I wanted”.

Overall the search was a key feature of the system helping as both a starting point and for getting new leads.

The Frequency list tab was suggested to the users as a starting point. This was done to avoid them being overwhelmed by the size of a data set they did not know much about. Most of them used it as a starting point by looking through it until they found something that picked their interest.

The Browse Biclusters tab was used when the subject had a more specific strategy in mind. One example is subject 2's "follow the money" strategy; he simple browsed for biclusters involving money rather than use the frequency list as a starting point.

The Browse Documents tab was included mostly for completeness and due to the efficient search it did not see much use. It was used in some cases for example some biclusters had document title as a type and the document list provided a quick and easy look up.

The Browse Links tab was barely used other than in the training phase. However the in-graph version of the feature "Show Bicluster Links" was used more frequently (see the Graph Workspace section for more details).

4.2.2 Preview

Every subject used the preview at least once since it is the only way to get data into the workspace when it is empty. They also used it to add data from subsequent queries into the workspace. Users with spectacles also seemed to have a preference for reading documents in the preview rather than adding it to the graph to read it, most likely due to the larger font used in the preview.

Subject 4 was the only one who used primarily in the data browser and preview. He read all his documents in the preview and kept a network of names on in his notes on paper. His method came down to alternate search strategies: he would search for an entity usually a place or person, read the documents then search for another entity and this time preview biclusters to find a connected entity with which he would start this process over. He was aware of the "show documents" feature implemented in the graph

but disliked the graph too much to use it. He agreed however that if given the ability to use it in the data-browser by selecting show documents instead of add to workspace it would have been useful with his strategy.

4.2.3 Graph Workspace

With the exception of subject 4 all the subjects did the bulk of their investigations in the graph space using the various context actions provided.

All of the participants used a relatively small number of elements in the graph compared to the data available to them. Table 1 shows the numbers of biclusters, documents, user links and highlights in the final state of the workspace for each participant. Out of 1001 biclusters none used more than 100, and perhaps more meaningfully, out of 111 documents none of the users inspected more than 50 documents. This suggests MineVis might be effectively helping users to filter down the documents and read only documents relevant to their investigation. We can also see that user links and highlights were barely used at all.

	Biclusters used	Documents used	User links used	Highlights used
Subject 1	36	29	0	1
Subject 2	34	24	3	0
Subject 3	77	47	0	5
Subject 4	45	31	1	0
Subject 5	87	29	6	1

Table 1: Element statistics for users final graph workspace.

Each of the participants had different strategies to navigate within the workspace. Table 2 shows the types of links in use in each workspace at the end. ‘Bic/Bic’ stands for bicluster-to-bicluster links obtained by using the ‘Add link to workspace’ feature or the ‘Show Bicluster Links’ feature. ‘Doc/Doc’ stands for document-to-document links that could only be created manually by the user. In fact they

links account for almost all the user created links in Table 1. 'Bic/Doc' and 'Doc/Bic' represent the bicluster-to-document links and document-to-biclusters links respectively, the first is created using 'Show Bicluster Documents' whereas the second is created using 'Show Document Biclusters'. We can see that in overall the participants preferred to see the documents linked to biclusters (show bicluster documents) rather than the other way around with the exception of subject 5 who even said he found them "equally useful". Also keep in mind that subject 4 stopped using the workspace early in his investigation.

	Total Links in graph	Bic/Bic links	Doc/Doc links	Bic/Doc links	Doc/Bic links
Subject 1	84	11	0	52	21
Subject 2	71	9	3	34	25
Subject 3	169	12	0	101	56
Subject 4	61	4	1	26	30
Subject 5	130	21	6	31	72

Table 2: Detailed statistics on the types of links used.

Users used the close feature in two circumstances: the "read a document" or "bicluster" and determined it was useful to their investigation. Or after using 'show documents' or 'show biclusters' they were overwhelmed with the amount of information and decided to explore another option but due to the lack of an undo feature they had to close all the newly opened biclusters or documents. In one case before the number of biclusters returned by 'show document biclusters' one of the subject spent 30 minutes closing all of the biclusters that opened as he did not want them to clutter his workspace.

Subject 5 made use of 'user links' than other users. He used biclusters and documents heavily throughout his investigation and towards the end he explained that there was a series of names he needed to investigate individually. He manually opened the documents to reach about each of the

individuals in his notes and then once done synthetizing the information he proceeded to manually link all the recently added documents together to represent he connections he learned about.

5 DISCUSSION & FUTURE WORK

In this section we explore some of the various behaviours observed during the study. The results show that all of the users integrated the biclusters in their strategies and used them to either reduce the amount of documents to read or to target which documents to read. We also gathered a lot of insight on how users conducted their investigation and how biclusters were used during the interview. The discussion will be structured in four sub sections first a general discussion of the results relating to visual analytics, then more specifically data mining concepts and information visualization concepts and last a short discussion of framework possibilities.

5.1 Visual Analytics

Biclusters were used by all of the users, even the one that did not use the workspace; this shows that they users were able to integrate them in their mental process to some extent even with limited training. Most of the users also believed their progress would have been slower without having documents. The answers they gave during the interviews also lead us to believe that biclusters almost function as a small hypothesis. The hypothesis consists of the relationship the bicluster is made off, when this spiked the interest of the users they usually investigated by using 'show documents'. They already knew the presence of the relationship but this allowed them to understand the nature of it and confirm or deny they bicluster hypothesis, one user even said that 'the documents are the evidence' when describing the differences between documents and biclusters. The way subjects used biclusters to connect documents and navigate the dataset confirms our first hypothesis, viz. that biclusters can be integrated in the analytic process and used in combination with source data. The bicluster as hypothesis and documents as evidence mental model confirms our second hypothesis, viz. that biclusters provide different insight from documents but do not replace them.

Investigating the hypothesis by looking into the related documents was part of the strategy of many users. This can explain why as shown in Table 2 there are more links from biclusters to documents than documents to biclusters. However we believe it can be due to the fact that the result of show documents was usually on par with their expectations of what it would be. The opposite action, show biclusters for a document, did not usually meet their expectations and, worse, overwhelmed them with a long list of biclusters (even after the results were limited to 15 entries). One possible cause for this is since there are more biclusters than documents, the number of results will naturally increase and due to the lack of a technique to rank biclusters the results are returned in a random fashion and arbitrarily limited to 15. This lack of consistency likely keeps the users from being able to build expectations as to the results and can be the reason for the poor adoption of this feature.

We also found that training plays an important role in the users efficiency at using the system. Users can be trained in different aspects separately and some of our users had different backgrounds. First a user can be trained as an analyst in investigation techniques, second a user can be trained in using visual analytics software and third a user can be trained to use and understand biclusters. We believe these are the main three types of training that apply to using MineVis and will explore each in more detail.

None of our users are analysts. While some have taken courses that covered visual analytics they still cannot be considered to be analysts. It is important to make that distinction when considering their performance or their investigation strategies.

Two of our users had previews experience with visual analytics tool. Since the graph workspace in MineVis is very similar to other analytics tool with a layout component it is possible that these users were able to reuse some part of previously developed strategies. When asked, one subject said that in the graph he reused his previous strategies and it enabled him to organize it better. Both the users who had previously used visual analytics tools seemed to be able to recall specific items in their layout better and quickly give an explanation of what they did.

None of our users were previously familiar with biclusters. Some of them were familiar with data mining but not biclusters. While the concepts were explained to them in training it might not have been sufficient to allow them to master the concept of biclusters. Three of the users believed that if given more time they would be able to make better use of the biclusters in their investigations. Our findings confirm that as we expected although users had difficulty understanding biclusters, the combination of limited training and hands on use still allowed them to find ways to make use of biclusters in their strategies and solutions.

Another pertinent topic to intelligence analysis is finding a starting point when investigating a new dataset. This "where to start" problem becomes more important as the size of datasets increases. In MineVis all the subjects were able to get started with relative ease; here we attempt to discuss the features that were helpful and how they helped attenuate the "where to start" problem. We recommended the frequency list from the data-browser as a starting point to all of the subjects, however as mentioned in the results section not all of them used it. We believe the key factor is that the MineVis system allowed the subjects to start with more concrete information on the dataset, entities and relationships as opposed to more abstract information like a list of documents with abstract titles. This allows users to focus more on their strategy for solving the problem at hand rather than their strategy for exploring the data, for example one subject decided to investigate the people in FBI reports, and then revised his strategy to investigating known Al-Qaeda associates. Another subject as mentioned before focused directly on following the money and then used the systems features to accomplish this task. Also note that MineVis is different from other visual analytics system in that it not only offers entities and relationships from entities to entities, but also groups the relationships into biclusters. Biclusters played a key role in allowing users to focus on their strategy to solve the problem and sensemaking rather than foraging. Users relied on biclusters to explore groups of relationships and then decided which of these should be further investigated. Without biclusters the users would have had to read the

documents in order to build these connections, and hence biclusters allow them to focus the strategy on the solution desired rather than the type of dataset. In MineVis the use of biclusters as connectors allows a user strategize about "how to start" rather than "where to start", shifting the starting point problem away from foraging and closer to sensemaking.

In this study of MineVis we focused on evaluating document-based intelligence datasets but other databases can be used with MineVis. The project configuration of MineVis was designed to be flexible and allows for configuration on many different types of data. With a few changes to the visualization workspace MineVis can be used on non document-based relational datasets. For example the IMDB database could be mined and used to explore connections and patterns between actors, producers, genres, or ratings in MineVis. Databases can be imported by either conversion to the MineVis project database format or upgrading the MineVis project configuration to be even more flexible. This provides opportunities for evaluating many new types of datasets.

5.2 Compositional Data Mining

The current implementation of MineVis has limitations on the data generated from the source dataset as well as the ways in which can be used. In this subsection, we discuss our findings and explore future possibilities regarding the data mining aspects of MineVis. First we will discuss issues related to biclusters (how they were used, alternative types of biclusters that could be used, as well as additional data that can be generated to measure biclusters). Second we will discuss the concepts of chaining multiple biclusters in MineVis, again from both usage and perspectives of future possibilities.

Biclusters were aimed at information reduction, but with almost ten times more biclusters than documents the risk of information explosion is real. Yet from the numbers in Table 1 we see that users where not overwhelmed by the sheer numbers. We can say that biclusters generate more data but it would be more accurate to say that biclusters generate more meta-data, as they do not add new entities they only group entities together. So to the user the many different biclusters offer many different ways

to look at the data but not more data. The users did not spend much time skimming lists of biclusters, instead they used them in the context of their current train of through. The result is that in most situations the number of biclusters the user looks at has already been filtered down using context. Biclusters clearly outnumber documents but when used efficiently in the system that number is hidden from the end user so that they see the benefits instead of being overwhelmed.

Information reduction is seen when biclusters are used in combination with documents, a key feature for most subjects was the ability to go from bicluster to documents. Once they knew what they wanted to find evidence, this feature allowed them to look at a reduced set of documents most of which would be relevant. It is important then to discuss which biclusters were most useful to the user and the meaning that they had. We identified three key attributes to biclusters: the size of a bicluster, the different kinds of biclusters (domains on each sides) and the content of biclusters (generic or specific) and we will discuss the meaning of each to the subjects. First it is important to look at how documents are related to biclusters in more details. For example, as shown in Figure 16 below, we have a set of four documents: for each we re-described the content as a list of relations from people to places and the bicluster generated from those documents. Each cell in the bicluster is color coded to match the document it originally came from. This not only shows how biclusters are related to documents but how the meaning of the relationship between biclusters and documents to the user will depend on the context. A user interested in the death star investigating this bicluster might find that 3 out of the 4 documents are irrelevant; a user interested in Tatooine only on the other hand will find that only half of them are irrelevant; and a user interested in Luke only will find that 3 out 4 contain information relevant to him. This shows how a bicluster can have a different meaning and usefulness depending on the intent of the analyst. Also note that sometimes information not directly relevant can prove useful as well, the last user above might find that in fact the document not mentioning Luke actually talk about him anonymously so the bicluster in this case would help him pick up on this less obvious fact. Last note that unlike in the

example, documents are not limited to representing columns but can represent parts or rows or even non-contiguous areas of a bicluster.

Figure 16 Documents and list of relations in each on the left and bicluster generated on right. Color coded by original document. Fictional data used.

We can now relate the concepts above to the size of the biclusters in the dataset. The example above had 3 columns and 3 rows; in our study each bicluster had a minimum of 3 rows and 3 columns but some were much larger (3 rows by 8 columns, 3 columns by 7 rows or 5 rows by 5 columns). When asked on their preferred bicluster size, three of the subjects had no preference, and the other two however noted that they preferred the smaller biclusters but made sure to note that they still found the larger one useful. One explained that it was because when using 'show documents' on a smaller bicluster the set of documents loaded would include more relevant documents, while larger biclusters returned more documents and sometimes several of them were irrelevant to the subject's current context of investigation. We believe it is not the size of the bicluster itself but the relevance of the entities in it that dictate how meaningful it is to the user.

The entities in biclusters can be considered generic or specific, the more generic entities in one the more generic the bicluster. Users skimmed biclusters to look for interesting relationships: if something piqued their interest they would investigate it further; this could happen even for a single specific relationship even if the rest of the relationships were more generic. Several of the users inquired about a way to hide rows or columns from a biclusters to hide generic information. Information considered generic were high occurrence entities such as countries or states, for example US or Virginia appeared in so many biclusters it was not very meaningful to the users; other information such as an individual's name or a specific place like a shop is considered more specific. The more specific entities a bicluster contained, the more meaningful the user found it. In one case a subject mentioned that he was looking for specific term in a certain area of his layout and others in a different area of the layout, suggesting that entities can be generic on a global level as well as on a local level. Trimming the dataset from some of the generic entities could yield more specific biclusters globally but might not improve at a more local level.

The last characteristic of biclusters we need to discuss is the type of relationships and the domains on its sides. All subjects expressed that they found biclusters between people and location most useful. Money was useful to two of the users who at least at some point attempted to "follow the money" as a strategy. Of all the entity domains it appeared that dates was the least useful. Some subjects found that it was simple not useful; others noted that skimming such biclusters was helpful but they generally did not feel like digging deeper into it as opposed to the other more domains they found more useful.

These results suggest that the various characteristics of biclusters or their values cannot be sorted into useful and non-useful at a glance. Context played a heavy role for users to decide how meaningful different pieces of information were. Also the users often depended on documents since they can only see relationships and not context in biclusters; but after reading documents they went back to the bicluster knowing what was meaningful and what was not. However when a user did know an entity in a

row or column of a bicluster there was no feature available for them to act on it; in some cases the other entities seemed interesting enough for the user to investigate; in other cases they would give up on investigating that bicluster further if they felt it had too many generic entities for the reasons mentioned above. This suggests that a dynamic filtering on a biclusters could alleviate these problems, for example in Figure 16 a user not interested in 'Death Star' might be able to filter or hide that row from the bicluster.

MineVis currently only uses complete biclusters, biclusters where all of the cells of the matrix are required, for example see Figure 17. Using an algorithm that can generate partial biclusters can present benefits through non-constant rows and columns [31, 32]. Integrating such an algorithm into the MineVis pipeline might offer benefits at the analysis stage. Currently if three people are related to four locations, but two of the people are also related to a fifth location, the fact that the third person is not related to that location will essentially truncate that information from the bicluster and possibly hide something that could be relevant to one's investigation. The potential of such biclusters should be investigated.

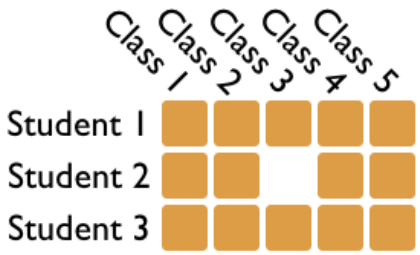

Figure 17 Example of a partial bicluster.

Another similar possible extension of the mining done in MineVis is to use non-binary relationships. Again coherent values might help just as in the non-binary case above [31, 32]. Currently the mining algorithm only preforms mining on binary relationships. For example to the current algorithm two entities are related to each other in 7 documents is just the same as two other entities that are only related

in 1 document. Using a threshold we can then determine what is valuable enough to be considered a bicluster: for example more than 5 only on a numeric scale (see Figure 18). This data could be used as the intensity of the relationship and provide better options for linking, filtering or even ranking biclusters and potentially again improve the investigation for the user.

Figure 18 Example of a bicluster highlighted in a non-binary matrix.

Another possible feature that was not implemented in the system is a bicluster ranking feature. Biclusters are used in search results, browsers and listed from either documents or other biclusters through links, however they currently a way to sensibly order them. Various measures could be automatically generated for biclusters such as interestingness or statistical significance and then used to rank biclusters. Statistical significance could prove useful when loading a list of biclusters linked to a document. Currently the first 15 query results of that feature are returned in an arbitrary order and subjects mentioned that they sometimes found the results either too similar or not relevant, statistical significance can alleviate this. Generating an interestingness value could improve the data-browser, rather than just listing the biclusters by their ids the interestingness would be displayed and could help users in selecting results to preview. The ability to rank bicluster offers the possibility to improve many of the features in MineVis and in turns further assist the analyst in their tasks.

Another important features of biclusters was chaining, i.e., the act of putting biclusters with matching entities side by side. Study subjects did not use the chaining features as much as the bicluster and document features; our expectations were that the users would be able to build long chains of multiple biclusters as described in Figure 3 or even longer. These chains would act as stories from the start to the end of their investigations, with documents branching of to the sides as evidence. In Table 2 we can see the number of actual links from bicluster to bicluster resulting from using the show bicluster link function, subjects 3 and 5 used it the most. However they did not use it as we had hypothesized. Their strategies consisted of looking at biclusters for interesting relationships and to then investigate the evidence by reading related documents. Then when a document warranted further investigation they opened related biclusters. It was then that they would use the chaining link function; if the biclusters related to the document did not show them information they were interested in they opened more biclusters and looked again. The bicluster as hypothesis and documents as evidence mental model most subjects used shows us that both are best in combinations. We can see on the graph layouts of subject 3 in Figure 13 and subject 5 in Figure 15 almost every single element has at least one connection creating a large network. A chain containing only biclusters would offer little evidence for the hypothesis it represents. This suggests our initial concept, using compositional data mining to build chains of biclusters only, did not make for useful features for the user. Instead our findings indicate that to build a good chain documents need to be included as well as bicluster in order to include not only the hypothesis, gained by skimming biclusters, but also the evidence found in the documents.

Currently the chaining data in MineVis is only used to show bicluster links manually. However queries could be used to build links between two items (entities, documents or biclusters). Given a start point and an end point links could be created by linking several biclusters to each other. Note that unlike our initial assumptions, as mentioned in the previous paragraph, chains should be made of both biclusters and documents, used alternatively if possible. This would enable a user to explore direct as well as less

obvious indirect connections between two items and the ability to automatically build long chains compared to the current manual option using the graph layout. It presents several technical issues however such as ranking the resulting chains. The similarity between two biclusters can be used to choose a tighter match, but whether it is sufficient to rank the results of search results containing many elements of different types has yet to be determined. Unlike a map where a user can easily chose between a preference for the shortest route or a route avoiding tolls, the usefulness of a route from an item to another through entities or document amongst others must be ranked by new preferences and measures such as the ones discussed above. We can use what was learned about the way biclusters and documents are used in combination to design ways to build chains automatically or interactively to support the user with what they need.

5.3 Information Visualization

This section focuses on information visualization specific concepts, however many of them were already addressed along with other concepts in the previous two subsections. Two important concepts not yet discussed are the integration of semantic interaction and the concept of a clean or neat workspace and what that can mean. First we will discuss how semantic interaction can help to avoid overwhelming users and second we will look at the idea of a neat workspace.

Semantic interaction [11] could be used to address some of the issues raised in section 5.1. For example if we added a feature to allow the user to highlight text and entities within a document, this information could be used by an algorithm on the server side when loading the biclusters for that document to filter out the ones not relevant to what the user through their highlight has indicated as relevant. This can be done without the need for complicated input parameters that would increase the mental load on the user. This technique could also be applied to the ‘show document’ action of the biclusters to further improve it by letting the user mark the rows or columns of a bicluster as important or not important. Many users actually expressed their wish for a way to ‘hide’ rows in a biclusters

during both the investigation and the interview. They saw it as a way to hide irrelevant entities but their willingness to perform such actions shows potential for semantic interaction. Links also present potential for semantic interaction. Most users did not use highlight but one user mentioned that he would have rather used arrows to show the direction of the connection. This suggests that the direction could be used as soft data. Perhaps using features such as show document and show link should indicate that direction automatically. Including these types of dynamic user interaction based fillers could allow the users to deal with even larger dataset by keeping the mental load low.

Supporting an analyst's desire for a neat/clean workspace offers an alternative way to look at information overload issues. The users really wanted an undo or function or a delete multiple item selection feature in the system especially when an action opened many new biclusters or documents. They expressed a concern for keeping their workspace neat and organized. We believe there might be more than just a usability need as they all asked for this in similar situation where they became overwhelmed with the data. This suggests that perhaps there is a two types of workspaces a long term workspace where data is neatly organized and approved to stay and a more short term workspace where new ideas and data is added, filtered and evaluated for admission into the long term workspace similar to the way long and short term memory works. It seemed to us as if the information could be displayed in multiple layers as shown on Figure 19. The bottom layer would represent long-term information, reviewed and accepted, while recent actions such as listing biclusters for a document would be on a separate layer, short-term layout, on top of the long-term one. Sharing the same view port the two layers would appear as one when viewed on the screen, however the separation would allow for discarding and hiding the short-term layout to pursue a different avenue while maintaining the long-term layout or to push specific elements to the long-term layout.

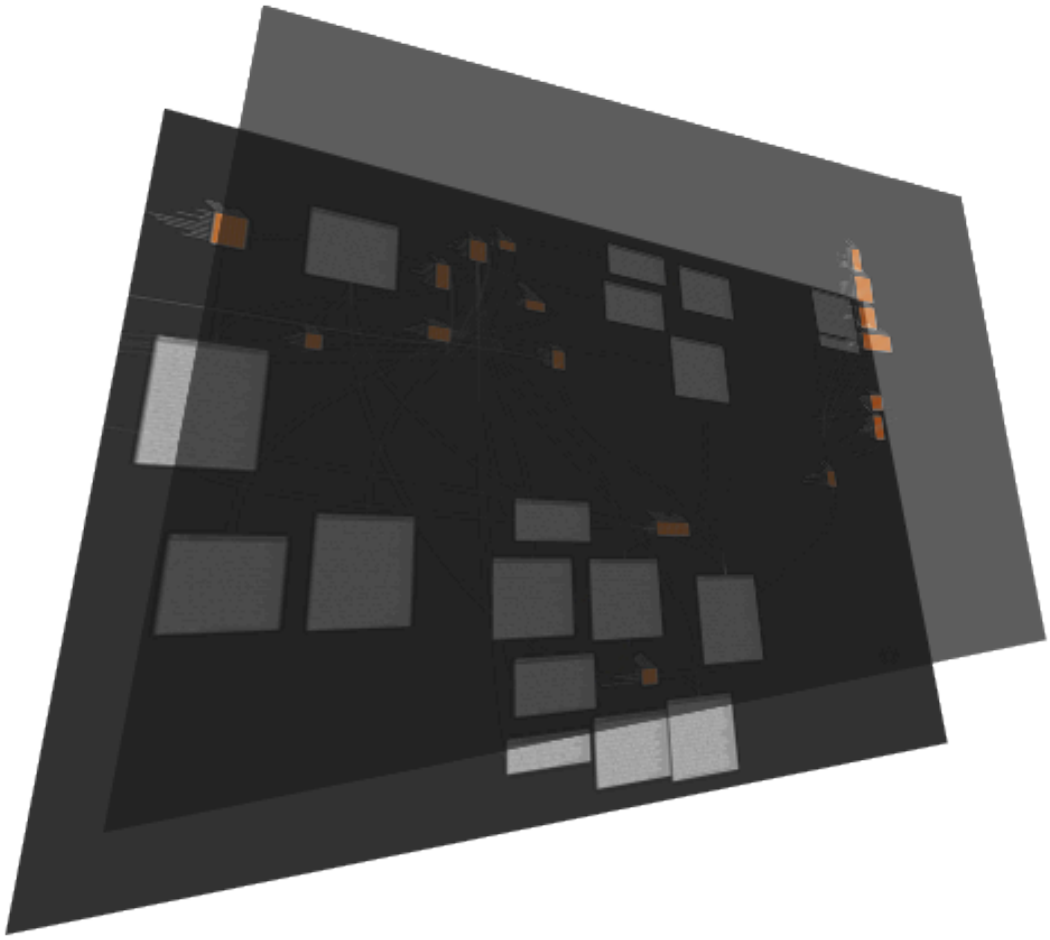

Figure 19 MineVis on 2 layers displayed at an angle for better paper representation. The Long-term info is on the bottom layout and the recently added list of biclusters is on a top 'temporary' or short-term layout.

5.4 MineVis as a Framework

The MineVis system provides a tight integration of data mining algorithm and visualisations through a common database storage type. The system was designed to allow multiple algorithms or visualizations at each stage of the pipeline. This however is limited to the current pipeline, refactoring MineVis as a framework rather than a system would enable for a wider range of data types, data mining algorithms as well as visualizations and analytics tools. There are many toolkits and frameworks available such as visualization toolkits like D3 or the InfoVis Toolkit [33, 34] and focus on easing development of visualizations. Or data mining frameworks like WEKA from the University of Waikato [25] focus on providing tools for data processing and mining. A MineVis framework would not focus on either

specifically instead it would focus on improving the integration of both. The MineVis system showed the viability of the integration of biclusters into visual analytics for intelligence analysis datasets. MineVis as a framework could enable many combinations datasets, data mining and information visualization technique, including integration of currently existing frameworks to be evaluated.

6 CONCLUSION

We have designed a system, MineVis, that combines the power of data mining and visual analytics in a seamless environment. It uses biclusters to improve the visual analytics workspace by allowing use of a new rich metaphor combining relationships.

Through a user study we found that while training was required to integrate biclusters into the user's mental model, the results showed great potential. All of our users were able to integrate biclusters in their strategies. This allowed them to explore relationships and links across the workspace, affording better non-linear investigation. Users were able to explore relationships between entities across separate documents and form hypothesis and then read the related documents to confirm or deny it. They were able to effectively focus on the connections and read documents in context as opposed to linearly or randomly.

MineVis successfully integrates compositional data mining and visual analytics. As the study results show, biclusters can be integrated into one's strategies and provide new interactions with text based datasets. We have also raised a lot of questions about the possibilities that this created. As the amount of data we deal with every day keeps increasing, systems like MineVis allowing higher level views of the data will become more and more essential for analysis tasks.

BIBLIOGRAPHY

[1] Thomas, J. J. and Cook, K. A. 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics. *IEEE Computer Society*.

[2] Pirolli, P. and Card, S. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *International Conference on Intelligence Analysis* (2005).

[3] Andrews, C. and Endert, A. et al. 2010. Space to think: large high-resolution displays for sensemaking. *Proceedings of the 28th international conference on Human factors in computing systems* (Atlanta, Georgia, USA, 2010). ACM, New York, NY, USA,

[4] Shipman, F. M. and Marshall, C. C. 1999. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. In *Computer Supported Cooperative Work (CSCW)*, 8, 10.1023/A:1008716330212 (1999). Springer Netherlands, 333-352.

[5] Endert, A. and Fox, S. et al. [insert 2012, Under Review of Publication]. The Semantics of Clustering: Analysis of User-Generated Spatializations of Text Documents. *In Proceedings of the AVI (2012, Under Review)* ([insert 2012, Under Review of Publication]). [insert City of Publication],

[6] Pike, W. A. and Bruce, J. et al. 2008. The Scalable Reasoning System: Lightweight visualization for distributed analytics. *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on* (oct. 2008). 131 -138.

[7] Wei, F. and Liu, S. et al. 2010. TIARA: a visual exploratory text analytic system. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (Washington, DC, USA, 2010). ACM, New York, NY, USA,

[8] Stasko, J. and Gorg, C. et al. 2008. Jigsaw: supporting investigative analysis through interactive visualization. In *Information Visualization*, 7 (April 2008). Palgrave Macmillan,

[9] Andrews, C. P. 2011. Space to Think: Sensemaking and Large, High-Resolution Displays. Virginia Polytechnic Institute and State University.

[10] Endert, A. and Fiaux, P. et al. 2011. Unifying the Sensemaking Loop with Semantic Interaction. In *IEEE Workshop on Interactive Visual Text Analytics for Decision Making at IEEE VisWeek* (2011).

[11] Endert, A. and Fiaux, P. et al. 2012. Semantic Interaction for Visual Text Analytics. *CHI '12* (2012).

[12] Henry, N. and Fekete, J.-D. et al. 2007. NodeTrix: a Hybrid Visualization of Social Networks. In *Visualization and Computer Graphics, IEEE Transactions on*, 13, 6 (nov.-dec. 2007). 1302 - 1309.

[13] Bach, B. and Pietriga, E. et al. 2011. OntoTrix: a hybrid visualization for populated ontologies. *Proceedings of the 20th international conference companion on World wide web* (Hyderabad, India, 2011). ACM, New York, NY, USA,

[14] Pati, A. and Jin, Y. et al. 2008. CMGSDB: integrating heterogeneous Caenorhabditis elegans data sources using compositional data mining. In *Nucleic Acids Res*, 36, Database issue (Jan 2008). D69-76.

[15] Ramakrishnan, N. 2009. Exploring Multi-Stress Scenarios using Compositional Data Mining. *Virginia Tech*.

[16] Jin, Y. and Murali, T. M. et al. 2008. Compositional mining of multirelational biological datasets. In *ACM Trans. Knowl. Discov. Data*, 2 (April 2008). ACM, New York, NY, USA,

[17] Zaki, M. J. and Hsiao, C.-J. 2002. CHARM: An efficient Algorithm for Closed Itemset Mining. *SIAM International Conference on Data Mining* (2002). 457-473.

[18] Zaki, M. J. 2004. Mining Non-Redundant Association Rules. In *Data Mining and Knowledge Discovery: An International Journal*, 9, 3 (Nov 2004). 223-248.

[19] Zaki, M. J. and Hsiao, C.-J. 2005. Efficient Algorithms for Mining Closed Itemsets and their Lattice Structure. In *IEEE Transactions on Knowledge and Data Engineering*, 17, 4 (Apr 2005). 462-478.

[20] Uno, T. and Asai, T. et al. 2003. LCM: An efficient algorithm for enumerating frequent closed item sets. *FIMI'03* (2003).

[21] Uno, T. and Kiyomi, M. et al. 2004. LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. *ICDM'04 Workshop FIMI'04* (2004).

[22] Uno, T. and Kiyomi, M. et al. 2005. LCM ver.3: collaboration of array, bitmap and prefix tree for frequent itemset mining. *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations* (Chicago, Illinois, 2005). ACM, New York, NY, USA,

[23] Uno, T. and Arimura, H. LCM. http://research.nii.ac.jp/~uno/codes.htm. Accessed Dec 24 2011.

[24] Beygelzimer, A. and Kakade, S. et al. 2006. Cover trees for nearest neighbor. *Proceedings of the 23rd international conference on Machine learning* (Pittsburgh, Pennsylvania, 2006). ACM, New York, NY, USA,

[25] Frank, E., Holmes, G., Mayo, M., Pfahringer, B., Smith, T., and Witten, I. WEKA. http://www.cs.waikato.ac.nz/ml/weka/. Accessed Dec 24 2011.

[26] Raden, K. Distance Measures. http://www.sequentix.de/gelquest/help/distance_measures.htm. Accessed Dec 24 2011.

[27] Potencier, F. Symfony Framework. http://www.symfony-project.org/. Accessed Dec 24 2011.

[28] JQueryProject JQuery. http://jquery.com/. Accessed Dec 24 2011.

[29] Baranovskiy, D. Raphaël JavaScript Library. http://raphaeljs.com/. Accessed Jan 17 2012.

[30] Hughes, F. J. 2005. 'All Fall Down' (Atlantic Storm) Case Study. *Joint Military Intelligence College*.

[31] Madeira, S. C. and Oliveira, A. L. 2004. Biclustering algorithms for biological data analysis: a survey. In *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1, 1 (jan.-march 2004). 24 -45.

[32] Tanay, A. and Sharan, R. et al. 2004. Biclustering algorithms: A survey. *In Handbook of Computational Molecular Biology Edited by: Aluru S. Chapman & Hall/CRC Computer and Information Science Series* (May 2004).

[33] Bostock, M. d3.js Data-Driven Documents. http://mbostock.github.com/d3/. Accessed Jan 18 2012.

[34] Fekete, J.-D. The InfoVis Toolkit. http://ivtk.sourceforge.net/. Accessed Jan 18 2012.