Parallel Sets in the Real World: Three Case Studies

Robert Kosara* Caroline Ziemkiewicz† F. Joseph Mako III‡ Jonathan Miles§
UNC Charlotte Mako Metrics Gloucestershire County Council
Kam Tin Seong¶
Singapore Management University

ABSTRACT

Parallel Sets are a visualization technique for categorical data. We recently released an implementation to the public in an effort to make our research useful to real users. This paper presents three case studies of Parallel Sets in use with real data.

Keywords: Information visualization, categorical data, Parallel Sets, case studies.

1 INTRODUCTION

Parallel Sets [1] were developed for the visualization of categorical data. We had long wanted to release the program as open source, but always wanted to add more features, clean up the organically grown code, etc. We finally released it[1] on June 1, 2009.

The program was written in Java, using JOGL OpenGL bindings and SQLite for local data storage. Given its dependencies, the program currently runs on Windows, Mac OS X, and Linux/x86. In addition to running, we wanted it to feel as much like a native application on each platform as possible. The program's look and feel adapts to the platform, and it stores its database in the correct user/application data location. It also has a standard Windows installer and installs its own data on first launch on other platforms.

As a step towards production software, we also added a mechanism for checking for new versions on startup, as well as a crash reporter. The latter sends a log of important application steps as well as Java exception stack traces to our server (after asking the user for permission). These reports have proved to be extremely useful in diagnosing and fixing bugs.

Within the first six weeks of being online, the program executables have been downloaded over 350 times. Below, we present three case studies from real users who agreed to share their stories.

2 SERVICES AWARENESS CAMPAIGN (JONATHAN MILES)

Gloucestershire County Council (United Kingdom) recently ran a marketing campaign in which it advertised its services to the public. Previous market research had indicated that the public were not fully aware of the provided services, and being tax payers they have a right to know what is on offer from their County Council.

A series of posters were placed around Gloucestershire, advertising many the services such as home carers, free Internet access, adopting and fostering, etc. to teach the public about council services they might not have known about. To measure the effect of the campaign, the County Council carried out public surveys alongside the marketing campaign that asked people two key questions: *Have you seen these posters?* and *Do you think Gloucestershire County Council provide great value services?* The idea was that by comparing the opinions of people who had seen the posters against those that had not, the effect of the posters could be measured.

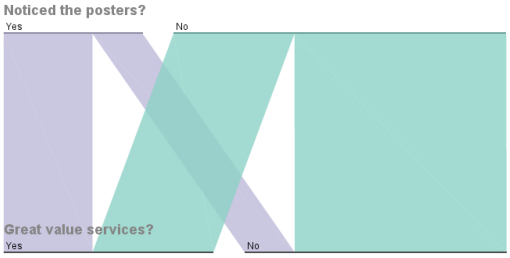

Figure 1: Opinions of services provided vs. having seen the posters.

Without Parallel Sets, plotting the results of both these questions in terms of each other would have been impossible. In a previous study, two pie charts were used: one showing the opinions of those who had seen the posters, and the other showing the opinions of those who had not. This was confusing to the reader, made visual comparison difficult, and also required an additional chart showing overall opinion about the provided services. Yet another pie chart was needed to show total exposure of the campaign: how many people had seen the posters? This would have resulted in four separate pie charts, all essentially giving glimpses of the same data set.

Using Parallel Sets, it was possible to plot all this information on a single graph (Figure 1). One can see at glance that the majority of survey respondents did not see the posters (71%), but out of those who did, the majority thought that Gloucestershire County Council provide great value services (62%); while the majority of those who did not see the posters thought the county council does not provide great services (64%). This shows that the campaign has had a small effect, and is definitely worth developing for the future.

Parallel Sets makes it easy to add dimensions, like gender and age. For the former, the visualization shows a nearly 50/50 split between males and females, and thus no effect. The latter leads to a rather pretty image, which suggests that age also has little affect on opinion, with the exception of the 31 to 40 category, in which the majority answered "No."

For Gloucestershire County Council, Parallel Sets have made presenting categorical data much easier, as they eliminate the need for multiple pie charts. For a research company for the public sector, being able to present data in a way that is easily understandable to the public is vital, and Parallel Sets provide a way to do that.

3 RECORD CLEANUP (JOSEPH MAKO)

We recently completed a complex records cleanup project, in which we kept track of each record in the approximately 20 steps. In the end, 98% of the initial records were fixed, the other 2% would need to be eliminated as they did not pass the required checks. Before we started using Parallel Sets, we would draw the process out in a flow chart with nodes for each check point. After reshaping our

*e-mail: rkosara@uncc.edu
†e-mail: caziemki@uncc.edu
‡e-mail: joemako@gmail.com
§e-mail: jonathan.miles@hotmail.co.uk
¶e-mail: tskam@smu.edu.sg

[1] http://eagereyes.org/parsets

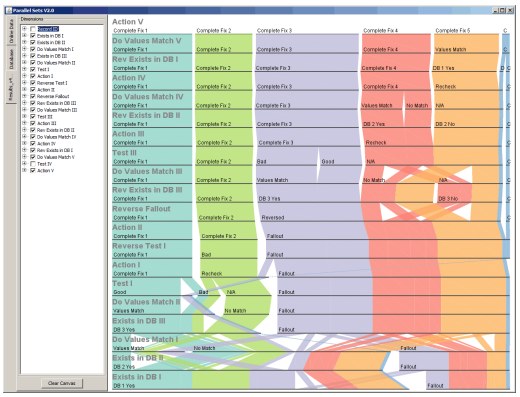

Figure 2: Tracking the steps of a data clean-up.

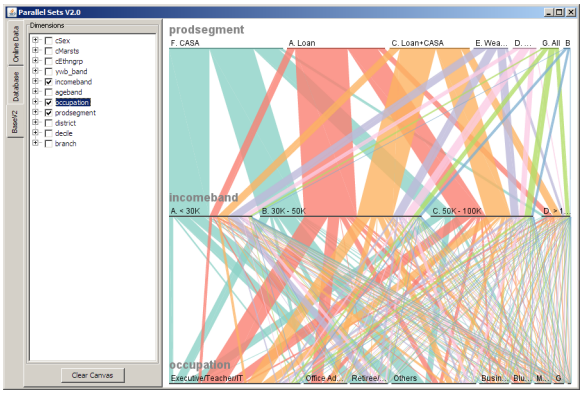

Figure 3: Exploring bank customers by income, age, and occupation.

data, we were amazed at the insight we were able to gain by putting this categorical data into Parallels Sets. We could clearly see the interesting steps, and which checks did little to affect the outcome. The most useful representation turned out to be adding each step in reverse to the canvas, last step done at the top and first worked at the bottom, which made the color directly related to the outcome and facilitated a reverse trace of records that were fixed in each of the five batches (Figure 2).

We found some interesting visual crossovers that prompted us to investigate the underlying data to ensure that no mistakes had been made. We first looked at the steps from *Reverse Fallout* to *Test III*, the crossover in the red and orange colors. We cleared the canvas, and only added a check mark next to the steps and values that we wanted to zoom in on. By focusing on this level of detail, we were able to easily confirm that all was well and that we understood what happened in this case.

We then investigated where all the fallouts occurred. We removed the ones that were fixed in batches 1 and 2 and the ones that could not be worked because they were exceptions. By limiting the display to only the first four steps and the final for coloring, we could see where the fallouts came from and how they got fixed.

Overall, passing our data through Parallel Set gives us a view of our data we did not have before, and in this case, clearly pointed out the interesting steps where the crossovers occurred so we could fully understand the threads.

4 Bank Customer Profiling (Kam Tin Seong)

Financial and banking institutions are increasingly turning to business analytics as a key means of survival. Yet, far-reaching, enterprise-wide programs are still an anomaly. This is because legacy business analysis and data mining tools are designed (or perceived as having been designed) for experts with Ph.D.s in statistics, and have largely failed to put analytics tools into the hands of data analysts who would most benefit from them.

Business users and everyday data analysts do not want to deal with advanced statistical concepts; they want straightforward visualizations and task-relevant outputs. In this case study, we explore the potential of using Parallel Sets to understand the customer profiles of different product segments and their interrelationship. This is part of a collaborative industry project between the School of Information Systems, Singapore Management University, and a global bank located in Singapore.

The dataset used in this study consists of 368,533 customer records. Six customer profile measures are included in this study: sex, ethnicity, age group, occupation group, income group, and number of years with the bank. Two parameters, product segment and customer rating in deciles, are also included in the analysis. The product segments employed in this study are divided into seven major groups: Current Account and Savings Account (CASA), Loan, Wealth Management, CASA and Loan, CASA and Wealth, Loan and Wealth, and All.

One of the questions of interest to retail banking analysts is: what are the income and education profiles of customers invested in the product segments currently offered by the bank? Knowing this information will allow the marketing team of the bank to formulate a campaign strategy to attract the targeted customer group, rather rely on the conventional "cold calling" approach.

The answer to the above question can be found in Figure 3 without the need for advanced statistical analysis. From the graphical presentation and coupling with the interactive exploration feature of Parallel Sets, the bank analyst quickly detected that customers that owned CASA only were mainly from the income range of $<$30K and 30–50K, whereas customers who had a loan only were mainly from the income range of 30–50K and 50–100K. Moving down to the third dimension, the visualization reveals that the Executive/Teacher/IT group with income range of 30–50K form the largest customer base for Loan. Customers with other and retiree occupation categories and income ranges of $<$30K and 30–50K form the majority of the CASA.

The bank analyst found the interactive selection feature of Parallel Sets particularly useful to focus on the general view to detect patterns. For example, moving the mouse over Loan at the top level showed the makeup of that customer group in terms of the income and occupation. Similarly for filtering: in order to focus on customers who owned the wealth related products, which are not revealed very clearly in the first image, the analyst just needed to uncheck CASA, Loan and Loan+CASA.

In conclusion, Parallel Sets have provided everyday data analysts at a bank with a powerful visual analytics tools to detect known patterns and discover unknown patterns from a large customer dataset without having to rely on expert statisticians. More importantly, the tool has encouraged them to combine analytics with their intuition to develop a more informed marketing strategy.

5 Conclusion

In addition to classical categorical data, Parallel Sets have proven useful as a kind of flow visualization and to illustrate simple two-dimensional datasets. User feedback and requests will guide its further development as much as research questions.

References

[1] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *Transactions on Visualization and Computer Graphics (TVCG)*, 12(4):558–568, 2006.