Visual Browsing of Remote and Distributed Data

Parthasarathy Krishnaswamy* Stephen G Eick† Robert L Grossman‡

National Center for Data Mining
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL, USA

Abstract

Data repositories around the world hold many thousands of data sets. Finding information from these data sets is greatly facilitated by being able to quickly and efficiently browse remote data sets. In this note, we introduce the Iconic Remote Visual Data Exploration tool(IRVDX), which is a visual data mining tool used for exploring the features of remote and distributed data without the necessity of downloading the entire data set. IRVDX employs three kinds of visualizations: one provides a reduced representation of the data sets, which we call Dataset Icons. These icons show the important statistical characteristics of data sets and help to identify relevant data sets from distributed repositories. Another one is called the Remote Dataset Visual Browser that provides visualizations to browse remote data without downloading the complete data set to identify its content. The final one provides visualizations to show the degree of similarity between two data sets and to visually determine whether a join of two remote data sets will be meaningful.

CR Categories: H.1.2 [Models and Principles]: User/Machine Systems—Human information processing; H.2.8 [Database Management]: Database Applications—Data mining; H.3.4 [Information Storage and Retrieval]: Systems and Software—Distributed systems;

Keywords: visual data mining, information visualization, visual data exploration, distributed data

1 Introduction

Data repositories around the world contain many tens of thousands of datasets and are growing exponentially. It is difficult now and in the future it may not be possible, even with high performance networks, to consolidate the data into a single repository for a data mining analysis. The reasons for this are:

- *Time and effort.* it is not practical to download all relevant data.
- *Size.* some datasets are too big to move.
- *Legal.* it may be against the law to combine certain datasets when the result will violate privacy laws.
- *Complexity.* the structure of some repositories may be too complex to be moved.
- *Data Revisions.* many datasets are constantly updated and downloading a snapshot could bias an analysis.

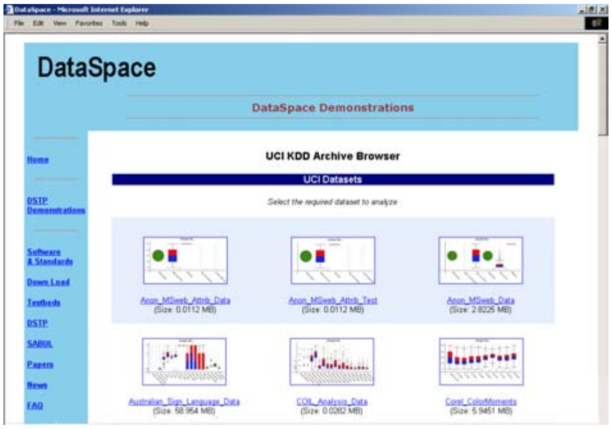


Figure 1: Dataset Icons for the UCI KDD Archive.

To address these problems we have developed a tool called IRVDX for distributed visual data browsing. It addresses three major questions:

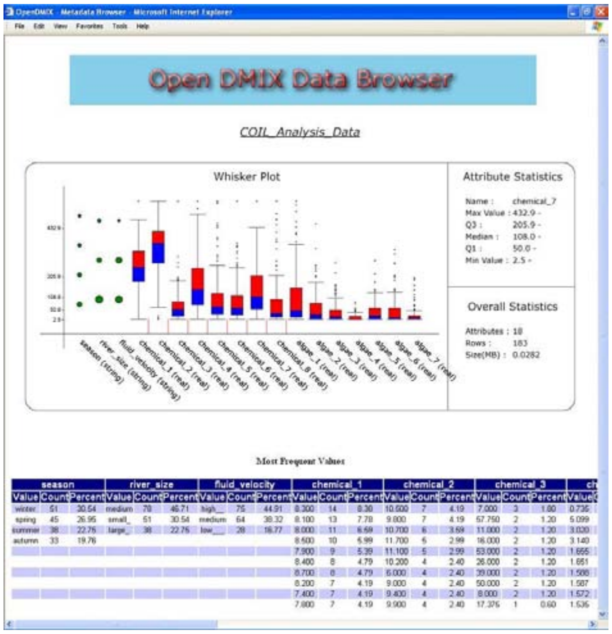
- How to identify remote and distributed data sets that may be relevant to an analysis problem?
- How to browse remote data without first downloading the complete data set?
- How to identify columns in distributed data sets that can be used to fuse the data sets in order to extract additional insights?

In IRVDX we introduce the concept of Dataset Icons to identify relevant datasets from remote distributed repositories. A Dataset Icon is a reduced representation of a dataset that shows summary and statistical characteristics of the data set that are important for data mining. A Dataset Icon provides a succinct visual summary of a remote dataset. Our first Dataset Icon, shown in Figure 1, represents a data set using modified box plots thumbnails. The icon shows the number of attributes in the dataset (columns), type of attribute (continuous or categorical), distribution of attributes in the dataset, and size of dataset. Our users, data miners, tend to be familiar with this representation and easily understand it.

The next problem for a scientist after he or she finds a dataset that looks relevant is to browse the data. Figure 2 shows our remote dataset browser. Users click on the icon to pull up our browser. The intent is to provide a simple remote browsing capability to help the scientist avoid pulling over irrelevant data.

The visualization in Figure 2 consists of two panes. The top pane summarizes the statistical distributions of the attributes. The summary information includes the name of the column, number of attributes, number or records, and size of the dataset. The lower

*e-mail: babuk@microsoft.com
†e-mail: eick@cs.uic.edu
‡e-mail: grossman@uic.edu



Attribute Statistics

Name : chemical_7
Max Value : 432.9 -
Q3 : 205.9 -
Median : 108.0 -
Q1 : 50.0 -
Min Value : 2.5 -

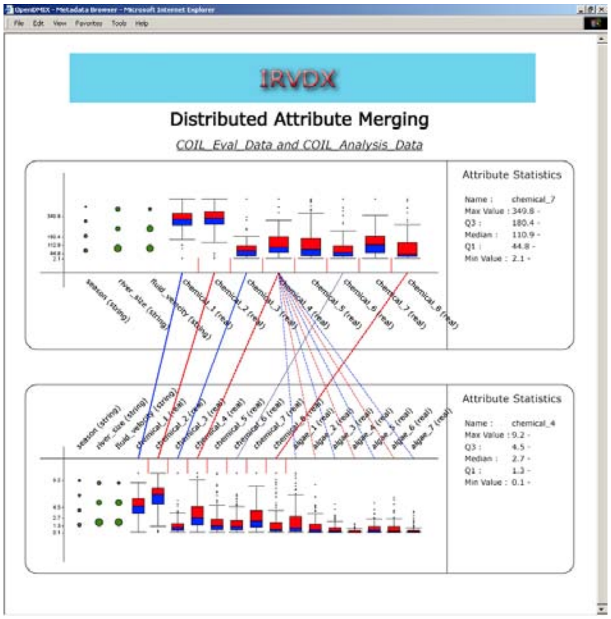
Overall Statistics

Attributes : 18
Rows : 183
Size(MB) : 0.0282

Most Frequent Values

season			river_size			fluid_velocity			chemical_1			chemical_2			chemical_3			ch
Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value
winter	51	30.54	medium	78	46.71	high__	75	44.91	8.300	14	8.30	10.500	7	4.19	7.000	3	1.80	0.735
spring	45	26.95	small_	51	30.54	medium	64	38.32	8.100	13	7.78	9.800	7	4.19	57.750	2	1.20	5.099
summer	38	22.75	large_	38	22.75	low___	28	16.77	8.000	11	6.59	10.700	6	3.59	11.000	2	1.20	3.020
autumn	33	19.76							8.500	10	5.99	11.700	5	2.99	16.000	2	1.20	3.140
									7.900	9	5.39	11.100	5	2.99	53.000	2	1.20	1.655
									8.400	8	4.79	10.200	4	2.40	26.000	2	1.20	1.651
									8.700	8	4.79	6.000	4	2.40	39.000	2	1.20	1.506
									8.200	7	4.19	9.000	4	2.40	50.000	2	1.20	1.587
									7.400	7	4.19	9.400	4	2.40	8.000	2	1.20	1.572
									7.800	7	4.19	9.900	4	2.40	17.375	1	0.60	1.536

Figure 2: Remote Dataset Browser.



Attribute Statistics

Name : chemical_7
Max Value : 349.8 -
Q3 : 180.4 -
Median : 110.9 -
Q1 : 44.8 -
Min Value : 2.1 -

Attribute Statistics

Name : chemical_4
Max Value : 9.2 -
Q3 : 4.5 -
Median : 2.7 -
Q1 : 1.3 -
Min Value : 0.1 -

Figure 3: Relationship between data sets.

pane shows the most frequent values, ranges, and a few other statistics. By looking at the display a data miner should be able to decide if the full dataset would be appropriate for further analysis.

Finally, Figure 3 shows the relationships between columns in two datasets. The relationships are calculated by comparing data types and using the statistical distributions to suggest columns that may be related. We currently use a simple pseudo metric based on the inter-quartiles of the distribution to determine similarity.

2 Related Work

Related work for this research is concentrated in the fields of information visualization and visual data mining. Papers [8] and [4] present well-known classifications of information visualization techniques based on the data types and visualization techniques. We mainly used 2-dimensional visualization techniques. There are other notable techniques for information visualization currently available [5],[7],[6]. Our notion of interaction and direct manipulation follows [1].

Iconic visualizations are also studied extensively in many fields. The paper [10] describes the use of icons as symbolic parametric objects to visualize feature attributes. Our approach is to use Dataset Icons as a tiny representation of the statistical and data mining characteristics of a data set. Our approach also makes use of some of the Visual Scalability techiniques presented in [3].

Existing visual data mining research [9] focuses on improved techniques for understanding various classes of data and on understanding aspects of the underlying data mining models. This research work focuses more on distributed and remote data sets and on the information about a data set, where we provide visualizations to figure out which data sets are relevant for a particular analysis.

3 Conclusion

We have implemented IRVDX and have it running at the National Center for Data Mining. We have created a meta-data repository where we store statistical properties of the datasets and other meta-data necessary to produce our icons. The images are pre-computed for performance. Our visualizations are implemented using SVG graphics for portability and run in essentially any browser.

Although we have not done a CHI-style formal user study, our users, data miner, find that our interface is significantly better than the textual descriptions of datasets and text-based searches that are commonly used by data repositories. Our experience is that it is much easier and faster to navigate through large repositories using dataset icons than textual dataset descriptions.

If network conditions permit, we will demo IRVDX live at InfoVis 2004 conference in Texas.

References

[1] C. Ahlberg and Ben Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of Human Factors in Computing Systems*, pages 313–317, 1994.

[2] Stephen G. Eick. Visual discovery and analysis. *IEEE Transactions on Computer Graphics and Visualization*, 6(1):44–59, 2000.

[3] Stephen G. Eick and A. F. Karr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, 2002.

[4] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions On Visualization And Computer Graphics*, 8(1):1–8, 2002.

[5] L. Nowell, S. Havre, B. Hetzle, and P. Whitney. Theme river: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[6] R. M. Pickett and G. G. Grinstein. Iconographic displays for visualizing multidimensional data. In *Proceedings of IEEE Conference on Systems, Management, and Cybernetics*, pages 514–519. IEEE, 1998.

[7] Ben Shneiderman. Tree visualization with treemaps: A 2d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.

[8] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[9] Ben Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[10] Th. van Walsum, F.H. Post, D. Silver, and F.J. Post. Feature extraction and iconic visualization. *IEEE Transactions On Visualization And Computer Graphics*, 2(2):111–119, 1996.