Bixplorer: Visual Analytics with Biclusters

Patrick Fiaux, Maoyuan Sun, Lauren Bradel, Chris North, and Naren Ramakrishnan
Virginia Tech

Alex Endert
Pacific Northwest National Laboratory

A prototype visual analytics tool uses data mining algorithms to find patterns in textual datasets and then supports exploration of these patterns in the form of biclusters on a high-resolution display.

Exploring relationships among people, places, dates, organizations, and other entities in large textual datasets is an important aspect of intelligence analysis. However, imprecisely formulated goals make meaningful knowledge discovery in unstructured text difficult—for example, to find suspicious activity, identify nontrivial coalitions of individuals and groups, and determine links between diverse entities created through chains of intermediaries.

Visual analytics tools facilitate this process by creating spatial abstractions of relationships—clusters, timelines, and so on—that researchers can examine and, by applying their domain expertise and intuition, obtain new insights. Toward that end, Bixplorer is a prototype visual analytics tool that uses data mining algorithms to find patterns in textual datasets and then supports exploration of these patterns in the form of biclusters on a high-resolution display.

BIXPLORER

Bixplorer first applies standard entity extraction algorithms on a dataset and posits relations between different entity classes via lexical or syntactic co-occurrence. It effectively models this extracted data as a relational database.

The tool then extracts *biclusters* from the discovered relations. A bicluster is a bipartite graph that "bundles" relations between individual entities into a pair of sets, wherein every vertex of one set is connected to all vertices of another set. For example, Figure 1 shows a bicluster of a set of students {S2,S5} attending the same set of courses {C3,C7,C8}. Biclusters are typically maximal: additional entities—in this case, students and classes—can't be added because they will not have a relation to one another in the original matrix.

Bixplorer can use numerous algorithms to discover biclusters, including CHARM (an implementation of closed association rule mining), and LCM (linear closed itemset miner). These algorithms typically find biclusters having one row, and then aim to grow them by adding more rows and observing how many columns, if any, are affected. The biclusters are organized along a lattice; the algorithms differ in how they traverse this lattice and the search optimizations they employ.

Bixplorer then chains the biclusters by matching sets of entities across common domains of interest, as Figure 2 shows. It defines a cover tree for each domain and then indexes the set of rows and set of columns for every bicluster into the

Published by the IEEE Computer Society

corresponding cover tree. Once all biclusters are indexed, it conducts similarity searches to reveal the nearest neighbors.

USER INTERFACE

As Figure 3a shows, the Bixplorer user interface consists of three main views. The *data browser* provides access to the various data structures generated by the system, as well as the raw source data; users can search document plain text, rank entities by frequency, and browse documents, biclusters, and links between biclusters. The *preview pane* lets users preview search results as well as add selected results to the *workspace*, where the bulk of visual exploration of the dataset occurs. Initially, the workspace is empty. Throughout the course of their analysis, users add documents and biclusters to the workspace to visualize and link data as Figure 3b shows.

Users can perform two types of actions on biclusters in the workspace. "Show documents" displays all the documents containing relationships within a bicluster, while "show biclusters" ranks the biclusters that contain one or more relationships within a given document. Bixplorer also provides many highlighting and annotation capabilities.

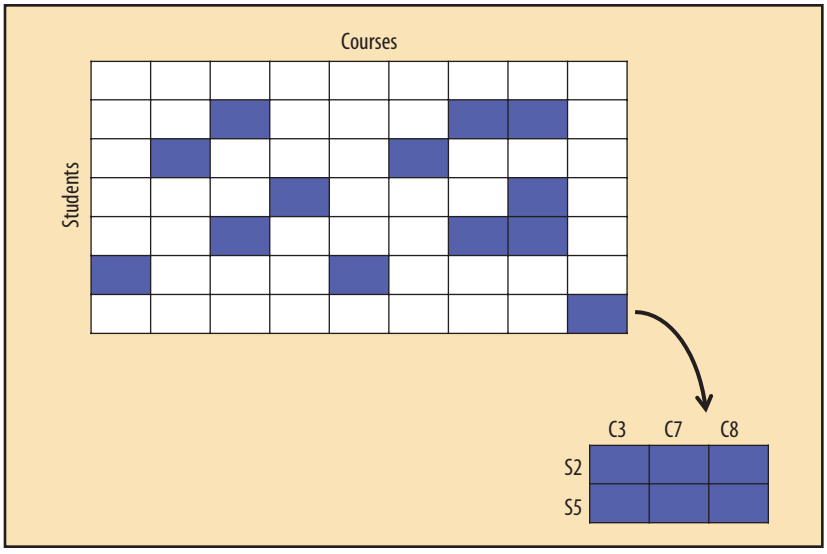

Figure 1. Example bicluster extracted from a matrix of relations between students and the classes they're attending.

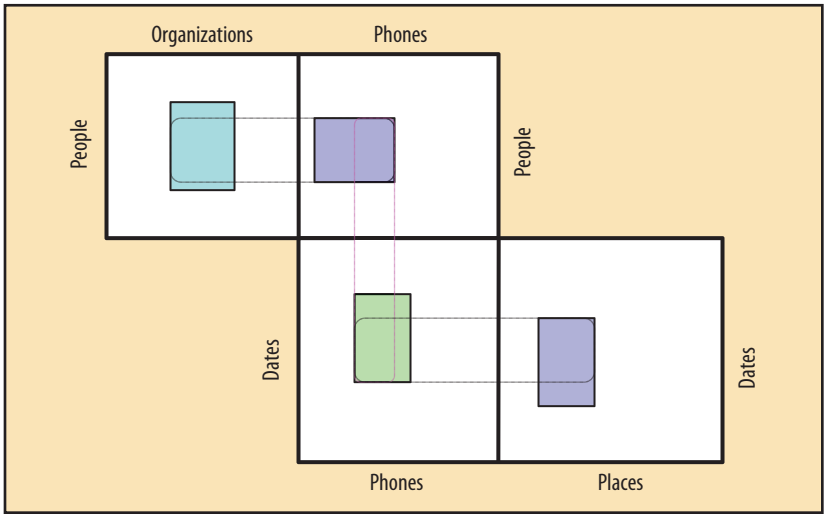

Figure 2. Bixplorer chains biclusters by matching sets of entities across common domains.

USER STUDY

To determine whether biclusters truly aid in sensemaking or impose additional hurdles, we conducted a small-scale study in which five participants, none of whom had prior experience with biclusters, used Bixplorer to analyze the Atlantic Storm dataset, a collection of 111 text documents, to uncover a fictitious terrorist plot.

The tool distilled the information in these documents into 437 unique entities using entity extraction algorithms. It discovered 4,257 relations among these entities and from these relations mined 1,001 biclusters. For this study, we set a minimum threshold for bicluster size of at least three rows and three columns, but with a lower threshold, Bixplorer could have generated many more biclusters due to their combinatorial nature.

We used a large high-resolution display as shown in Figure 3b outfitted with eight 30-inch LCD monitors, which provided a 10,240 × 3,200 pixel workspace. Almost no virtual navigation (scroll bars) was needed to navigate the workspace with the exception of the data browser, which can require vertical scrolling in searches with many results.

The study consisted of three parts. We first explained the nature of the Atlantic Storm dataset to participants and trained them in the use of Bixplorer on a separate training dataset. Next, we asked them to assume the role of an analyst and, in the space of two hours, investigate the main dataset for any indications of planned attacks. Finally, we asked the study

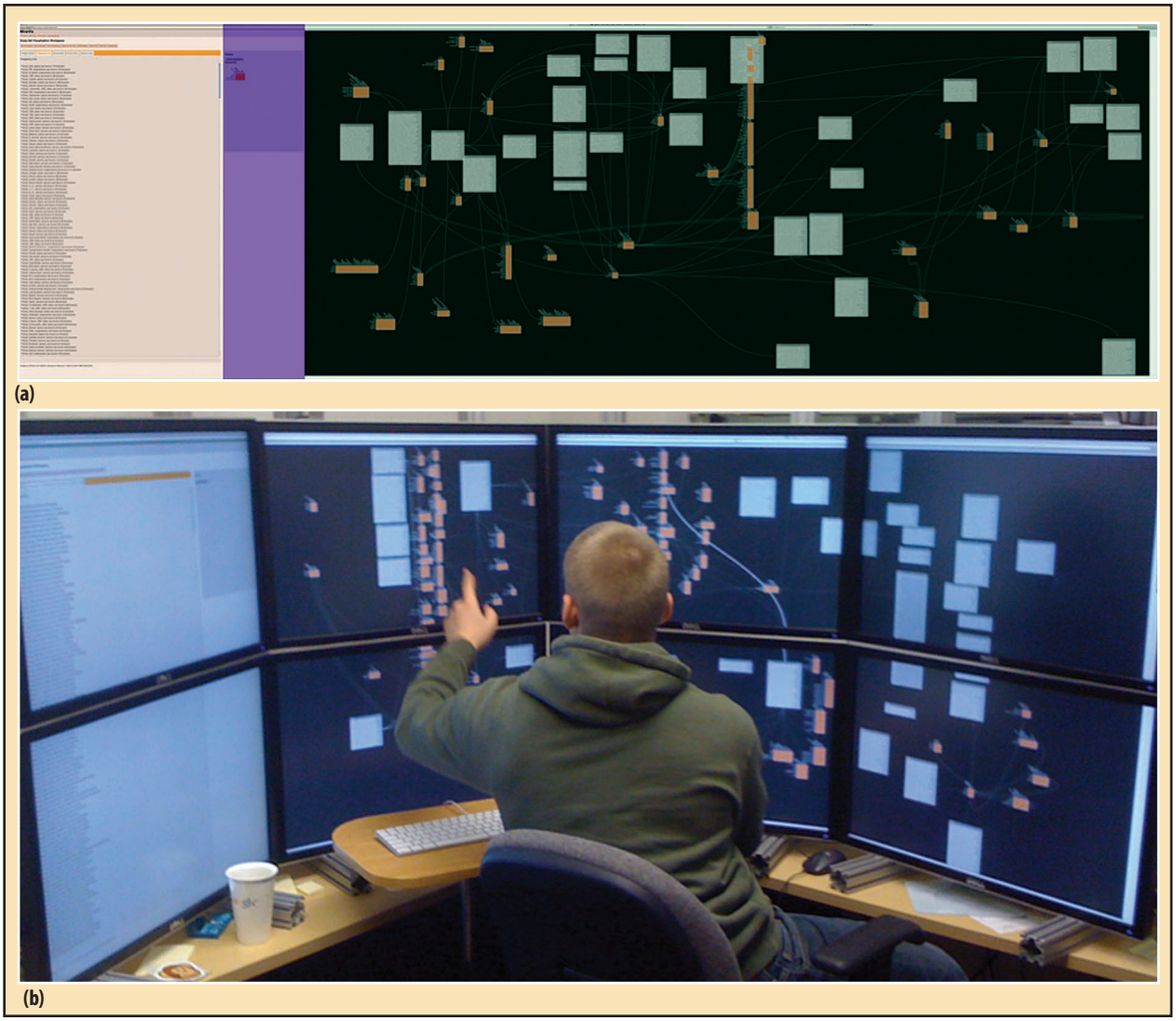

Figure 3. Bixplorer user interface. (a) The interface consists of three main views: a data browser (left), preview pane (middle), and workspace (right). (b) A user explores biclusters and documents in the workspace.

participants to report their findings and recommendations, and then interviewed them on their use of Bixplorer to determine whether and how biclusters benefitted their analysis.

STUDY RESULTS

Table 1 summarizes the study results. We examined each user's analytic process as well as the product of that analysis to assess how Bixplorer's features were used and the effectiveness of biclusters in supporting sensemaking.

Table 1. Bixplorer user study results.

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

Analytic process

All study participants integrated biclusters into their spatial analysis of the dataset, leveraging the visual representation of relationships in various ways.

Feature use. *Search* was the most heavily used feature in the data browser due to its versatility in returning results as both biclusters and documents. Study participants often used it to search for a particular entity. They also used it to investigate a specific person related

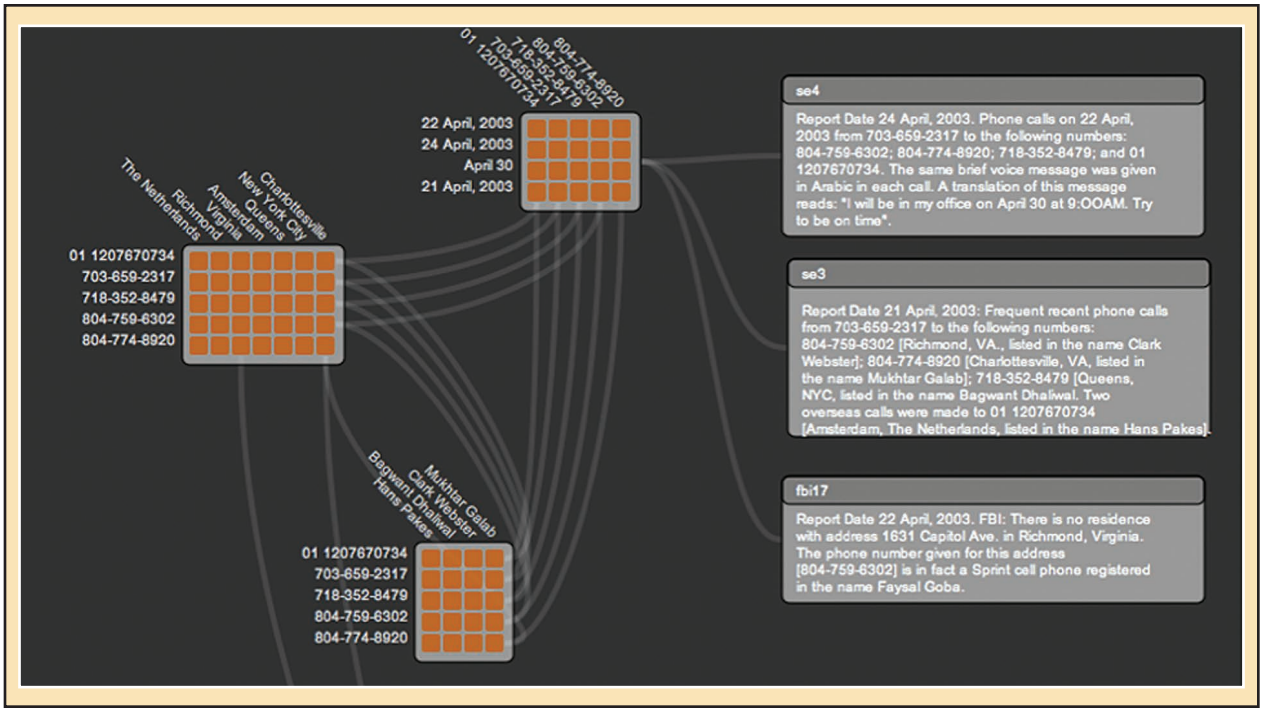

Figure 4. Spatial workspace of a Bixplorer study participant showing a chained set of biclusters along with connections to supporting documents.

to a document that might not have showed up in nearby biclusters.

Study participants generally used the *frequency list* feature to find a starting point for analysis. However, they didn't blindly select the most frequently occurring entities: "If it's everywhere," one user noted, "it can't be that important."

When they had a specific strategy in mind, study participants alternatively employed the *browse biclusters* feature. For example, one user began his analysis by browsing for biclusters involving money.

Finding leads. An important function of a visual analytics tool is to present users with cues to explore additional information. As one participant observed, "Biclusters give you a lead and then you have documents to look at."

Transitioning between biclusters and documents. Study participants exploited Bixplorer's ability to right-click on a bicluster to explore the documents involved and gain more context. One user explained that he "add[ed] a bicluster [to the workspace] when looking for relationships" and "the documents when [he] want[ed] the details."

Transitioning from documents to biclusters was also useful both to provide potential leads and to ensure adequate investigation of a specific topic or keyword. For example, one study participant scanned the biclusters of the current topic before exploring another topic because it "helped make sure I knew it all before moving on."

Rapid exploration of topics. Study participants used biclusters to rapidly move between complex topics. "[Biclusters] let me do quick jumps," one user stated. For example, finding the terms "Atlanta" and "students" in a given document might lead to other documents containing "student." At the same time, browsing the biclusters that contain "Atlanta" could reveal a broader set of related terms beyond those co-occurring in a single document.

Analytic product

Of the 1,001 biclusters Bixplorer extracted in this study, all of the users maintained less than 100; of the 111 documents in the dataset, none of the users' final layouts contained more than 50. These results suggest that Bixplorer helped users filter out unnecessary documents from their investigation. Participants also rarely relied on user-generated links and highlights, indicating that the bicluster links were adequate.

All of the users found Bixplorer to be helpful at obtaining evidence of potential terrorist activity by making it easier to navigate the large number of connections between entities in the dataset. As Figure 4 shows, one study participant used biclusters to reveal that four persons called the same phone numbers and that these numbers were associated with the same dates and same cities, and from these biclusters found documents pointing to possible collusion.

Users relied on terms along the edges of the biclusters to help recall specific findings. "They're like tags for the area I was in," one user explained.

BICLUSTER PROPERTIES

We identified three key bicluster attributes based on the user study results: area, entity domains, and content.

As indicated earlier, each bicluster in our study had a minimum size of three rows and three columns. Some were much larger in size—for example, three rows by eight columns, three columns by seven rows, and five rows by five columns. When asked about bicluster size, three of the study participants had no preference, and the other two preferred smaller biclusters. Although larger biclusters returned more documents, many of these were irrelevant to the user's investigation. In contrast, smaller biclusters included more relevant documents.

The types of relationships between entities on the sides of biclusters were also important to the study participants. All of them found biclusters between people and locations to be the most valuable. Money was useful for two of the users. In our study, dates appeared to be the least helpful, indicating that none of the users focused on a temporal strategy.

Study participants found entities that occurred rarely or were narrowly focused—for example, a city, an individual's name, or a street address—more meaningful than high-occurrence entities such as countries or states. Users inquired about ways to filter biclusters based on the specificity of content to make the workspace more manageable.

Although none of the users in our study had previous knowledge of biclusters, each was able to quickly integrate them into their analysis. These encouraging results show that biclusters could function not only as useful visual representations of relationships between entities but as information-rich glyphs with which users can easily interact.

Acknowledgments

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

Patrick Fiaux *is a graduate student in the Department of Computer Science at Virginia Tech. Contact him at pfiaux@cs.vt.edu.*

Maoyuan Sun *is a PhD student in the Department of Computer Science at Virginia Tech. Contact him at smaoyuan@cs.vt.edu.*

Lauren Bradel *is a PhD student in the Department of Computer Science at Virginia Tech. Contact her at lbradel1@vt.edu.*

Chris North *is an associate professor in the Department of Computer Science and director of the InfoVis Lab at Virginia Tech. Contact him at north@vt.edu.*

Naren Ramakrishnan*, Discovery Analytics column editor, is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of the university's Discovery Analytics Center. Contact him at naren@cs.vt.edu.*

Alex Endert *is a visualization researcher at Pacific Northwest National Laboratory. Contact him at alex.endert@pnnl.gov.*