PerCon: A Personal Digital Library for Heterogeneous Data

Su Inn Park, Frank Shipman
Center for the Study of Digital Libraries
Department of Computer Science and Engineering
Texas A&M University
1-979-862-3216
{joshsuihn, shipman}@cse.tamu.edu

ABSTRACT
Systems are needed to support access to and analysis of large heterogeneous scientific datasets. We developed PerCon, a data management and analysis environment, to support such activities. PerCon processes and integrates data gathered via queries to existing data providers to create a personal digital library of data. Users may then search, browse, visualize and annotate the data as they proceed with analysis and interpretation. Interpretation in PerCon takes place in a visual workspace in which multiple data visualizations and annotations are placed into spatial arrangements based on the current task. The system watches for patterns in the user’s data selection and organization and through mixed-initiative interaction assists users by suggesting potentially relevant data from unexplored data sources. PerCon’s data location and analysis capabilities were evaluated in a controlled study with 24 users. Study participants had to locate and analyze heterogeneous weather and river data with and without the visual workspace and mixed-initiative interaction, respectively. Results indicate that the visual workspace facilitated information representation and aided in the identification of relationships between datasets. The system’s suggestions encouraged data exploration, leading participants to identify more evidence of correlation among data streams and more potential interactions among weather and river data.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Relevance feedback
H.3.7 [Digital Libraries]: Systems issues, user issues

General Terms
Design, Experimentation, Human Factors.

Keywords
Heterogeneous data, management, data analysis, visual interpretation, spatial hypertext, mixed-initiative interaction.

1. INTRODUCTION
People need help in collecting, managing and interpreting data. A crucial issue for providing useful results from the current data explosion [6] is facilitating interactions with heterogeneous data sources. Advances in sensors along with new software for scientific data management/delivery means more data of more data types is available than ever before. Additionally, increases in interdisciplinary research put greater demands on scientists to bring together datasets from independent communities to better understand phenomena.

Tools supporting heterogeneous data management and analysis often focus on domain-specific representations and interfaces [6] or create separate silos for data of each data type (e.g. GenBank [3]). Domain-oriented visualizations are used for exploring and locating data. Indexing and classification of data most often occurs in relation to predefined structured representations developed for specific domains. We are interested in supporting the collection, management, and interpretation of unanticipated collections of data types—the type of idiosyncratic collections that occur during the formative stages of exploratory research. We liken these personal and small group data collections to personal book and document collections. Thus, our long-term goal is to support the ingestion, management, indexing, and interpretation of ad-hoc collections of data.

As a step towards this vision, we developed a personal or small group digital library system called PerCon (Personalized and Contextual Data Environment) that allows users to process, manipulate, analyze, and interpret diverse and interrelated data. PerCon is unique in that it features a persistent visual workspace and engages in mixed-initiative interaction with the user. The visual workspace is provided so that a user can explore and translate data into information visibly in multiple presentations (e.g. temporal, thematic, and spatial composition), and discover knowledge from information. To improve human data analysis, PerCon observes user behavior to infer user interests and locates and recommends related data within the current collection.

The following section overviews related work. Section 3 describes PerCon’s architecture, interface, and analysis capabilities. The evaluation of PerCon and evaluation results are discussed in Sections 4 and 5. Finally, Section 6 addresses contributions and future work.

2. RELATED WORK
Work informing our efforts comes from a variety of subfields: digital libraries for scientific data; data analysis and interpretation tools; and implicit relevance feedback.

2.1 Digital Libraries for Scientific Data
Digital libraries for scientific data often take the form of generic holders of unindexed data, modeling and providing access via generic metadata attached to data files via DSpace or other institutional repository software. In contrast, domain-oriented digital libraries can make use of domain-specific representations to index into datasets. In practice, these libraries often provide siloed access, limiting user queries to a particular type of data. For example, based on data integration, researchers in bioinformatics
have developed databases and computational/statistical analysis tools to explore different types of large-scale genome sequencing. Genbank [3] in the U.S is an implementation of huge databases functioning as a type of fine-grained digital library system. Diverse geography datasets have also been incorporated into domain-oriented data libraries. For instance, the Alexandria Digital Library [15] provides search services from collections of geographically referenced materials. Healthcare informatics digital libraries, such as Microsoft HealthVault [11], are used for managing and sharing personal health records collected from different devices. The capabilities of these systems are necessary for our vision of personal data digital libraries but they assume users have the expertise and tools for merging and analyzing the data located. The separation of data access from data analysis also reduces the likelihood that systems provide proactive support across this access/analysis border (e.g. recommendations of new data based on the user’s analysis activity).

2.2 Data Analysis and Interpretation Tools
Software tools for data analysis can be grouped into two broad categories: general purpose analysis tools and representation-specific tools. General purpose tools include MATLAB and GNU-based implementations such as Octave and R. These computing environments are widely used in research circles and have a large number of user-contributed libraries, but are too sophisticated for non-programmer end users. Spreadsheets (e.g. Microsoft Excel) are also commonly used for data analysis. These tools enable computation and visualization but require users to locate, import, and structure their data by hand.

Beyond such general purpose tools, data visualization and analysis capabilities have been included in domain-oriented digital libraries. For example, Bernard et al. [4] explore metadata visualization with respect to content-based similarity to lead the user to find relationships between data items. Abbasi and Chen [1] apply a linguistic feature-based visualization technique for analysis and categorization to represent textual data. Rowe et al. [13] represent spatial data visually to model complex 3D geometric data via interactive and sketch-based interfaces. Also, Booker et al. [5] visualize geo-temporal data that are visually presented by combining geospatial node maps and a timeline view with their GIANT system. PerCon builds on such capabilities but places their results in an analytic workspace in which users can organize and interpret data objects, much like the visual expression enabled in our prior work on the Visual Knowledge Builder (VKB) [14].

2.3 Implicit Relevance Feedback
Most data location and analysis tools leave the users in control, and on their own, during the performance of their task. But users’ actions in the system provide evidence of their interests that can be used to generate recommendations. Such recommendations have been well studied in the context of information location environments [9]. Conventionally, the system infers a model of the users’ interests, based on patterns in metadata or contents, and uses this model to locate related information. For example, the Curious Browser [8] uses observations of mouse usage, keyboard usage and the time spent viewing documents to generate suggestions and our prior work on the Interest Profile Manager builds a model based on activity in multiple applications [2].

3. PERCON: A PERSONAL DIGITAL LIBRARY FOR HETEROGENEOUS DATA
PerCon is more than a typical digital library since it integrates data management with data manipulation, presentation, and analysis. This section describes PerCon’s architecture, overall interface, visual workspace, and mixed-initiative recommendation subsystem.

PerCon was originally developed as a tool for a local research group to manage and analyze data from an investigation looking for patterns in physiological data from a variety of wearable sensors (e.g. heart rate), contextual data from portable devices (e.g. geocode and sound pressure level data), and behavior data (e.g. users’ answers to questions). PerCon has since been extended to work with data from other domains, as is the case in the evaluation reported in Sections 4 and 5.

3.1 PerCon Architecture
PerCon’s architecture for managing the interconnections and interoperations among the diverse software components and data resources consists of three layers: the resource layer, the middleware layer, and the application layer. Figure 1 shows the

Figure 1: PerCon’s architecture and software components in three layers
core capabilities at each layer and the components within it.

The resource layer provides functionality to store and preserve the original data objects, computed and filtered datasets, and metadata. It is implemented via a combination of a local repository and a database. The repository is designed to include loosely connected data sources. It stores and manages the raw and processed data objects as well as data relation objects that record similarities and differences between data objects. The database stores descriptive metadata, visual thumbnails of the digital objects, and indexes into the data objects stored in the repository. The resource layer also includes a web server to provide network applications access to the data resources in the repository and database. Finally, the resource layer records information about data provenance.

The middleware layer of the architecture includes functionality for data ingestion, data access, automated data analysis, and lower-level components that enable data visualization the data workspace, which is a modified version of VKB [14].

The data ingestion component provides three services: data processing, data integration, and provenance recording. The processing service includes a data reliability check to ensure that the data object content correlates with that expected for the data object type. The processing service also provides algorithms for processing existing data to generate computed data objects and metadata, and to populate the data relations objects in the repository that record correlations and other similarities/distances between data objects. The data integration service populates database tables used to index the data object repository. For example recording high and low values per day to enable searching the repository based on these content features of the data. Finally, the provenance recording service extracts metadata regarding the data provenance. This includes provenance of data ingested from other sources (e.g. where it came from and when), data entered from local sources (e.g. project and user IDs, data types, dates) and computed data (e.g. recording what operations were performed on which existing data streams and when).

The second functionality of the middleware layer is to enable the application layer components to access and interact with the resource layer's contents through a set of external APIs. A query-processing module parses requests and determines the communication necessary with the resource layer (e.g., database and repository) to fulfill the requests.

The third functionality of the middleware layer is data visualization, which instantiates data as a visual object in the workspace. The initial visual and spatial attributes of the visual object are assigned depending on the data type and query.

The final functionality of the middleware layer is data analysis. PerCon’s analysis framework associates data objects in the resource layer with metadata and visual properties determined by the user and workspace, ranging from the highly interactive to the highly automated.

PerCon’s application layer enables end-users and external systems to access the content in the digital library. This layer implements the user interfaces for browsing, searching, and visualizing the contents of the resource layer. A data registration interface enables adding, updating, and deleting data and datasets. The application layer also includes a data publication interface that enables remote access to certain contents of the digital library.

3.2 PerCon Interface

PerCon’s main interface is composed of three interface components: a repository browser, a visual workspace, and a suggestion viewer. Figure 2 shows this interface being used to analyze precipitation and river data.

The repository browsing interface organizes the digital objects (i.e., data files) into a hierarchy. The hierarchy ensures that raw data and the data objects computed from it are found together. Users can filter the view by selecting the types of data or files that are of interest. A property often important when analyzing heterogeneous data is locating concurrent data sources, i.e., temporal overlap in data capture. Users of the repository browser can bring up the list of overlapping data objects from a selected element in the hierarchy.

The hierarchic browser also allows users to preview the data objects in precomputed thumbnails. Previewing was found crucial
in early testing of PerCon as it aids the ability of users to rapidly locate initial data pertaining to their task while reducing undesired activity and complexity in the workspace.

Much of PerCon’s interface is a workspace for visualizing and organizing data objects. Users can drag objects from the repository browser into the workspace to generate a new manipulable visualization of the data. More than one visualization can be available for individual data types. The initial visualization is based on the data type and query used to locate the data but can be changed by the user. The workspace is described in more detail in the next subsection.

A history mechanism records user actions in the workspace. This history can be replayed to facilitate the comprehension of visualized information and to formalize interrelated knowledge. Based on user activity within the workspace, PerCon infers the user’s interests resulting in additional data being presented in the suggestion viewer. The suggestion viewer displays thumbnails of each recommended data object. The user can drag a suggested data object from the suggestion viewer into the workspace. The suggestion viewer preserves the recommendation history for future reuse. This recommendation history can reflect shifts in user interest or transitions between subtasks.

The displays of the three components of the user interface communicate to ensure presentation consistency. When a data object is selected in one component of the interface, it is indicated in all of the interface components.

In addition to the main persistent interface, PerCon includes a query interface for locating data within the collection. As the data objects increase in number and size, locating data within the main tree view can be challenging. The query interface supports searching for particular types of data in particular date ranges. Figure 3 shows the calendar view of query results. Data that matches the query is shown as a set of labels and visualizations on the days from which the data comes. The data visualizations in this interface use color to represent the type and content of the data. Each data type maps to its own unique color with different tones of that color used to represent different data values. Data entities in the calendar view can be dragged into the workspace and opened for a more detailed view.

### 3.3 Integrated Visual Workspace

PerCon extends the capabilities of our prior visual workspaces in that it includes a model for selecting among multiple applicable data visualizations according to different requirements. For example, the same stream of quantitative data can be presented as a plot showing the value over time or as a bar chart showing the relative frequency of values in different ranges. Besides the system-generated visualization of the data, each data object includes visual and spatial attributes (e.g. border/background colors, font styles) that users can manipulate to express interpretations of the data.

PerCon’s visual data object separates the base data object that is used for user expression and the application object which is used for data visualization. The base object model provides a method to add application objects onto the base object without constraints. Since movement, resizing and other event-based workspace interactions are managed by the base data object, only application-specific interactions must be considered when adding a new application object type. The result is a combination of human visual expression via base data objects and data visualization via application objects. Application objects can enable varied interactions with the data (e.g. moving from high-level to detailed views) using appropriate methods (e.g. zooming) within the application object portion of the data object.

A variety of application objects have been integrated in PerCon. Data editing and plotting tools, a HTML viewer, a multimedia data streaming application, interfaces to database tables, and timelines are available. In many cases, metadata standards and structured languages like XML and JSON are employed for encoding characteristics of the data objects stored in the repository. Hence, PerCon also includes XML and JSON viewers.

![Figure 3: Calendar visualization of query results](image3.png)

Figure 3: Calendar visualization of query results

![Figure 4. Multi-datastream synchronized viewer](image4.png)

Figure 4. Multi-datastream synchronized viewer
One PerCon-specific application object type is the multi-datastream synchronized viewer shown in Figure 4. This was developed for our original application domain of analyzing physiological and contextual data. Since these datastreams are recorded from different sources in parallel, an application object that integrated and visualized data from multiple data sources from the repository was desired to help identify patterns/relationships, as well as to form and assess new hypotheses.

Interoperating with the history mechanism, the workspace records all interactions. The captured records can be used to revisit and replay workspace activity [10] and are used for mixed-initiative interaction as described next.

### 3.4 Mixed-Initiative Recommendations

Knowing what content is available is a challenge for users of most any library but is particularly difficult in a library containing a large quantity of heterogeneous data. To improve data exploration and analysis in this context, PerCon includes a mixed-initiative interface for recommending data objects to the user. The key functions of the recommender subsystem are (1) understanding relationships within the heterogeneous datasets, (2) recognizing user interests from a record of user activity, and (3) associating the relationships with the user interests. The agent’s framework to accomplish these three functions is shown in Figure 5.

To avoid making recommendations that are perceived as random, we initially focused on getting users attention to the data they had not already seen that is most like what they have been recently examining. Future efforts can alter this strategy.

#### Understanding Relationships in Data

To understand relationships within the heterogeneous data collection, the agent generates precomputed tables of data object similarities. These tables are stored in the repository. Because different notions of similarity are important for data selection in different tasks, the processing framework computes five similarities/distances with values from 0 to 1 for all combinations of data objects of the same data type:

- Pearson correlation coefficient – standard measure of similarity of values over time in data elements;
- cosine similarity – breaks the data elements’ timestream into segments and aggregates data in each segment into a single value; the two sequences of values are compared as vectors;
- temporal similarity – how close together in time are the two data elements, value is 1 if the data elements overlap and is 0 if the time gap exceeds a preset threshold;
- trendline similarity – computes the Euclidean distance between the two-dimensional tuples composed of (the slope of linear regression of data readings, the sum of values of readings); and
- variance similarity – difference between the variances of the two data elements.

Before the similarities can be computed, interpolation and smoothing are necessary for comparing data elements with different sampling rates and/or missing values.

Overall similarity of two data elements is a weighted sum of the five similarities above. The initial weights for the five similarity metrics are set heuristically based on experience with the system. These weights are modified based on user interaction with recommendations by increasing the weight for similarity metrics that correlate with accepted recommendations.

The similarity assessment mechanism is extensible. Our initial measure of overall similarity was much simpler, using a combination of the mean values and variances, but the resulting recommendations were quite bad. Iterative testing of the mechanism led to the set of similarity metrics described. Additional metrics can be included in the overall assessment if experience indicates the need to include new features.

#### Recognizing User Interests and Selecting Recommendations

PerCon records user activity in the workspace as a sequence of events (e.g. dragging data elements from the repository browser into the workspace to create a new data object, resizing the data object, etc.). In addition, each event type has been heuristically assigned an evidence weight (e.g. data object creation is 0.1, resizing a data object is 0.01, etc.). Using the event log and the list of weights, PerCon generates a model of user interest as follows:

1. Each event in the workspace is added to a table recording the event type and features of the data object involved in the event (e.g., data type, application object type, color1, color2, annotation) and the evidence weight for the event is added to an evidence tally.
2. When the evidence tally exceeds a threshold, the table is used to train a Bayesian network composed of nodes modeling the event features in order to predict the data type of greatest interest.
3. The data objects in the workspace of the predicted data type are ranked based on the table of activity and those above a threshold are selected as reference data objects.
4. Similarity between the data elements of the predicted data type not already in the workspace and the reference data objects is computed based on the weighted sum of the five similarity metrics.
5. The five most similar objects that are not already in the history of recommendations are added to the suggestion viewer as thumbnails and the evidence tally is reset to zero.
Rather than building a model around a fixed notion of feature importance, the probabilistic model aims to capture the unforeseeable characteristics of practical action [16]. The Bayesian network represents relationships between a data object and the features (e.g. visual/spatial attribute, application, events) and computes the (conditional) probabilities based on a sequence of events. As users proceed through a data analysis task, the sets of data that are their focus, the applications they use to view the data, and the interpretive coding and annotation they apply in the workspace are likely to change [17] [18]. By including the event and data object features in the Bayesian network, recommendations can take into account such shifts in behavior.

A last component to the recommendation system is controlling the frequency of recommendations. As described in the steps above, the accumulation of activity into the evidence tally is one method for controlling recommendations. Users have to perform enough actions in the workspace to (re)fill the evidence tally before a recommendation is made. In order to build recommendations that represent different lengths of user activity, the system includes shorter-term and longer-term evidence tallies. This means there can occasionally be multiple recommendations simultaneously or quite near together.

In addition to inferring user interests, the mixed-initiative interface allows a user to express his or her interests explicitly and to ask for recommendations. This is important because user activities in the workspace may not reflect user interests in cases when there are sudden task shifts, when the user performs multiple tasks in parallel, or when significant effort is required to examine data to locate data of interest. Also, a user may want to explore data different from that which was recently examined. To request recommendations, users select a workspace data object and request related data. Subsequently, the explicitly expressed interest is included in the agent’s user interest model and thus affects future system-generated recommendations.

4. User Study

PerCon is a large software environment and the result of many different design hypotheses. We conducted a user study to observe data analysis in PerCon focusing on two central hypotheses: (1) the visual workspace helps a user to manage data and to translate data into knowledge about the domain, and (2) the mixed-initiative recommendations improve a user’s ability to explore and analyze data. The participants were asked to perform several specific tasks with/without the visual workspace and with/without the mixed-initiative interaction.

4.1 Participants

PerCon is meant to support people who need to analyze heterogeneous datasets. Thus, our population of convenience, the students and researchers at a large academic institution, is representative of the target population. The twenty-four participants included one undergraduate, four Masters and sixteen PhD students, and three postdoctoral researchers. The participants ranged in age from 24 to 36 and represented a variety of disciplines: computer science, computer engineering, electrical engineering, soil hydrology, biomedical engineering, industrial engineering, and management information systems.

4.2 Domain Data for Participant Analysis

Data for the study was selected to include relations that are intuitive in nature but complex in detail. In particular, we selected weather and river data along two different geographic regions of a river system to provide two comparable sets of tasks with which to examine the effects of alternate configurations of PerCon.

Since weather and river data are recorded hourly or daily over decades, they provide a voluminous dataset. Weather data includes many variables that exhibit a great deal of spatial and temporal correlations with one another. Many variables in weather data also have an observable impact on environmental variables. River stream level is a representative environmental variable significantly affected by weather conditions. Depending on geographic relationships and intermediate reservoirs and dams, river stream levels exhibit a strong relationship between upstream and downstream values. Data was collected from two public repositories: the National Oceanic and Atmospheric Administration [12] and the Brazos River Authority in Texas [7].

Two years of weather and river data (from 2011 to 2013) were ingested into PerCon for this study. The weather data consists of five elements: temperature, precipitation, relative humidity, wind speed, and wet bulb temperature. The river data includes two data elements: the river level and discharge recorded. Data was collected from six different locations in Texas along the Brazos River and its tributaries. One data set included the data from College Station, Waco, and Temple while the other included data from South Bend, Seymour, and Fort Griffin.

The value for each data element was ingested for each hour, resulting in about 25 MB of raw data and 229 Mbytes of computed similarity data for each of the two tasks. To facilitate access to the data, a directory in the repository browser for each data element provided participants access to annual, monthly, and daily segments of data.

4.3 Tasks

Participants were asked to carry out three tasks with each of the two datasets. These tasks motivated participants to perform a complete cycle of data exploration, manipulation, management, analysis, and interpretation. Throughout the tasks, we provided the participants with a basic approach and methodology; each task included step-by-step procedures. To observe and discover how weather and river data are correlated, one possible approach is to classify and organize data objects based on individual changes, trends, and patterns in the data.

Task 1: Participants had 20 minutes to organize and classify river level and precipitation data according to common trends, quantities, durations, or other user-perceived criteria.

Task 2: With the classified weather and river data from Task 1, participants were then asked to investigate the implications and identify correlations among the classified data. In particular, participants were asked to investigate the data correlations focusing on: what and how weather factor(s) affects river level, and how the river data from different locations (e.g. Waco, Cameron, and College Station) are correlated. They had 10 minutes for this task.

Task 3: With the discovered evidence of relationships among the weather and river data elements from Task 2, participants were asked to explain river level changes/trends based on weather and upstream river flow conditions for 5 minutes. For example, some participants estimated river factors that caused these changes/trends (such as delay time) based on past weather data and other river stream conditions. Other participants interpreted why some river level changes are more or less affected by other river level changes.
4.4 System Conditions
For this study, our main hypotheses concern the effects of the workspace and mixed-initiative recommendations on data exploration and interpretation. To evaluate our hypotheses, we compared four PerCon configurations varying the availability of the visual workspace and recommendations. These are shown in Figure 6. Without the visual workspace, the top two configurations in Figure 6, the system provided up to two information objects (i.e. application windows) with spatially preset size and location. The left two configurations in Figure 6 show when the mixed-initiative data recommendation system was turned off.

When participants performed the tasks without the visual workspace, they could not use visual attributes and spatial organization to express themselves. In these configurations, they were allowed to use other application(s) such as MS Word or Excel to write down notes, intermediate and final task results, etc.

4.5 Procedure
Before conducting the given tasks, participants were trained in the use of PerCon. This involved watching a 10-minute video tutorial to follow along with a written tutorial/manual to ensure consistent training. After being led through use of features important to the tasks they would be given, participants had 5-minutes to try out features on their own.

Each participant was asked to perform the three tasks in Section 4.3 in two of the four system conditions. The order of exposure to system configuration and data set were balanced to account for learning effects, interactions between experiences with configurations, and complexities inherent in the two data sets. Table 1 shows the six evaluation groups covering the combinations of different interface modes. Four participants are in each evaluation group.

Table 1. Evaluation groups and interface modes

<table>
<thead>
<tr>
<th>Group</th>
<th>System interface modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Configuration 1 and 2</td>
</tr>
<tr>
<td>Group 2</td>
<td>Configuration 1 and 3</td>
</tr>
<tr>
<td>Group 3</td>
<td>Configuration 1 and 4</td>
</tr>
<tr>
<td>Group 4</td>
<td>Configuration 2 and 3</td>
</tr>
<tr>
<td>Group 5</td>
<td>Configuration 2 and 4</td>
</tr>
<tr>
<td>Group 6</td>
<td>Configuration 3 and 4</td>
</tr>
</tbody>
</table>

At the end of the user study, the participants were given a questionnaire to explore the effects of the visual workspace and recommendations in each condition. The overall study duration for each participant was 120 minutes: learning how to use the system for 15 minutes, performing the given tasks for 70 minutes, and answering the questionnaire for 20 minutes.

4.6 Result Data Collection
Data about participant activities and experiences was collected from four sources: (1) Likert-scale responses about the task and PerCon, (2) open-ended questions, (3) the final and intermediate user-created workspaces, and (4) a record of time-stamped events/interactions with PerCon throughout the tasks.
Table 2. Likert-scale questions

<table>
<thead>
<tr>
<th>Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workspace</strong></td>
</tr>
<tr>
<td>Q1 I had enough support to understand data content in the workspace</td>
</tr>
<tr>
<td>Q2 I had enough support to express relationships in the way I wanted</td>
</tr>
<tr>
<td>Q3 It was easy to interpret and characterize given/created objects in the workspace</td>
</tr>
<tr>
<td>Q4 I had enough support to effortlessly / quickly browse and select data</td>
</tr>
<tr>
<td><strong>Mixed-initiative interaction</strong></td>
</tr>
<tr>
<td>Q5 I was satisfied with the data suggested</td>
</tr>
<tr>
<td>Q6 I was satisfied with the suggestion request</td>
</tr>
<tr>
<td>Q7 I had enough support to find and interpret data I was interested in</td>
</tr>
<tr>
<td>Q8 I had enough support to find correlations within the dataset</td>
</tr>
</tbody>
</table>

Table 2 lists the eight statements included as 7-point Likert-scale responses (1 means “strongly disagree” and 7 means “strongly agree”) concerning the effect and usefulness of the workspace and mixed-initiative interaction. Participants in Configurations 1 and 2 were told to consider the fixed layout configuration of the workspace combined with their chosen external tools as their effective workspace.

5. RESULTS

The data from the Likert-scale questions will be reported first to give a sense of user perceptions of the different configurations. This will be followed by a more detailed analysis of the activity logs and workspaces to examine how the different configurations objectively changed data analysis practice.

5.1 Perceptions of Participants

Figure 7 presents the mean and standard error of participant assessments for the four workspace-related statements after using each interface configuration. The distributions for the four statements are all similar. The means for responses for Configurations 3 and 4, which include the workspace, vary between 5.5 and 6.5 out of 7. The means for Configurations 1 and 2, which did not include the workspace, were between 1.75 and 3.25. In all cases, the difference is statistically significant (p<.001 for the closest, Q3). In the case of question 3, a few participants valued the larger fixed-size graphs when analyzing data, explaining it provided a “bigger plot” and was “easy to check the pattern” in their open-ended responses.

Figure 8. Responses to questions related to recommendations

5.2 Participant Work Practices

To investigate the effects in each interface mode, we also examined the work practices of each participant group. We first examine the effect of interface configuration on the number of data elements examined by participants. Next we look at the distribution and pattern of activity in the repository browser and workspace in the different configurations. Finally, we will look at the interleaving of system-initiated and user-requested suggestions.

Figure 9. Average number of data objects classified or analyzed during the tasks

Number of Data Elements Examined. Figure 9 shows how many data elements were examined during the 35 minutes spent with a single data set in the four conditions. Without either the workspace or the suggestions, the average was around 11. When only suggestions were added, the mean rose to 17 and when only the workspace was added the mean went to 31. With both the workspace and suggestions, participants examined about 41 data elements on average. It is clear that having a drag-and-drop interface that supports the visualization of an open-ended set of data elements facilitates rapid data analysis. The effect of the workspace on number of data elements analyzed is strongly significant (p<.001, t-test) when comparing the activity in...
The effect of mixed-initiative recommendations on number of data elements examined was also significant (p<.04).

The more data elements the participants explored, the more evidence they discovered to identify and explain relationships between data elements. Thus, the results of their Tasks 2 and 3 were strengthened. Furthermore, participants receiving recommendations were able to substantiate correlations over broader time periods.

Comments in the open-ended questions confirmed that the workspace supported rapid and persistent analysis:

“Obviously, there is still quite a large amount of data and sources to sift through. Being able to collect data objects together, stack them, and reshuffle at will allowed for more opportunity to see potential correlations in data that otherwise might have gone unnoticed.”

Distribution of Activity. Given the same tasks but different interface configurations, participants showed different work practices. Without the visual workspace and recommendations (Config. 1), 40% to 97% of the activities/interactions occurred in the repository browser. Without the visual workspace, users focused on and spent more time searching for data using data previews in the repository browser (avg = 0.71, std = 0.24). On the other hand, with the visual workspace and without recommendations (Config. 3), 11% to 69% of the activities were recorded in the repository browser (avg = 0.41, std = 0.21). Figure 10 shows the event sequences in the repository browser and the workspace for the four participants in Group 2. The availability, or lack thereof, of the workspace had a strong effect on individual work practices.

Figure 10. Ordering of user events in repository browser and workspace shows distinct patterns of work.

Without the visual workspace, users focused on and spent more time searching for data using data previews in the repository browser. They then relied on their short-term memory or notes taken in other applications to get back to data. However, with the visual workspace, users spent more time exploring potential data relations in the workspace. In particular, visual data objects for correlated data elements were visually annotated and interacted with in ways that made their relationships significant and meaningful, easing Task 2 and 3 activities.

Recommendations. As already indicated, participants were more willing to make use of recommendations when their configuration had a workspace. Thus, to evaluate the effectiveness of recommendations, we examine the twelve uses of configuration 4, which included both the workspace and recommendations. All twelve of the participants in this configuration accepted and explored some of the suggested data. Indeed, 11% to 49% of the data objects in their workspaces were from recommendations.

Figure 11. A sequence of suggestion events

Ten of the twelve participants used the ability to request suggestions. Figure 11 shows the temporal sequence of suggestion events triggered by the system and requested by users during the task. This shows at least half of the participants requested suggestions fairly frequently over the 35 minute period. This continued use implies the recommendations were seen as being valuable. This was corroborated by open-ended comments such as:

“When I wanted to find data with similar pattern/trend, the recommendations reasonably provided me with data objects that resembled what I was looking for. There were other times when I wanted to have [data with new patterns] recommended, e.g. only considerable amounts of rain in College Station, and this is where I tended to have to look more for myself or hunt more in the requested recommendations to find what I was looking for.”

6. DISCUSSION AND CONCLUSIONS

We started this project with the goal to build an environment that would integrate much of the data ingestion, management, and analysis activities for an on-going data-intensive research project. As such, the original system architecture and development focused on the system-side perspective of storing, processing, and visualizing data. As development continued, our focus shifted to the human activities of locating, annotating, and interpreting data.

In terms of data location, PerCon’s main interface provides metadata-based access to the collection via the repository browser. A weakness of our initial implementation was that it just showed the limited metadata of the data elements (e.g. data type, date/time, location or name). This completely obscured the data contents, making it hard to know what the data was really like until it was brought into the workspace by the user, resulting in workspace management issues (e.g. continually rearranging and deleting content). This observation led to the addition of thumbnails for previewing data in the repository browser.

In addition to browsing the collection, PerCon supports queries that filter the content presented in the browser. Because many domains require understanding temporal relations among data elements, query results can also be examined in a calendar view. The calendar visualization uses a combination of labels and colorized thumbnails to present the types and contents of data elements.
PerCon is built around a main workspace for interpreting data. The workspace enables the visualization of heterogeneous data types and the expression of interpretation through annotation and visual structure. The user study showed that the workspace has a large effect on data analysis practice in terms of the number of data elements classified, the time spent locating vs. time spent interpreting data, and the users’ perceptions of system support. The ability to develop persistent task-oriented workspaces from data collections seems crucial to efficient and effective data analysis.

Data location is traditionally a user-driven activity. Our efforts to recommend data objects based on user activity in the workspace aimed to overcome the difficulty of users not knowing what data is available in a collection. The development of a multi-faceted approach to similarity assessment and probabilistic reasoning about user interests combine to generate recommendations. The user study showed that recommendations increased the number of data elements classified with and without the workspace and that many of the users valued the recommendations enough to take the initiative to ask for recommendations.

There are a number of directions for future work: improving workspace interactions, extending the recommendation subsystem, and exploring the use of PerCon in new domains and with new user communities.

With the ability to add most any Java application as a data object in the workspace, PerCon has the potential for broad use. In practice, most data analysis proceeds via a simple graphing object type. We want to explore data visualization objects that are more dynamic and tailorable. For example, workspace objects that synchronize the presentations of time-aligned data elements. Another example is extending the data graph object to allow users to merge data elements into a single presentation. If the user drags one data graph object into another data graph object, the two are presented in a single graph. This new data object type would allow the user to pull a data line back out of the graph to separate the two again. This mode of interaction is common in tabbed web browsers where users can put windows together and pull tabs out to create new windows.

The recommendation subsystem shows the potential for data analysis environments to be more proactive in supporting users. The current approach to recommendation generation assumes the user wants more data of the same data types with similar characteristics to those already included in the workspace. A natural extension is to look for similarities across data types. Because the phenomena being measured in the data are often dissimilar, an alternative similarity metric is likely to be needed.

PerCon was originally designed with the expectation that users are either professionals engaged in this type of work or are otherwise knowledgeable about data types and data manipulations. The user study showed the system was usable by such a population. We would like to explore the use of a PerCon-like interface for more casual use in the domain of personal health. Enabling people to explore the data collected on their smart watches, health monitoring devices, cell phones, etc. could lead to improved mental models about how their lifestyle and health are intertwined.

7. ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation grant DUE–0938074.

8. REFERENCES