Parallel Sets: Visual Analysis of Categorical Data

Fabian Bendix[*]
VRVis Research Center

Robert Kosara[†]
Inte:Ligand

Helwig Hauser[‡]
VRVis Research Center

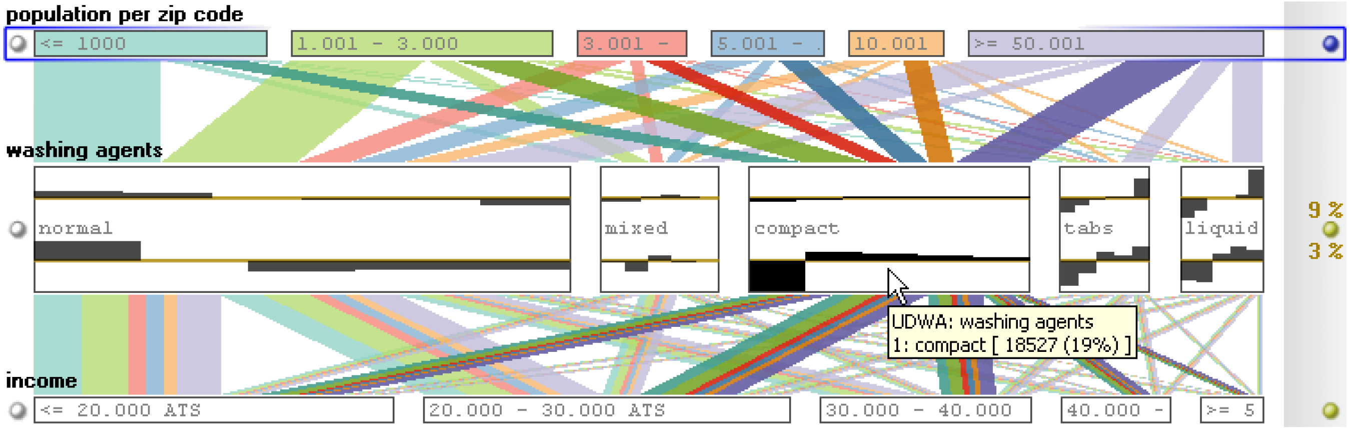


Figure 1: A visualization of a CRM data set which contains 93.872 data records; three categorical dimensions are displayed as Parallel Sets. Complex relations within the data can be revealed using interactive dimension composition, interactive subset highlighting, and the integration of histograms.

Abstract

The discrete nature of categorical data makes it a particular challenge for visualization. Methods that work very well for continuous data are often hardly usable with categorical dimensions. Only few methods deal properly with such data, mostly because of the discrete nature of categorical data, which does not translate well into the continuous domains of space and color.

Parallel Sets is a new visualization method that adopts the layout of parallel coordinates, but substitutes the individual data points by a frequency-based representation. This abstracted view, combined with a set of carefully designed interactions, supports visual data analysis of large and complex data sets. The technique allows efficient work with meta data, which is particularly important when dealing with categorical datasets. By creating new dimensions from existing ones, for example, the user can filter the data according to his or her current needs.

We also present the results from an interactive analysis of CRM data using Parallel Sets. We demonstrate how the flexible layout eases the process of knowledge crystallization, especially when combined with a sophisticated interaction scheme.

CR Categories: I.3.m [Computer Graphics]: Miscellaneous—; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information Filtering

Keywords: categorical data, meta information, interaction

1 Introduction

In the last decades, the technological evolution has led to a constantly increasing amount of digital information being stored. Also, the awareness has grown that the data potentially contains useful and valuable information.

As a consequence, the research field of *knowledge discovery in databases (KDD)*, that started in the late 1980s, has gained much interest. The computer-aided analysis of data can be roughly divided into: (1) data mining that provides an automated approach to gain insight into the data [8] and (2) interactive visual analysis (or information visualization), in which the exploration of the data is *user-driven*, i.e., the user's domain knowledge is involved in the process of gaining new information.

Information visualization (InfoVis) techniques can be classified according to the data they are capable of displaying. One interesting data domain is the domain of heterogeneous and high-dimensional data that can be found in market research, customer relationship management (CRM), surveys, census, etc.

Usually, there are many dimensions, which all have different qualities. The distinction between different data types, for example, plays an important role in visualization (e.g., the distinction between ratio, interval, ordinal, and nominal data [2]). Especially challenging for visualization is the class of *categorical* data.

A categorical variable is characterized by the facts that (1) it is discrete, (2) there is a usually small number of different values (classes) that define the variable, and (3) there is often no implicit relation among these classes (whereas it is for numerical variables in terms of ordering and distances among values).

Traditional InfoVis techniques, like scatterplots and parallel coordinates (Figure 2a,b), best work for continuous data variables, because (1) there is a natural one-to-one mapping of data values to visualization parameters like positions and colors and (2) these continuous parameters better match the continuous characteristics of the screen (in the spatial, temporal, and chromatic dimensions). In this paper we present a new approach to information visualiza-

[*] e-mail: Bendix@VRVis.at
[†] e-mail: Kosara@inteligand.com
[‡] e-mail: Hauser@VRVis.at

IEEE Symposium on Information Visualization 2005
October 23-25, Minneapolis, MN, USA
0-7803-9464-X/05/$20.00 ©2005 IEEE.

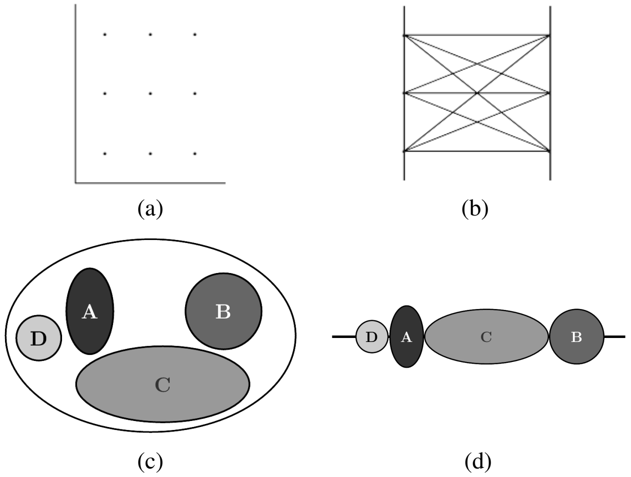


Figure 2: Top: the viewer has no idea how many data records are visualized; the illustration shows the problem that can arise if two categorical variables (three categories each) are displayed by a traditional scatterplot (a), or by parallel coordinates (b). Bottom: a Venn diagram provides a better way of displaying categorical values (c) and (d).

tion, called *Parallel Sets*, which is optimized for categorical data.

The main contribution of our work is a new technique for interactive visual analysis that combines the advantages of two well-proven InfoVis techniques. Parallel Sets combine the flexible layout of parallel coordinates, i.e., treating all displayed dimensions visually independent from each other (in contrast to recursive space-subdivision approaches like the Mosaic Display), with the idea of displaying frequencies as representatives for the categories (in contrast to the one-by-one items-based visualization of data).

With Parallel Sets, the dimensions are displayed side by side, the frequency-based representation of categories and relations reveals even complex information about the data, and meta information is provided (that is used to store additional information about the data [7]).

The following sections present the related work and the idea of Parallel Sets in terms of the visual metaphor and of the inherent interaction schemes. The workflow is explained to gain interactive visual analysis of heterogeneous and high-dimensional data. It is important to stress the fact that this interaction scheme is an integral part of Parallel Sets which is necessary to make use of this approach. Finally, we demonstrate the use of Parallel Sets to reveal interesting information in a CRM dataset.

2 Related Work

Parallel coordinates [12] are a visualization technique, in which the axes are not arranged orthogonally, but they are placed side by side. An n-dimensional data point is represented by a polyline, which intersects the parallel axes at points which represent the values of the individual data dimensions. This view is capable of displaying high-dimensional data, because the axes are visually independent from each other (compare Figures 3a and c).

Initially, parallel coordinates were designed to display continuous variables [14], but recent approaches have tried to integrate categorical variables into this visualization as well. Rosario et al. [17] suggest transforming categories to numbers by techniques similar to Multiple Correspondence Analysis (MCA). By this, the space on each axis is used more efficiently, because the spacing becomes meaningful (similar categories are positioned close to each other).

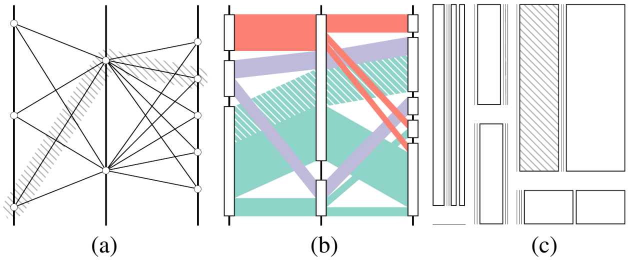


Figure 3: Three different visualization techniques displaying the same data: (a) the categories are represented by points on continuous axes in parallel coordinates, (b) Parallel Sets show the frequencies of categories and relations, and (c) a Mosaic Display provides a compressed view of the data (the hatched parts represent the same subset).

A simpler approach is proposed by Soon et al. [21]: for each category, an interval is constructed on the continuous axes to make more polylines visible. By this, the space is used to give the user an impression of how many data items are visualized. One problem remains for all parallel coordinates techniques: the visualization implements a continuous design model, which does not match the discrete user model of the data. This discrepancy of user imagination and presented image is eliminated by the use of frequency-based techniques: categories are represented by visual entities that are scaled according to their corresponding frequency.

There are several techniques that follow this approach: the Mosaic Display [11, 9] is a recursive space-subdivision technique (similar to Dimensional Stacking [16]), in which the frequency values of categories are represented by particular areas (“tiles”) on the screen – interactive mosaic plots [22, 13] provide an even better approach for visual exploration, because they make use of the user’s domain knowledge; Bargrams [23] and InfoZoom [20] are techniques that display the dimensions row by row and the categories are mapped to boxes, whose widths are scaled according to their frequency. The drawbacks of these frequency-based techniques are: (1) space-subdivision methods introduce a ranking of the displayed dimensions and are limited in the amount of dimensions that can be displayed, and (2) for the latter kind of visualizations, the relationships between dimensions are not shown explicitly, but the vertical alignment encodes the relation of different dimensions’s categories, which sometimes makes the view difficult to understand when investigating multi-dimensional relations within the data.

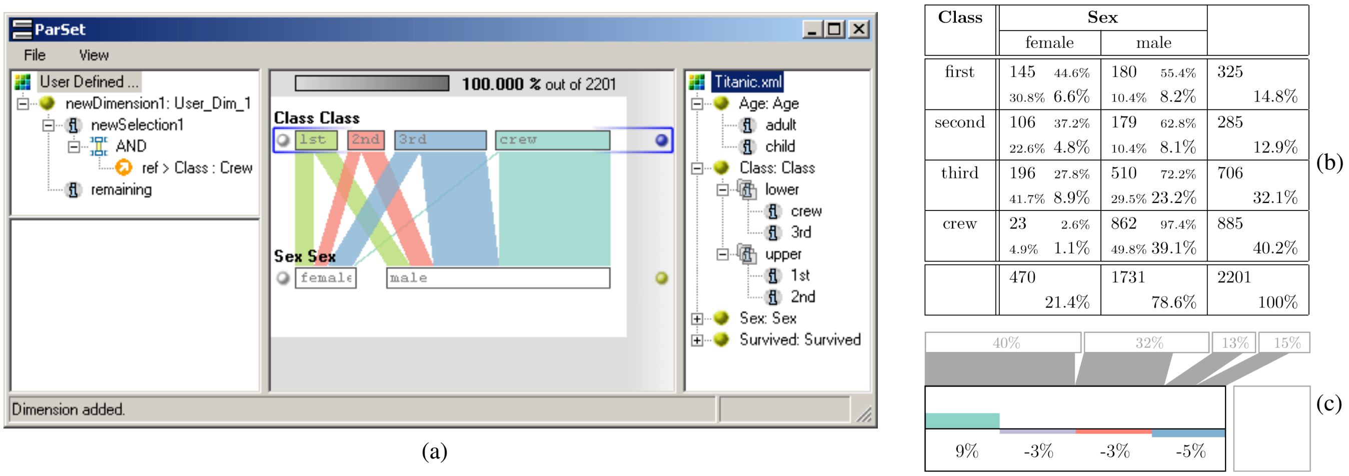
The Parallel Sets technique combines the advantages of frequency-based techniques (implementing a discrete design model and displaying the frequencies of categories) and parallel coordinates (treating dimensions independently).

3 Parallel Sets

This section presents our approach to serve the purpose of interactive visual analysis: (1) Parallel Sets is both a new visual metaphor and an interaction framework, (2) the visual metaphor serves as a natural way of mapping categorical variables to visual entities, (3) interaction options are a key issue to make exploration possible, and finally, (4) some additional features are introduced.

3.1 Basic Idea

The basic idea of Parallel Sets consists of two parts: (1) the visual metaphor that properly deals with categorical dimensions and (2) an interaction concept that facilitates the exploration of the data as well as the creation of new information about the data.



Class	**Sex**		
	female	male	
first	145 44.6% 30.8% 6.6%	180 55.4% 10.4% 8.2%	325 14.8%
second	106 37.2% 22.6% 4.8%	179 62.8% 10.4% 8.1%	285 12.9%
third	196 27.8% 41.7% 8.9%	510 72.2% 29.5% 23.2%	706 32.1%
crew	23 2.6% 4.9% 1.1%	862 97.4% 49.8% 39.1%	885 40.2%
	470 21.4%	1731 78.6%	2201 100%

Figure 4: (a): our Parallel Sets prototype (with the *titanic* data set [3] loaded) consists of four panels (from left to right): the user panel (showing the dimensions the user has created), the exclusion panel (for filtering), the visualization panel, and the data panel (showing the source data); (b): the crosstabulation for the displayed dimensions, which builds the basis for the visual mapping transformation; (c): histograms show the deviation of marginal frequencies to relative frequencies, visualizing the degree of independence of the category to all neighboring categories.

Concerning the visual metaphor, Figure 3 relates our approach to parallel coordinates and to the Mosaic Display. As mentioned in Section 2, these approaches do not optimally deal with categorical data: either the frequency information is not visible or a ranking is imposed on the visual mapping transformation [6], influencing the human perception.

Parallel Sets share the layout with parallel coordinates, but the continuous axes are replaced with sets of boxes that represent the categories. These boxes are scaled according to the frequency of the corresponding category (see Section 3.2) and are initially ordered according to the meta information (see Section 3.3). Using the frequency information means to utilize an aggregation [15] of a large categorical data set, reducing the amount of data to be displayed without information loss.

Because these sets of categories are placed independently side by side, the connections between categories (representing the associated attribute combinations) are also scaled according to their frequency values. Two features characterize Parallel Sets especially well. Firstly, the visualization is not restricted to categorical variables. By means of binning or clustering, a continuous variable can be easily transformed to fit into this kind of visualization. So, the display is not limited by the amount of data records (as for instance parallel coordinates are), but by the amount of categories of the displayed dimensions (usually a small or at least smaller number).

The second speciality of Parallel Sets is its support of interactive visual analysis. As already mentioned, a categorical dimension is a classification of the data. However, it also is only one possible classification out of many. Hence, it is useful to give the user the possibility to create a new classification which implies that he or she can build new meta information for the data. The key word here is *user-driven*: the user utilizes his or her domain knowledge to enrich the meta information about the data; this new information can consequently be used for further exploration.

Since exploration is a cyclic process (knowledge crystallization [6]), it is useful to store user-defined dimensions together with other meta information. As shown in Figure 4a, a tree view (on the right) provides access to the data itself: the dimension names, the associated category names, and hierarchy information (among the dimensions *and* categories); another tree view (on the left) represents the user-created dimensions, which are built from logical combinations of user selections (see Section 3.3).

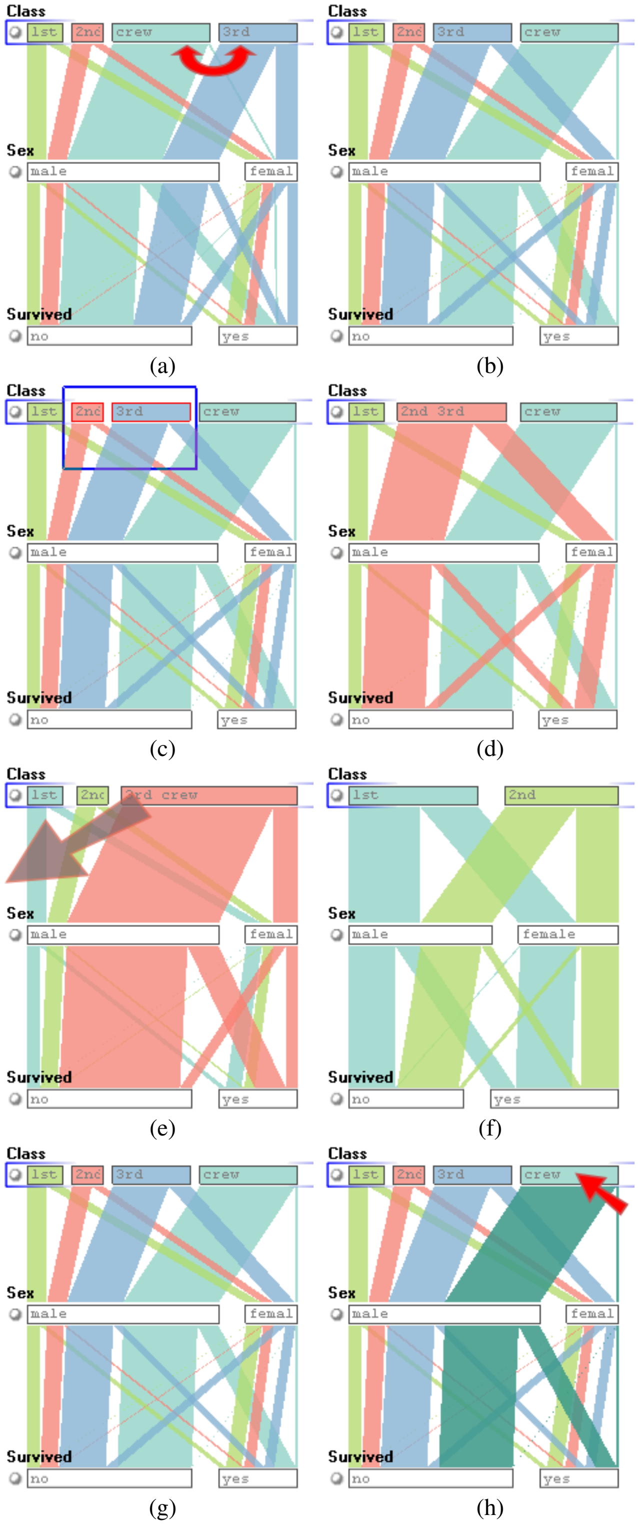
3.2 Visual Metaphor

The information that is provided by the visualization is obtained by a crosstabulation [2]. Statistical examinations deal with categorical data quite frequently and usually there is a look at frequency (contingency) tables first to get a quick overview. Figure 4b shows an example of a two-way table: what is displayed by the visualization is the information obtained by multi-way tables.

In each cell of the crosstabulation, the top left values show the occurrences f_{ij}, the bottom right numbers show the absolute frequencies $p_{ij} = f_{ij}/f_{++}$ (where $f_{++} = \sum\sum f_{ij}$), and the remaining two show the individual row frequencies $r_{ij} = f_{ij}/f_{i+}$ and column frequencies $c_{ij} = f_{ij}/f_{+j}$ (where f_{i+} and f_{+j} are the marginal row and column frequencies, respectively). Figure 4a shows our Parallel Sets prototype visualizing this data. The crosstabulation, which is calculated for each attribute combination of the displayed dimensions, builds the basis for the visual metaphor: each category is scaled according to the corresponding marginal frequency p_{i+} and p_{+j} respectively, and the connection between each two categories is scaled according to the absolute frequency p_{ij}. The visualization of actual data records is exchanged with that of frequency information, which still gives the user insight into the distribution of the data records.

The screenshot in Figure 4a shows the visual mapping of such a crosstabulation. The visualization for two dimensions is illustrated in a comprehensible way and color is used to distinguish between different relations. At any point in time, there is one selected dimension, the *active* dimension, which defines the color-coding of the connections. Each category of the active dimension gets one color (predefined, well-spread, iso-luminant colors are used to differentiate the connections [1]) and all connections obtain the color of the respective active category. Then a visual ordering of the displayed dimension is introduced: starting at the active dimension, neighboring dimensions split the connections into sub-connections according to their number of categories. This is analogous to imagining a subset with a particular attribute (e.g., first-class passengers) and subdividing it according to a second feature (e.g., gender), then a third feature, and so on.

In this flexible display only the absolute frequencies are visualized. Hence, there is room to offer more information: the user can vertically resize the boxes (representing the categories) and inside



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5: Because there need not to be a natural order among categorical values, each dimension's categories can be reorder interactively (a,b); selected categories can be grouped to facilitate a hierarchical organization (c,d); filtering is implemented by hiding particular categories (e,f); highlighting can be used to emphasize certain connections visually, if the user moves the mouse over a category (g,h).

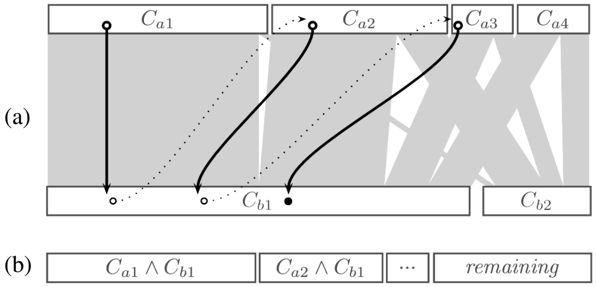


Figure 6: An example of dimension composition: the user is interested in a classification into the following four categories: $C_{a1} \wedge C_{b1}$, $C_{a2} \wedge C_{b1}$, $C_{a3} \wedge C_{b1}$, and *remaining*. The displayed interaction path in (a) illustrates the sequence of selections (the dotted lines indicate that the user finishes the current brush and starts a new category); the resulting user-defined classification is displayed in (b).

this additional space histograms can provide a more detailed view of the data. Aside from the absolute frequencies, the individual row and column frequencies of the contingency table (Figure 4b) can be integrated into the visualization by the use of histograms [12] for the selected dimensions. In statistical terms, these relative frequencies are conditional probabilities. Because comparing conditional probabilities can be misleading (similar to Simpson's paradox [4]), the relative frequencies have to be standardized. One way is displaying the deviations of conditional probabilities and the a-posteriori probabilities ($\Delta P_i = P(A_i \mid B_j) - P(A_i)$). If the deviation is zero, then the particular category (with associated probability B_j) is independent to all categories of the neighboring dimension. Figure 4c shows an example of dependent relations: one can see the absolute distribution of the upper dimension and additionally, how the particular frequencies change if only data records of the lower left categorical attribute are considered. For instance, the positive difference (9%) means that data records of the associated category are more frequent in the considered category (49%) than in the absolute distribution (40%).

3.3 Interaction and Workflow

There seem to be quite different opinions of what interactivity means [13] (e.g., concerning the temporality of change). A compact definition is given by Becker et al. [5]: *direct manipulation and instantaneous change*. Thus, in this section, features of Parallel Sets are presented that are triggered by human-computer interaction and result in immediate visual feedback: the use of selection and highlighting, interactive querying, filtering and reordering of dimensions as well as categories. The interaction scheme also implements Shneiderman's visual information seeking mantra *overview first, zoom and filter, then details-on-demand* [19] (Figure 5).

The interactive data exploration starts with an undirected investigation of the available data variables. The user chooses interesting data variables, adds them to the visualization panel, and explores their relationships. During this step, it is necessary that the user can factor more practical views, because the visualization can easily become very complex if the number of displayed categories increases to more than just a few. New classifications for data variables often help to structure the data. Later, the user wants more detailed information about the relation between two or multiple dimensions.

This high-level view of interactive visual analysis is implemented by Parallel Sets. The investigation starts with choosing interesting data variables: the data panel offers the data dimensions and the user panel offers the user-defined dimensions. The user can drag dimensions from both panels, drop them in the visualization panel, and create his or her own view of the data. The dynamic layout permits the reordering of dimensions with immediate visual feedback which is useful to look at the relationship of different dimensions more closely. Also, the categories can be reordered along the respective axis, as there not necessarily is a natural ordering among categorical values (Figure 5a,b). Having added interesting dimensions to the visualization (overview), the user can group selected categories together (zoom and filter: Figure 5c,d), by which he or she can organize categories hierarchically. The user can also drag uninteresting categories into the exclusion panel to filter the data (Figure 5e,f). Thereby, the entire screen space can be used by the remaining categories more effectively.

One fundamental idea behind the design of Parallel Sets is dimension composition. The use of this feature is (1) reduce the dimensionality of the visualization – both screen space and human perception limit the maximal dimensionality of the visualization – and (2) to build more practical categorizations (Figure 7a). In contrast to *data-driven* approaches (like PCA [10] or VHDR [24]), dimension composition enables the integration of the user's domain knowledge. A categorical dimension is a classification of all data records according to a particular data attribute (e.g., regarding the attribute age, a binning could classify the data into ten years intervals). In general, the data can be classified according to multiple aspects of the data. Hence, during the exploration process, it is useful to enable the user to build his or her own classifications of the data and to also *reuse* this information during further exploration and analysis. Figure 6 gives an example of the process: a new classification is created by selection activities. The path illustrates the sequence of selections; firstly, the category C_{a1} is selected, then the category C_{b1}, and so on. These selections are recorded by the user panel: for the first selection, a new dimension, an *active category* (equal to the selected category), and a *default category* (which contains all the remaining data items) are created. All successive selections are added to the current active category (by default, all selected categories are combined by a conjunction). In the example, after every two successive selections, the user indicates to start a new category (not visible). The result of the process is a new categorical dimension with four categories that represents a new classification of the data. This dimension (the user's domain knowledge) can be dragged into the visualization again and the user can continue working with just this one dimension, because it contains all the information the user considers to be relevant. Generally, two concepts are utilized: the new categorization can either contain all possible attribute combinations (specialization), or contain a subset of these combinations (generalization).

The final step is to have a closer look at interesting relationships and to get detailed information. Details actually are filtered data records that are the output of the visual analysis. Usually, once the user has found out some interesting relations within the data, he or she wants to get back to the original data items and to see all the details, e.g., in a standard table view. Concerning the investigation of relationships, Parallel Sets offers two schemes: histograms and highlighting. Histograms show statistical parameters to analyze relations in detail, highlighting is realized as mouse-over effect, i.e., all connections that *pass through* the box under the cursor (i.e., relations that feature the corresponding data attribute) are emphasized by drawing them more colorful and in front of all other connections. By this, the dimension can be applied to all other displayed dimensions, multi-dimensional relationships become visible, and interactive multi-dimensional exploration is enabled.

3.4 Additional Optimizations

This section summarizes further features of Parallel Sets that are implemented in addition to the base features as presented in the previous section to improve and to reflect different aspects.

One important feature of Parallel Sets is the treatment of continuous data dimensions since usually not all of the data dimensions are categorical. Continuous variables are binned to fit into the visualization scheme of Parallel Sets and the number of bins can be adjusted individually for each binned dimension. For the visualization of the relations, triangles are used instead of parallelograms to indicate the different nature of the data.

During a typical data exploration, multiple dimensions are visualized concurrently. With an increasing number of displayed categories, the view becomes very complex. Thus, a secondary mode for drawing connections is provided (Figure 8): all connections between each two categories are displayed in parallel, which results in a tidier visualization. The number of intersections remains the same, but the connections are bundled and the relations become clearer for better perception. The tidier visualization is achieved at the price of losing the nexus of the connections between more than two dimensions, because the horizontal alignment is lost in this mode.

Also, it can happen that the connections between categories become so oblique that it is difficult to visually compare the represented frequencies. If this is the case, histograms can help, because they provide a very comprehensible visualization for frequency data that facilitates better comparison. In addition to the mode explained in Section 3.2, relative frequencies can be displayed by this auxiliary plot: like in traditional histograms, the bars directly represent the frequency information.

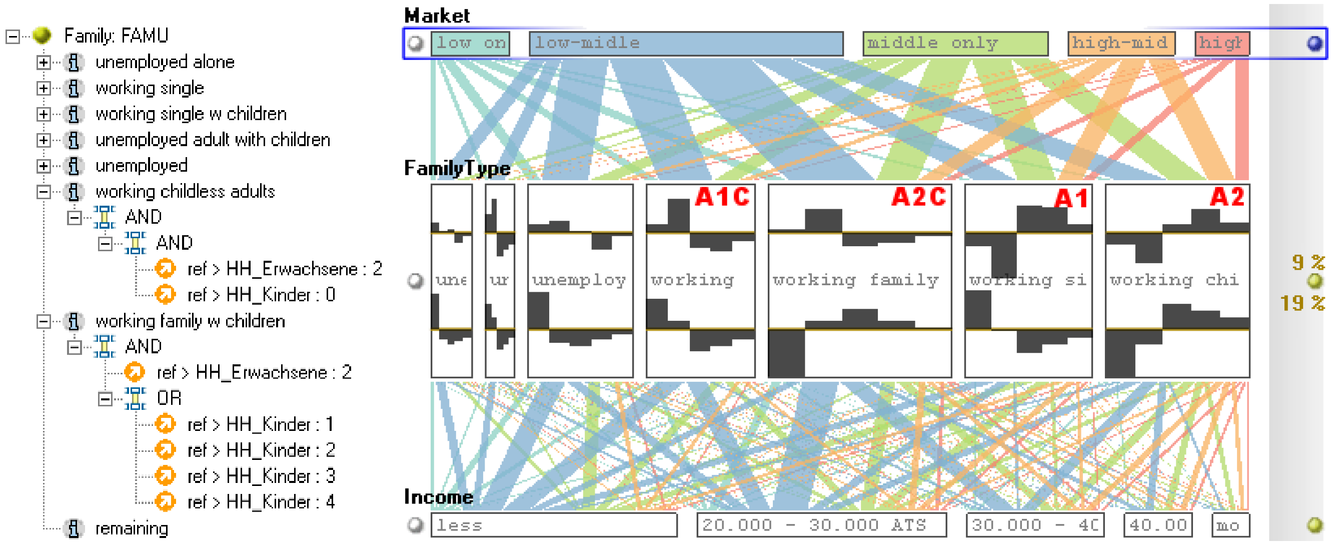
4 Case Study

In collaboration with our CRM partner, we have analyzed a questionnaire data set consisting of 99 dimensions and containing information about 93.872 households under investigation. The following portrays how Parallel Sets support the user in gaining important insight into the data.

The data contains information about people's living standards, shopping habits, pet care, and so on. It is a data set, which exhibits some of the challenging features of categorical values: (1) many data dimensions are given in binary form (e.g., yes or no, true or false), (2) the data has a hierarchical structure – that means, questions are grouped together according to particular topics, and (3) it is quite frequent that the questions are not answered, so most data variables have one category called *unknown* (or similar) – it could be interesting to analyze the data records in such categories, to find out whether there is a special reason why this information is not given (e.g., is it possible to specify a particular group of people that all have kept back their *real* data). The interactive visual analysis deals with the following tasks (questions):

- Extracting information about the households and bringing the data in a form the user can best work with: is it possible to categorize the households meaningfully?
- Are there meaningful relationships among the household types and other data dimensions that could give a clearer view of the households?
- Is it possible to make statements about people who buy a particular washing agent?
- Can we draw conclusions from the customer behavior to concrete attributes of these customers.
- Is it possible to make statements about people who buy a particular cat food brand?

There are two basic tasks that the user performs: undirected exploration and directed visual analysis. The first task is to bring the available data into a form so that the user can best work with it. In



(a) (b)

Figure 7: (a): the meta information for the user-defined dimension *household type*; the categories are build out of available dimensions (e.g., *working childless adults* is a logical combination of the categories *two* (number of adults) and *unknown or none* (number of children); because the meta information is processed top-down, all unemployed persons are already classified by the above category. (b) the type of household is shown in relation with the income and the user-defined favorite supermarket dimension: concerning income, the classes representing households with two adults (*A2C* and *A2*) are more likely to have a higher income than the corresponding classes with one adult; concerning favorite markets, the behavior depends on whether there are children living in the household or not (*HH_Kinder* is the identifier of the dimension that offers the number of children, and *HH_Erwachsene* provides the number of adults in a household).

this phase, interactive assembling of the visualization is an important tool for the user to get an overview of the available variables and of the relationships between the dimensions. The dimensions and categories can be reordered and as a final step, dimension composition can be used to process the dimensions to reflect the needed information. For instance, the data set contains information about the number of adults in the household (*unknown*, *one*, and *two*), the number of children (*unknown or none*, *one*, *two*, *three*, and *more*), and employment (*unknown*, *unemployed*, *half-day*, and *full-time*). This information characterizes the household and it is preferable for the user to have one dimension (e.g., called *household type*) that classifies the data records in the needed classes (instead of working with three dimensions). With the use of dimension composition, the user can define these categories successively. Figure 7a shows the meta information (displayed in the user panel) for the new classification: because the data records are classified top-down, the two expanded categories need not contain a logical combination concerning employment. For the subsequent analysis, it is sufficient to deal with these categories in terms of household types.

The goal of the user is to collect information about the households. Figure 7b shows one relation during this exploration. The relationship between household types, people's income and people's favorite supermarkets. The histograms visualize the frequency distribution of the market and income classes relative to the household types. Because similar histograms mean similar dependencies between the particular household types and the neighboring dimension, the top histograms show that households with children are *equally* distributed compared to the types without children concerning their favorite supermarkets (similar top histograms for the categories labeled *A1C* and *A2C*). The histogram distribution reveals that people living in children households are more likely to buy their goods in low and middle class supermarkets in contrast to non-children households. It turned out that this *affinity* is not only true for favorite supermarkets, but also concerning the living place: children households can be found more frequent in the countryside, non-children households are more frequent in larger cities. Thus, such a relation would look quite similar to the distribution of the top histograms of Figure 7b. The children households are equally distributed no matter it is a single household or a family household. The relation to the income is inverse: the histograms show that if two adults are living in a household the income is higher (related to single households) regardless of the number of children (similar histograms for the categories *A2C* an *A2*). It should be mentioned that the market classification is also a user-created dimension, because originally the data contains only the information whether a particular supermarket is the favorite market. The dataset contains sixteen dimensions – one for each supermarket with two categories each: *yes* or *no*. This kind of data variables is typical for CRM datasets: such questions (data variables) are optimized for the people who answer the questions, and not for those who analyze them. In this case, these sixteen questions are reduced into one dimension with five categories, because this generalization is sufficient for the user. The user can apply his or her own (or a well-known) classification of what market belongs to which class.

The user's goal is to link this knowledge about the households with concrete statements. With respect to customer relationships management (CRM), it would be interesting if there is a relation between household types and types of washing agents people buying preferentially. The different washing agents are again given in binary form (if people's favorite washing agent is of powdery, liquid, compact or normal type). Figure 8 shows the relations between household types, washing agent types, living place, and income. The absolute frequencies are represented by the connections that are displayed using the *bundled* mode (Section 3.4). Concerning the washing agent types, the histograms for the categories *tabs* and *liquid* are very similar and state: both types are more frequently bought by non-children households and by people that live in larger cities (similar distributions of the top and bottom histogram for the left and right type of washing agent). Also, the latter discovery is correlated with the fact of having a higher income. For the right income classes, the histograms are monotonically increasing, which means that the frequencies for right categories of the *population* dimension are higher than the displayed proportions. To summarize, one can state that the overall number of people that buy liquid washing agents, or tabs respectively, are more likely to live in urban regions and have a higher income compare to others living in the countryside. This statement holds because of the degree of independence; the deviation of conditional probability and marginal probability is 12% for the income-to-population relation and 11% for the population-to-washing agent relation.

We have found several relationships such as this, all making the group of customers clearer, because with each new relation, new attributes are found. One goal of the visual exploration could be to find concrete characteristics of people who buy a particular washing agent. As already mentioned, Figure 8 shows one characteristic: although the connections from household types to types of washing agents do not reveal any associations, the histograms for the latter dimension really show a clear correlation between these two dimensions. The user has reordered the top dimension to introduce a monotonic distribution of the histograms – to see the descending probability that the washing agent is bought by the household types on the right side. Concluding, the target group for liquid washing agents in terms of marketing and advertising are people who are employed, do not have children, live in larger cities, and have above average school qualifications (not presented here, but also found during the analysis). The information agrees with common knowledge in customer relationships, because for instance liquid washing agents are generally more expensive (therefore, it is surely not the favorite washing agent of people who do not have much money) and they are available in smaller sizes (the preference for town population).

Even though we only can describe a limited subset of our analysis here, we are able to demonstrate the workflow when visualizing with Parallel Sets: firstly, the user's domain knowledge is integrated by structuring the data and creating more feasible classifications of the data, then the user explores the data by relating interesting dimensions with each other and finally, histograms offer a powerful technique to identify relationships in detail. This workflow enables the user to accomplish several tasks: (1) identifying hot spots and major trends (they can be found easily, because of the visualization of frequency information), and (2) finding relations between dimensions and correlations between categories (histograms provide a handy tool to explore distributions – Parallel Sets support the comparison of relative frequency distributions and of deviations of marginal from relative frequencies).

In this regard, one of the weaknesses of Parallel Sets, which also is subject to further improvements, is that outliers cannot be seen in the visualization. Because frequency information is visualized, those categories, that feature few data records only, vanish. For the visualization, these categories pose the same challenge as hidden (filtered) categories (Section 3.3). It could be a good idea to integrate some special icon or glyph for these categories; but on the other hand, the visualization of these values could make the view more complex, than they are interesting for the user. In comparison to InfoZoom (similar idea of displaying categorical data), one drawback is that Parallel Sets are not capable of displaying $n : n$ relationships (in contrast to InfoZoom), only because InfoZoom is basically a hierarchical space-subdivision technique that facilitates these features, whereas Parallel Sets is based on parallel coordinates that only support the display of $n \times m$ data tables.

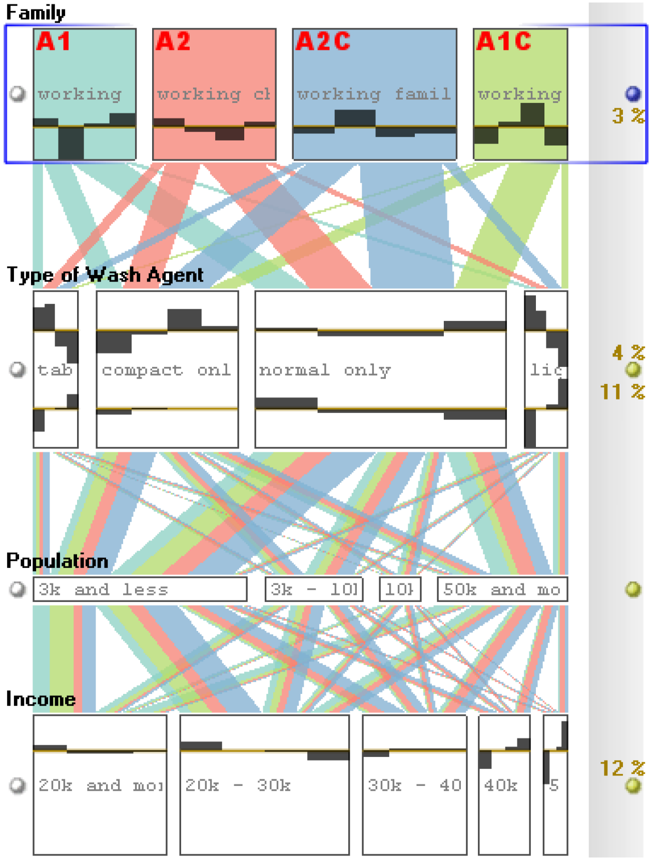


Figure 8: Four dimensions of the CRM dataset are visualized: the upper two dimensions are user-defined, and the lower two are data variables. The connections are drawn in *bundled* mode – i.e., the connections between each two categories are parallel (tidier display). The histograms for the selected dimensions show, for instance, that there is a correlation between non-children households and *liquid* washing agents, and this kind of washing agents is most frequently bought by people living in larger cities (the biggest portion of the rightmost category of the *Population* dimension).

It also should be pointed out that performance is a key issue in interactive visual data analysis, because the user has to be able to explore and change the view quickly. The analysis that is presented here was executed on standard consumer hardware (2GHz CPU, 500MB RAM, GeForce4 graphics chip) and worked interactively. The construction of the view is accomplished immediately, as long as the number of displayed categories is limited to approximately twenty to thirty categories (the delay is directly proportional to the number of categories that have to be aggregated). It should be noted that this hardly is any limitation as too many categories would overload a visualization anyway. After the categorical data is transformed to the frequency information, the reordering of dimensions and categories, the highlighting, the animation between rendering modes (Section 3.4), and the changing of the active dimension is well below one second.

A second key issue is the learning curve of the user. The analysis of our CRM partner has shown that there is some work left that should be addressed. People that already have worked with some other visualization frameworks quickly feel familiar with Parallel Sets, because the integrated interaction concepts are common and state-of-the-art (drag-and-drop, filtering, highlighting, tooltips, brushing, etc.). For newcomers, the interaction interface is easy to learn, but the visual metaphor is harder to understand, especially dimension composition is a quite complex task. Explaining the necessary steps for dimension composition is best done by the use of Venn diagrams (Figure 2c). This implies that space-subdivision techniques (such as Venn diagrams and Treemaps [18]) are more comprehensible, but they are less powerful and less flexible when the composition task deals with more than two dimensions at the same time.

5 Summary and Conclusion

Interactive visual analysis implies two requirements for a visualization technique: (1) an adequate visual metaphor that offers the user a comprehensible mental vision of the abstract data, and (2) a powerful, user-friendly, and user-driven interaction scheme. Parallel Sets fulfills both requirements. It adopts the layout from parallel coordinates (that makes the displayed dimensions visually independent from each other), but uses a frequency-based representation for categorical data variables, since frequency data is best represented by areas and not by individual data points (thus, the visualization becomes independent to the number of displayed data records). For data exploration, the dynamic layout and our sophisticated interaction scheme are important: adding dimensions to the view by drag-and-drop, reordering dimensions and categories, dimension composition, highlighting, and so on.

To conclude, it remains to be mentioned that the presented technique is an innovative idea in coping with categorical data in terms of knowledge extraction that should influence future approaches as preceding approaches have influenced this work. InfoVis is only one part of data analysis; thus, it is necessary that techniques include InfoVis ideas, as well as statistical, or data mining concepts.

6 Acknowledgements

This work has been done in the scope of the basic research on visualization (http://www.VRVis.at/vis/) at the VRVis Research Center in Vienna, Austria (http://www.VRVis.at/), which is funded by an Austrian research program called Kplus.

References

[1] ColorBrewer (http://www.colorbrewer.org).

[2] Electronic Statistics Textbook (http://www.statsoft.com).

[3] Titanic data set (StatLib – Datasets Archive: http://lib.stat.cmu.edu/s/harrell/data/descriptions/titanic.html).

[4] Alan Agresti. *An Introduction to Categorical Data Analysis*. Wiley & Sons, 1996.

[5] Richard A. Becker, William S. Cleveland, and Allan R. Wilks. Dynamic graphics for data analysis. *Statistical Science*, 2(4):355–395, 1987.

[6] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. Using vision to think. *Readings in information visualization: using vision to think*, pages 579–581, 1999.

[7] Helmut Doleisch, Martin Gasser, and Helwig Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proceedings of the 5th Joint IEEE TCVG - EUROGRAPHICS Symposium on Visualization*, pages 239–248. Eurographics Association, 2003.

[8] Usama Fayyad, Georges G. Grinstein, and Andreas Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann Publishers Inc., 2001.

[9] Michael Friendly. Visualizing categorical data: Data, stories and pictures. *SAS User Group International Conference Proceedings*, pages 190–200, 1992.

[10] Keinosuke Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., 1990.

[11] J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, 1981.

[12] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 127–130. IEEE Computer Society, 2002.

[13] Heike Hoffmann. Exploring categorical data: Interactive mosaic plots. *Metrika*, pages 11–26, 2000.

[14] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization '90*, pages 361–378. IEEE Computer Society Press, 1990.

[15] Daniel A. Keim. Visual techniques for exploring databases. *International Conference on Knowledge Discovery in Databases*, pages 1–121, 1997.

[16] Jeffrey LeBlanc, Matthew O. Ward, and Norman Wittels. Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization '90*, pages 230–237. IEEE Computer Society Press, 1990.

[17] Geraldine E. Rosario, Elke A. Rundensteiner, David C. Brown, Matthew O. Ward, and Shiping Huang. Mapping nominal values to numbers for effective visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 80–95. IEEE Computer Society, 2003.

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

[19] 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 Computer Society, 1996.

[20] Michael Spenke and Christian Beilken. Visualization of trees as highly compressed tables with InfoZoom. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 122–123. IEEE Computer Society, 2003.

[21] Soon Tee Teoh and Kwan-Liu Ma. PaintingClass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 667–672. ACM Press, 2003.

[22] Martin Theus, Heike Hofmann, Bernd Siegl, and Antony Unwin. MANET: Extensions to interactive statistical graphics for missing values. In *New Techniques and Technologies for Statistics II*, pages 247–259. IOS Press Amsterdam, 1997.

[23] Kent Wittenburg, Tom Lanning, Michael Heinrichs, and Michael Stanton. Parallel bargrams for consumerbased information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 51–60. ACM Press, 2001.

[24] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the 5th Joint IEEE TCVG - EUROGRAPHICS Symposium on Visualization*, pages 19–28. Eurographics Association, 2003.