MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Visualizing Set-valued Attributes in Parallel with Equal-height Histograms

Wittenburg, K.; Malizia, A.; Lupo, L.; Pekhteryev, G.

TR2012-040 May 2012

Abstract

Visualization of set-value attributes in multi-dimensional information visualization systems remains a relatively unexplored problem. Here we introduce a novel method for visualizing set-value attributes that we call the singleton set distribution view and integrate it into an interactive multi-dimensional attribute visualization tool utilizing parallel bargrams (aka equal-height histograms) as its main visual motif. We discuss our design rationale and report on the results of an evaluation study.

International Working Conference on Advanced Visual Interfaces (AVI)

Visualizing Set-Valued Attributes in Parallel with Equal-Height Histograms

Kent Wittenburg*
Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139 USA
wittenburg@merl.com
*and U. Carlos III de Madrid

Alessio Malizia
Universidad Carlos III de Madrid
Avda. de la Universidad, 30
28911-Leganés, Madrid
Spain
amalizia@inf.uc3m.es

Luca Lupo
Universidad Carlos III de Madrid
Avda. de la Universidad, 30
28911-Leganés, Madrid
Spain
llupo@inf.uc3m.es

Georgiy Pekhteryev
Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139 USA
gpekht@gmail.com

ABSTRACT

Visualization of set-valued attributes in multi-dimensional information visualization systems remains a relatively unexplored problem. Here we introduce a novel method for visualizing set-valued attributes that we call the singleton set distribution view and integrate it into an interactive multi-dimensional attribute visualization tool utilizing parallel bargrams (aka equal-height histograms) as its main visual motif. We discuss our design rationale and report on the results of an evaluation study.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Design, Experimentation, Human Factors.

Keywords

Information visualization, set-valued attributes, multi-dimensional visualization, equal-height histograms.

1. INTRODUCTION

Set-valued attributes are a frequent, naturally-occurring data type in a wide variety of domains. Objects in digital libraries can have set-valued attributes such as authors, references, and citations. Patents naturally have one or more inventors, related patents, related products, assignees, and backward and forward citations. While relational databases might represent such information with a set of inter-related tables involving duplicated rows, such representations are not appropriate to present directly to users.

Multi-faceted browsing methods are commonly deployed that do incorporate set-valued attributes. For example, FacetLens [4] presents data in a variation of a space-filling hierarchical layout in which set-valued attributes may be shown and used as the basis for exploration and filtering. However, there is no direct visualization of the full *distribution* of attribute set values.

An alternative is to use histograms (either equal-width or equal height) for interactive visualization and exploration of multi-dimensional data [5][7][8][9][11]. Histograms have the property that they do reveal the full relative distribution of attribute values. The Focus system [8] and InfoZoom [5] use equal-height histograms to reveal multi-attribute value distributions and allow users to explore and filter the space flexibly. Attribute Explorer [7][9] and EZChooser [11] use equal-width and equal-height histograms, respectively, and incorporate interactive features for "previewing" queries, affording a "see and go" interactive paradigm rather than a "go and see" paradigm [6]. The "see and go" paradigm allows users to discover trade-offs and acquire mental models of data prior to filtering the set of items under consideration. As for set-valued attributes, however, none of the histogram-based systems cited above handle set-valued attributes except by treating each possible set value as a separate Boolean attribute.

(a) (b)

Figure 1. Prior work in set-valued histograms: (a) the Set'o'gram [3]; (b) the Co-occurrence view [10].

Two previous proposals to integrate set-valued attributes into histogram-oriented multidimensional visualization tools are shown in Figure 1. (For schematic comparison purposes, think of (b) as a 90° rotation of (a).) The Set'o'gram technique [3] decorates mostly equal-width histograms of value counts with information on the distribution of set sizes via width differences. This method would reveal whether a given value most often appears by itself or in combination with other values across sets of given sizes. The Co-occurrence view [10] splits value bars to reveal the co-occurrence of set values. Such a view would be able to reveal trade-offs as in "If I choose set value X, what other values come with it?"

Both these two proposals for visualizing set-valued attributes within histogram motifs are hampered by a demand for vertical visual real estate, negating one of the main strengths of multi-attribute visualization tools based on equal-height histograms, namely, that many attribute types can be compactly visualized in parallel in a clear and visually consistent manner. Therefore, we were motivated to invent another view for set-valued attributes that would take up no more vertical real estate than a normal equal-height histogram.

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

2. DISTRIBUTED SINGLETON SET VIEWS

The challenge in finding a visualization method using equal-height histograms is dealing with duplication of items. For regular histograms, of course, a set of items is partitioned into some number of bins based on values or value ranges. A given item will appear only once in one bin and the total number of items is constant. However, for set-valued attributes, a given item can appear multiple times across multiple bins if the bins are based on singleton set value instance counts. One might think that the power set of a set of values could be used to partition the items. However, the number of sets in a power set can be huge. (Think of the power set of authors, for example, in anything other than a trivially sized collection.)

We define a singleton set distribution (SSD) as follows: Let P be the range of values in some set-valued attribute function A. Let S be the set of singletons in P. For a set of items I as the domain of P, the singleton set distribution SSD is a function over S and the range of A that sums the count of each member of S in each subset of A(I). We also count the number of null sets N in A(I) as a special case.

For example, assume that an attribute function A takes its range from the power set of P = {a, b, c, d}. S = {{a}, {b}, {c}, {d}}. Let A(I) be the function as follows:

A(I1) = {a, b}

A(I2) = {a, b, c}

A(I3) = {b}

A(I4) = {a, c}

A(I5)) = {}

The SSD of the above case is shown in Table 1.

Table 1. Example SSD

Singleton Value	Frequency	Percentage
a	3	33%
b	3	33%
c	2	22%
d	0	0%
{}	1	11%

Visualizing the SSD in a way that would equalize the width of each attribute row requires a normalization of the count in the sum of the second column of the SSD. We simply sum the column and then compute the value that determines the width of each value cell as the percentage of the total count.

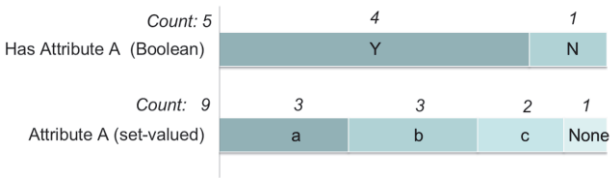

Figure 2. Standard stacked bar chart showing a Boolean and a set-valued attribute applying to the data in Table 1.

So far, the design seems straightforward. The data in Table 1 could be represented in a set-valued attribute row as shown in row 2 of Figure 2. But when paired with a non-set-valued attribute such as the Boolean attribute "Has Attribute A," i.e., Attribute A is non-null, shown in row 1 of Figure 2, we may have a problem.

As any user of business graphics software knows, one can graph a table of values with two different stacked bar chart types: (1) a standard bar chart with absolute counts (and typically unequal bar lengths) or (2) a bar chart showing percentages, where the bars will all have equal lengths. The example in Figure 2 shows a percentage-based bar chart but with a difference. The wrinkle is that the same set of objects is being shown in both bars but because of duplication, their total counts are different. There is a potential for a false implicature because of the strong tendency of the human visual system to interpret graphical length and position as an indication of quantity [1][2]. In Figure 2, it would appear from length and position that the quantity of items in the last cell of row 1 is greater than that in the last cell of row 2. But of course their counts are in fact the same.

One could use bars with absolute (not percentage) counts, but the problem is that the total count of occurrences of singleton set values is not particularly meaningful for any tasks that we can see. One would end up with bars of wildly varying lengths where the lengths didn't reveal much of anything. Valuable visual real estate would be wasted.

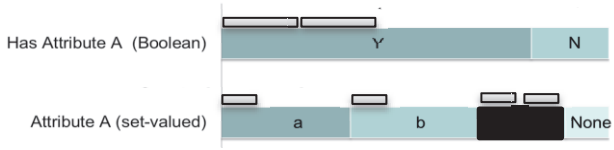

Figure 3. Selecting value "c" (shown in black) selects items that have value "c" but also have values "a" and "b".

A further complication of the design choice in Figure 2 arises when considering the selection behavior of an interactive system. In [11], selection of a value cell button selects the items that have that value, and an item row just above is highlighted accordingly. However, selection of a set-valued attribute cell selects not just the items with that attribute value alone, but may select others also. As shown in Figure 3, if a user selects value "c" then items are chosen that also have value "b" and "a." Such system behavior may be confusing, particularly if the set-valued attribute type is not made visually distinct from the other non-set-valued types.

Figure 4. Screen shot from BarExam for data in Table 1.

Thus we propose a design in which we draw the set-valued attributes with parallelograms rather than rectangles. We hypothesize that the perceptual tendency to associate horizontal position with quantitative information across rows will be weakened by introducing non-parallel lines. And we also hypothesize that the differing visual properties will help to mitigate confusion when users encounter the differing selection behavior. A screenshot from our system that illustrates this design for the data in Table 1 is shown in Figure 4.

3. USE CASE

Here we describe briefly a use case with parallel interactive bargrams that include set-valued attributes. This use case

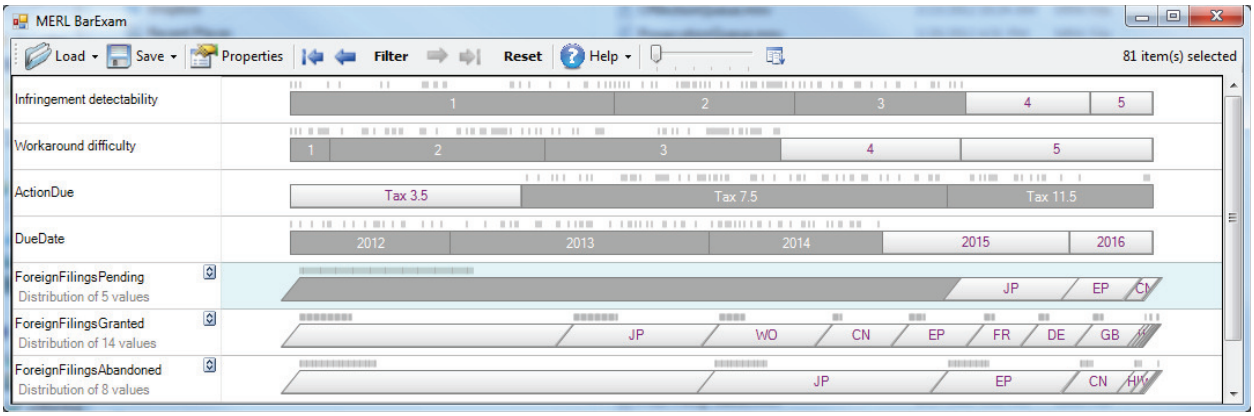

Figure 5. Use case involving reducing maintenance fees in the management of a patent portfolio. The user has already filtered a much larger set of patents to those that require maintenance. Now the user is exploring constraining the set further with the goal of ending up with a short list that can be examined in more detail.

occurs in deployment of BarExam at Mitsubishi Electric Research Laboratories (albeit with far more attributes). The task is to review the patent portfolio in order to reduce costs by abandoning the maintenance of some patents. Prior to the snapshot in Figure 5, the user has filtered a much larger set of patents in the portfolio to those that require further maintenance by selecting *ActionDue* values and hitting the *Filter* button. Out of these remaining 373 patents shown in Figure 5, the user explores how to limit the set further by finding a combination of values that will yield the maximum savings in the coming years while minimizing value loss to a global portfolio. The values of *ActionDue* determine cost—more mature patents cost more to maintain. Other attributes are indicators of value (*Infringement detectability, Workaround difficulty*). The decision to abandon US patents will also be affected by the status of foreign counterparts shown in the bottom three set-valued rows. The user explores abandoning US patents whose foreign counterparts are not actively being prosecuted in foreign countries, a selection shown in the *ForeignFilingsPending* row. After exploring many alternatives, the user ends up with a set of restrictions that represents a reasonable tradeoff for choosing a set of patents to consider abandoning that will next be examined in detail by a larger group of people in the organization.

4. EVALUATION

In this section we discuss an evaluation of our methods for set-valued attributes and also a more general usability study of BarExam. The studies were not expected to yield statistical significance, but rather produce usability and design feedback.

4.1 Participants

A total of 16 undergraduate and graduate participants at University Carlos III de Madrid were recruited for the evaluation. Their characteristics are shown in Table 2.

Table 2. Participants' characteristics

Age Range	18-22 (12.5%), 23-26 (50%), 27-34(37.5%)
Sex	75% Male, 25% Female
Education	Computer Science Bachelor (37.5%), Telecommunication Bachelor (12.5%), MSc Computer Science (25%), PhD Candidate (25%)
Use of computer as main activity	More than 5 years (75%), 1 to 3 years (25%)
Data visualization experiences (courses/seminars)	No (50%), 1 course/seminar (25%), more than three courses/seminars (25%)

4.2 Procedure

For our evaluation, we used a dataset of 200 items from a car models database with nineteen attributes, including model, price, length, horsepower, warranty years, color, and infotainment systems. Two of those were set-valued attributes. The attribute *color* could take up to three values and the attribute *infotainment system* could have zero or more options from the following: *MP3 Player*, *Navigation System*, *Bluetooth*, and *Traffic info*. Each participant session included two parts: (1) exploration of the design regarding parallelograms vs. rectangles for set-valued attributes and (2) using BarExam to accomplish typical tasks.

4.2.1 Parallelograms vs. Rectangles

We presented the subjects with static screen shots of our cars dataset with two alternative design choices for the set-valued attributes. The first used rectangles for set-valued attributes that were exactly like non-set-valued attributes (see Figures 2-3). The second used parallelograms as shown in Figures 4-5. We explained the difference between set-valued attributes and non-set-valued attributes including selection behavior discussed in Section 2. We then asked the participants to complete the answers for three questions:

Q1: Which of the two design variants is preferable and why? [Only rectangles, Rectangles and Parallelograms]

Q2: How much do you agree that parallelograms represent a valid alternative to rectangles to overcome the strong human perceptual tendency to interpret equal length bars as indicating equal quantitative values? [1-7 Likert scale]

Q3: Please list other alternative visualizations other than parallelograms we might have used to overcome the strong human perceptual tendency to interpret equal length bars as indicating equal quantitative values? [fill in]

4.2.2 Usability Study

The total time of this part of the study was approximately 50 minutes. After a 15 minute tutorial, we asked the participants to perform two tasks using BarExam.

T1: You want to buy a new car and you already know which are the characteristics you prefer. Your budget is around $30,000. Considering the size of your garage you need a car that is shorter than 180 inches. You require at least 4 years of warranty and an at least 15 miles/gallon in the city. Use the tool to select one car or at least to narrow the possibilities to 5 vehicles.

T2: You work in a car company that is currently designing a new full-size sport-utility vehicle. Your manager asked you to conduct an analysis to discover how many warranty years your competition offers for the same class of vehicles.

Afterwards, we asked the participants these questions:

Q4: Please rate how satisfied you are with this tool in helping you with your choice task given the information provided. [1-7 Likert scale]

Q5: How confident are you that you made the best choice given the information presented? [1-7 Likert scale]

Q6. How likely would you be to use the BarExam tool in the future if it were available for making choices among sets of the large sizes? [1-7 Likert scale]

4.3 Results and discussion

With respect to Q1 (rectangles vs. parallelograms for set-valued attributes), only one user out of sixteen indicated the first alternative as preferable, that is, set-valued and non-set-valued attributes both rendered with rectangles. The reason given was that "there is not much difference between parallelograms and rectangle shapes, so a small difference can be more confusing than useful to the user." The other users agreed that using different shapes might help users to understand that the attributes in play are of different type. On question Q2 (parallelograms as a choice to overcome perceptual tendencies), 69% of the participants answered positively, 19% answered negatively, and 12.5% neutrally. For Q3 (suggest alternative designs), seven users proposed the use of a progress bar inside the parallelograms as an alternative to item vectors above in the bars in case of set-valued attributes. These subjects felt this design would further distinguish set-valued from non-set-valued attributes.

Therefore, to the question *is our design supported,* we can answer positively. However, the support for parallelograms was not universal, and use of progress bars deserves further study.

Table 3 presents the results of the usability questions (Q4-Q6).

Table 3. Survey results for questions Q4-Q6

Likert scale	Q4	Q5	Q6
Strongly agree	43.75%	50.00%	25.00%
More than agree	31.25%	43.75%	62.50%
Agree	25.00%	6.25%	6.25%
Not sure	0.00%	0.00%	6.25%
Disagree	0.00%	0.00%	0.00%
More than disagree	0.00%	0.00%	0.00%
Strongly disagree	0.00%	0.00%	0.00%

The results of the usability part of our study were encouraging. Satisfaction (Q4) was high as was confidence in the result (Q5). Likelihood of future use (Q6) was also positive though not as high overall as Q4 and Q5.

The usability study captured a few other minor suggestions for improvement, particularly in labeling.

5. CONCLUSION

In this paper, we have introduced a novel method to visualize set-valued attributes in the context of a multi-attribute visualization system utilizing parallel equal-height histograms (bargrams) as the main visual motif. Our method is motivated by the desire to accommodate set-valued attributes into our general scheme without requiring as much vertical visual real estate as previous methods. However, to do so we have to overcome some strong tendencies in human visual perception and what may be perceived as inconsistent system behavior in selection actions. Our user study gave support for our design, but also yielded some alternatives we will investigate. In terms of general usability, the system was rated positively, but we would note that users were first given a tutorial in how to use it. In future work, we will continue to refine the designs discussed here but also seek to address other limitations of parallel bargrams. In particular, focus and scalability in the number of attributes is an ongoing issue, and we believe that other visualization methods related to temporal and spatial attributes should be investigated and integrated into systems like that presented here.

6. REFERENCES

[1] J. Bertin, The Semiology of Graphics, ESRI Press, 2010 (original publication in French 1967).

[2] W.S. Cleveland and R. McGill, Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Models, Journal of the American Statistical Association 79, 387, pp. 531-554, 1984.

[3] W. Freiler, K. Matkovic, H. Hauser, "Interactive Visual Analysis of Set-Typed Data," Visualization and Computer Graphics, IEEE Transactions on , vol.14, no.6, pp.1340-1347, Nov.-Dec. 2008.

[4] B. Lee, G. Smith, G. G. Robertson, M. Czerwinski, and D. S. Tan. Facetlens: exposing trends and relationships to support sensemaking within faceted datasets. In CHI, pages 1293–1302, 2009.

[5] Lindner, H-G. Knowledge Reporting with InfoZoom, SAP Design Guild, Innovation Second Edition, 12/22/ 2000, http://www.sapdesignguild.org.

[6] R. Spence, Information Visualization: Design for Interaction, Second Edition, Pearson/Prentice Hall, 2007.

[7] R. Spence and L. Tweedie, The Attribute Explorer: information synthesis via exploration, Interacting with Computers 11, pp. 137-146.

[8] M. Spenke, C. Beilken, and T. Berlage, "FOCUS: The interactive table for product comparison and selection," in Proc. UIST '96, pp. 41-50.

[9] Tweedie, L.A., Spence, R., Williams, D., and Bhogal, R. The Attribute Explorer. In Video Proceedings and Conference Companion of CHI 1994 (Boston MA, USA, April 1994), ACM Press, 435-436.

[10] K. Wittenburg, "Setting the bar for set-valued attributes", in Proc. AVI, 2010, pp.253-256.

[11] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton, "Parallel Bargrams for Consumer-based Information Exploration and Choice," in Proc. UIST '01, pp. 51-60.