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

Singleton Set Distribution Views for Set-Valued Attribute Visualization

Wittenburg, K.; Pekhteryev, G.

TR2011-073   October 2011

Abstract

Visualization of set-valued attributes in multi-dimensional information visualization systems remains a relatively unexplored problem. Here we introduce a novel method for visualization set-valued attributes that we call the singleton set distribution view and integrate it into an interactive multi-dimensional attribute visualization tool utilizing bar-grams (aka equal-height histograms) as its main visual motif.

IEEE Conference on Information Visualization (INFOVIS)

Singleton Set Distribution Views for Set-Valued Attribute Visualization

Kent Wittenburg, *Member, IEEE*, and Georgiy Pekhteryev

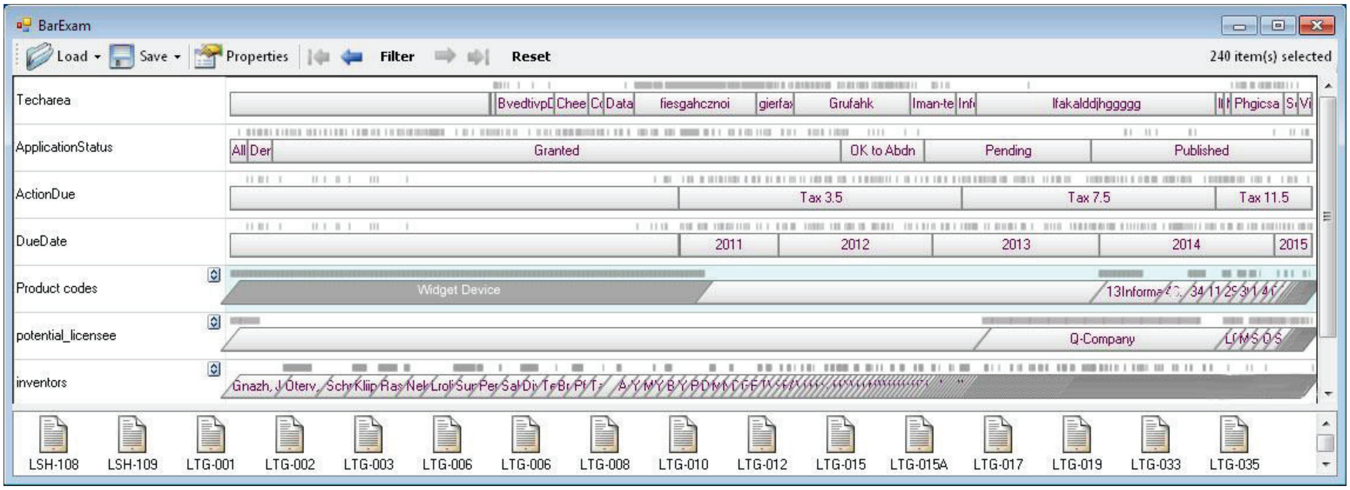

Fig. 1. Singleton set-value distribution views (shown as parallelograms) are shown here for a fictional corporate patent database in a parallel attribute visualization tool called BarExam. Other non-set-valued attributes are shown as rectilinear bargrams. The product code "Widget Device" value is chosen, and the dark lines above the rows indicate the 240 items (inventions) thus selected. Note that there are items with multiple product codes selected in this case, reflecting a typical property of set-valued attributes.

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 bargrams (aka equal-height histograms) as its main visual motif.

Index Terms—Set-valued attributes, multi-dimensional visualization, bargrams, equal-height histograms, patent visualization.

1 Introduction

Set-valued attributes are a frequent, naturally-occurring data type in a wide variety of domains. For instance, within corporate patent portfolios, many attributes of an invention are naturally represented as set-valued. Values for these attributes naturally take zero, one, or more members from a larger set. For example,

- Inventors
- Patent applications (US and foreign)
- Related patents
- Products currently utilizing the invention
- Products potentially utilizing the invention
- Licensees
- Potential licensees

Visualizing a corporate patent portfolio as a whole is valuable for a number of tasks related to portfolio management. For example, culling the portfolio in order to control maintenance costs; mining the portfolio for patents related to cross-licensing negotiations, litigation, or licensing opportunities; valuing the portfolio; and establishing areas of strength for strategies for future patenting. Visualization is also useful for maintaining data integrity to the extent that it reveals data anomalies, inconsistencies, and incompleteness. Visual analytics methods further complement visualization methods alone for portfolio management.

Compared to work on numerically-valued attributes, the related InfoVis literature on set-valued attributes is relatively sparse. The study in [1] provides a review of related work and proposes a method for visualizing set-valued attributes called the set'o'gram that decorates equal-width histograms of value counts with information related to the distribution of set sizes. This method would reveal whether a given value most often appears by itself or in combination with other values across sets of given sizes. In [2] another set-valued attribute view was proposed that was designed to reveal the co-occurrence strength of specific 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?" A suitable task for this type of view would be selecting one item from many in, say, a car-buying context.

In our experience applying set-valued attribute visualization to patent portfolio management, we found a need to go beyond both these previously proposed methods. Both the set'o'gram and the co-occurrence view have the disadvantage that they can take up

- *Kent Wittenburg is with Mitsubishi Electric Research Laboratories and University of Madrid King Carlos III. E-Mail: wittenburg@merl.com.*
- *Georgiy Pekhteryev is an independent contractor formerly with Mitsubishi Electric Research Laboratories., E-Mail:gpekht@gmail.com.*

Manuscript received 27 June 2011; accepted 1 August 2011; posted online 23 October 2011; mailed on 14 October 2011.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

extensive vertical visual real estate compared to the typical bargram (equal-height histogram) in multi-attribute visualization tools [3][4]. For set'o'grams, vertical real estate is necessary to show equal-width histograms that would normally be more parsimoniously shown as equal-height histograms. In the case of co-occurrence views, an attribute that normally would take up only one row in a matrix of attribute rows would require as many rows as there were occurring set values. This need for vertical real estate defeats one of the main strengths of parallel attribute visualization tools utilizing bargrams, namely, that many attribute types can be visualized simultaneously 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 bargram.

2 Distributed Singleton Set Views

The challenge in finding a visualization method for set-valued attributes that uses no more visual real estate than a normal bargram is that the total count for the attribute row as a whole might vary. 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, for example, the bins are based on singleton set values. 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 inventors, for example, in a typical corporate setting).

Figure 1 shows an example of our proposal for singleton set distribution views in the context of a parallel bargram visualization tool we call BarExam. The singleton set distribution (SSD) is defined 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:

$A(I_1)$ = {a, b}
$A(I_2)$ = {a, b, c}
$A(I_3)$ = {b}
$A(I_4)$ = {a, c}
$A(I_5)$) = {}

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 (to be consistent with parallel bargram layouts) 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, as shown in the third column. The actual drawing function will determine the width of a given attribute row based on the total item count in the data set being visualized as well as the available horizontal visual real estate, and the SSD views will be normalized to that same width.

If we made no other changes to the visualization method, we would introduce a false visual implicature related to Bertin's observations about positional encoding of quantitative information [5]. That is, we will perceive vertically aligned positions as indicative of equal quantitative information, in this case, the count of items. It would be implied that the total count of items was the same across all types of attributes and also that any subcounts of value bins would be comparable. To illustrate, consider Figure 2, in which we have introduced a Boolean attribute "Has_A" along with our set-valued attribute A from the previous example. If we rendered the set-valued attribute A as usual, it would appear that the number of items with null values in A was less that the number of items marked "N" in the Boolean "Has_A" attribute. But of course this is false. The upper bars (referred to as item vectors) for set-valued attributes is not a count of items but a sum of frequencies of occurrences of singleton set values, as shown in Table 1.

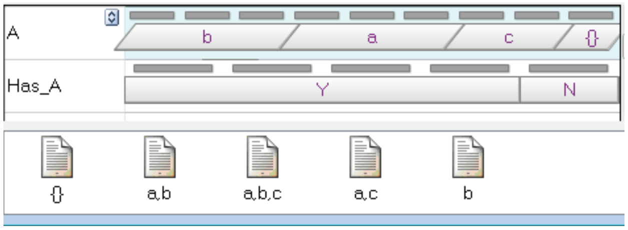

Fig 2. Example showing non-comparable counts.

In order to counteract the potential false implicature, we draw the singleton set views with slanted rather than vertical lines. We hypothesize that the perceptual tendency to associate horizontal position with quantitative information across rows will be weakened or defeated by introducing non-parallel lines.

Our Singleton Set View method has the merits that it is relatively parsimonious in its visual real estate demands and yet it reveals information useful to forming mental models of a multidimensional dataset. In the view in Figure 1, for instance, one can infer that this portfolio is relatively strong in protecting products in the "Widget Device" domain compared to the other product types of the company. One can see that there are a few inventors who stand out as the most prolific. And the strongest potential for licensing is revealed to be the company labelled "Q-Company." A limitation of our method is that it requires interaction to uncover correlations across attributes, an issue common to basic equal-height histogram approaches. Scalability in the number of attributes is also a problem.

3 Conclusion

We have introduced a novel method to visualize set-valued attributes in the context of a parallel attribute visualization system utilizing bargrams as the main visual motif. Our method appears to yield new insights into set-valued data without requiring as much visual real estate as previous methods. We have given examples from the domain of corporate patent portfolios. Whether the potential false visual implicature that we discussed is actually overcome by our presentation methods deserves future study.

References

[1] 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.

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

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

[4] 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.

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