Your Knowledge Partner™

Accident Data Analysis with PolyAnalyst

Pavel Anashenko
Sergei Ananyan

Megaputer Intelligence, Inc.
120 West Seventh Street, Suite 310
Bloomington, IN 47404, USA
+1 812-330-0110

www.megaputer.com

Contents

Why invest in data mining software?

The State Traffic Safety Information System (STSI) estimates 6.4 million accidents occur annually within the US, with about 40,000 associated fatalities, resulting in economic losses of $230.6 billion per year (1). These gloomy figures expose the current need for government agencies to better understand root causes of accidents and to try eliminating at least some of them.

State Departments of Transportation receive reports capturing information about majority of occurring accidents, a few hundred thousand records on average, and store these data including structured data and text narratives in databases for further analysis. The objective is to discover repetitive patterns and root causes of accidents in historical data in order to increase safety in the future. One strives to look beyond obvious individual accident factors, such as speed, alcohol, weather and visibility conditions, which might be just the tip of the iceberg. Given the large volume of accidents and potentially complex interrelationships between separate accident factors, it is becoming increasingly difficult for agencies to analyze accident data without the aid of advanced data mining computer systems.

Megaputer's PolyAnalyst™ is a data mining software package designed to tackle such tasks: automatically discover and report important patterns from the analysis of large volumes of data. Upon deploying a data mining system, Department of Transportation or related government agency can make better decisions incorporating additional knowledge automatically derived from the analysis of historical traffic records.

An investment in data mining software can be very rewarding, not just in monetary terms, but in saving human lives. Data mining software is capable of discovering patterns that cannot be feasibly replicated by traditional manual approaches. These new patterns can improve the understanding of accident factors and lead to making informed decisions aimed at decreasing accident rate. Deploying a data mining system reduces the amount of human resources required to manage the analysis, reducing project cost. While information technology does not intend to fully replace human intelligence, it certainly can improve efficiency of the decision making process by automating the most time consuming routine operations and letting analysts concentrate on the comprehension of obtained results and the decision making process.

Accident Analysis: A Case Study

A particular state's Department of Transportation collects around 200,000 accident reports per year and stores these data in a database. Once collected, the data is cleaned, prepared and then utilized for analysis. The set of data in this study is real, but has been disguised to protect individuals' privacy in accordance with federal regulations. Each row in the data represents an accident, and each column is a characteristic of the event, such as the type of weather, road surface and condition, the day of the week, or contributing circumstance. A year worth of data from the database were loaded in PolyAnalyst data mining system to discover and visualize correlations and patterns in data and summarize key findings in convenient reports.

A typical first task is to identify relationships between factors in accidents. PolyAnalyst Link Charts can be used to aid the analyst in automatically identifying correlations in data and visualizing them in an interactive graph. Figure 1 displays a collection of most significant correlation between specific days of the week and contributing circumstances found by PolyAnalyst.

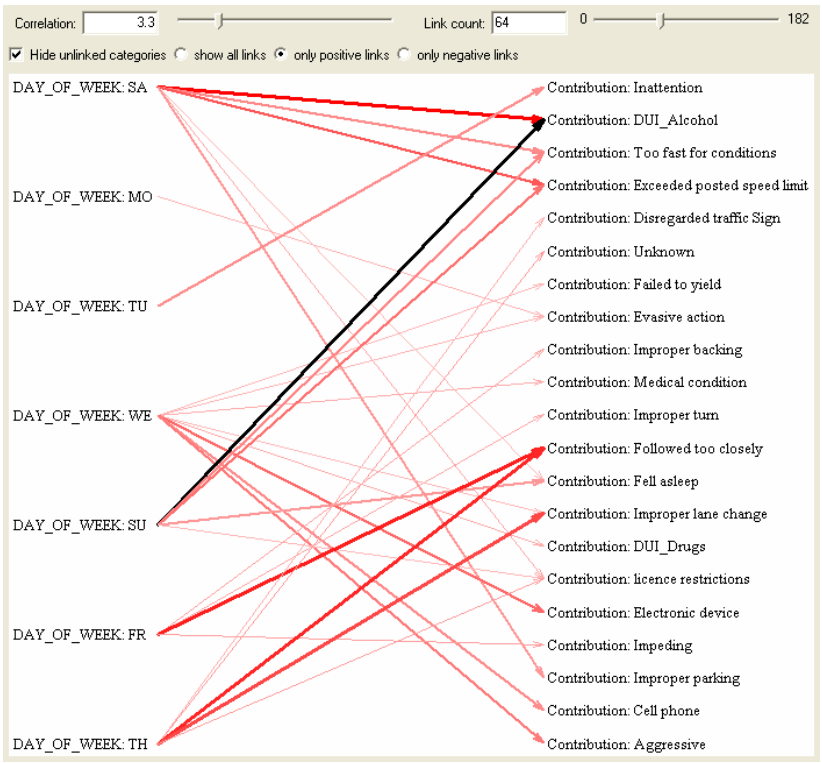

Figure 1: *Link Chart*

The left hand side of the graph displays individual days of the week an accident occurred, and the right hand side – a selection of different contributing factors. Lines in between represent discovered relationships. The darker and redder the line, the stronger the significance of the pattern. This initial graph created in a matter of seconds without any human supervision, allows the analyst to make first interesting conclusions.

Apparently, Saturdays and Sundays have strong correlation with several factors in common, such as "driving under the influence of alcohol", "too fast for conditions", and "exceeding posted speed limits". At the same time, drivers seem to be "falling asleep" more frequently on Sunday, while "improper parking" has stronger correlation with Saturday. "Following too closely" turns out to correlate strongly with Friday and Thursday.

It is important to note that PolyAnalyst calculated all these correlations from the evaluation of raw data itself, without any guidance from the analyst re: what correlations to concentrate on.

By choosing different levels of detail, the analyst is able to vary the number of links displayed on a Link Chart. The graph is interactive, making it easy to "drill-down" on any link to isolate all data records supporting the selected link. For example, clicking a link connecting "Sunday" and "Alcohol" brings up all relevant records (Figure 2). For the subset of accidents involving this pattern, some general statistics are calculated and displayed.

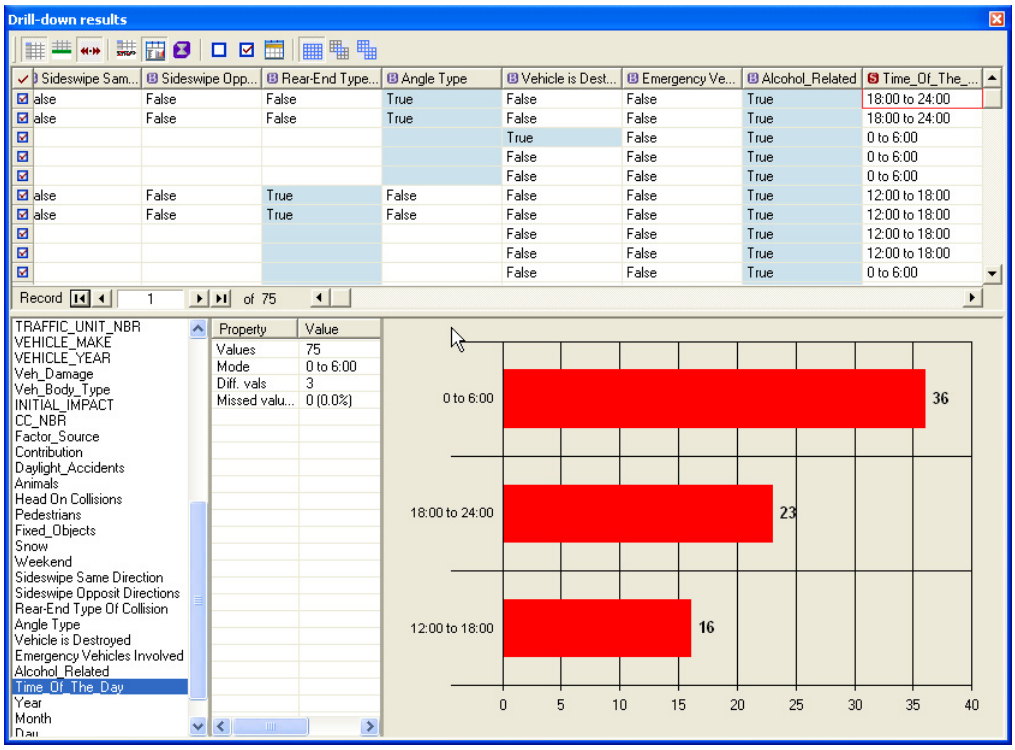

Figure 2: *Drilling down*

Continuing the process of drilling down by clicking on a bar in a histogram that represents time of the day when the accident occurred, reveals that the majority of alcohol-related accidents happening on Sunday are taking place between midnight and 6 am. Fewer alcohol-related accidents were happening Sunday evening (6 pm to midnight) and none of them occurred on Sunday mornings (6 am – noon). Using the same Link Chart and drill down capabilities, the analyst can easily identify correlations between individual values of other factors, such as location, weather, and road condition, and use this information to reduce causes of accidents in particular areas.

Link Analysis

Sometimes one might need to look at relationships between several types of factors at once. In this case, PolyAnalyst Link Analysis is the tool of choice. This tool generates a graph revealing correlations between any given number of factors at once. For example, suppose the analyst runs the algorithm on attributes containing names of streets and highways involved in accidents to expose the most accident-prone intersections (Figure 3).

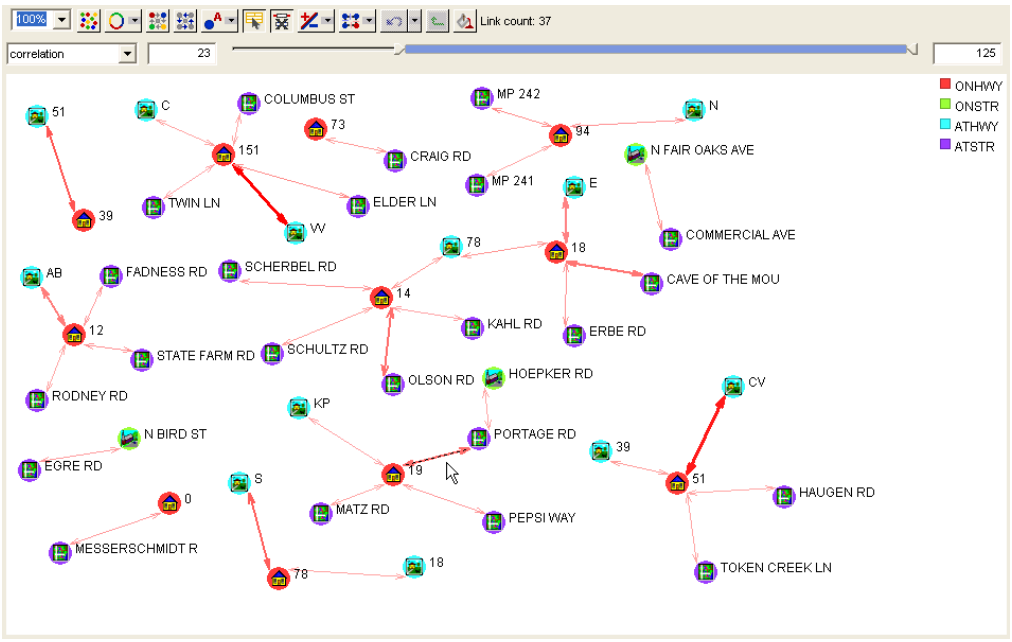

Figure 3: *Link Analysis: accident-prone intersections.*

This diagram relays some of the "most dangerous intersections" where the number of accidents is significant, given the number of accidents involving each of the intersecting roads. Circles of different colors represents names of roads or highways reported. Lines represent relationships between pairs of roads or highways. As with the Link Chart, the thicker the line, the more important the pattern.

In the following example, link analysis was performed on such attributes as the "manner of collision", "speed limit", "contributing circumstances" and "accident class". Figure 4 also illustrates how the analyst can drill down on a particular link of interest for further exploration.

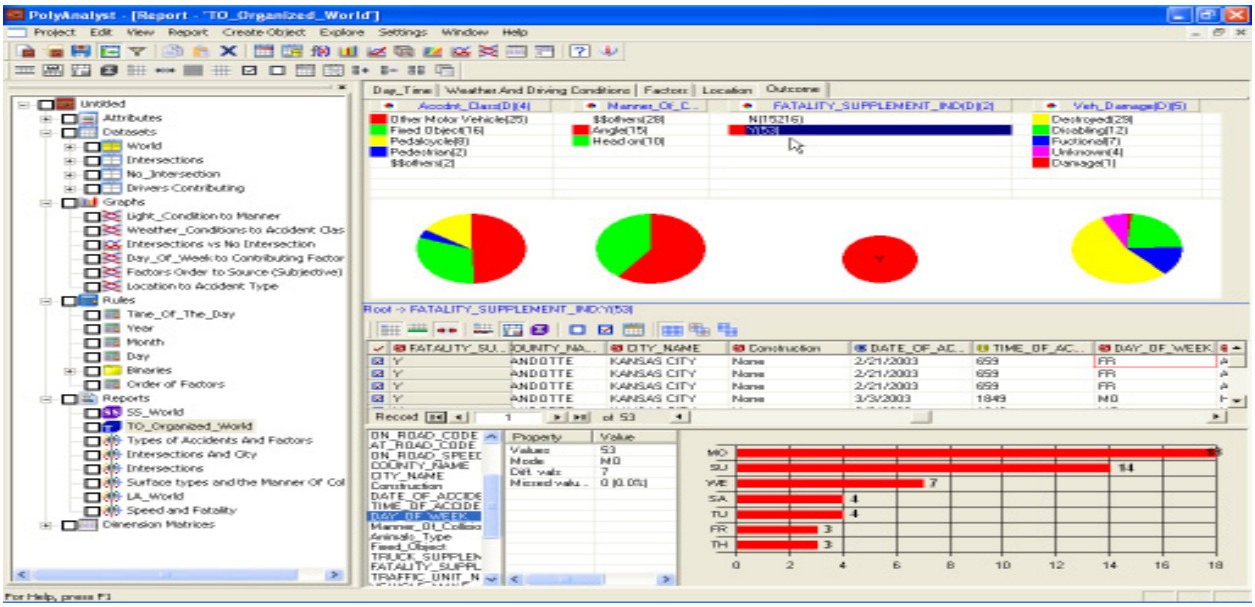

Figure 4: *Link Analysis: manner of collision, accident class, and contributing factors.*

The results show that the "Head-on" and "Angle" types of collision are correlated with the lower speed limit of 30 mph. The significant contributing circumstances to "Angle" collisions are "Improper turn, "Disregarded traffic sign" and "Failure to yield". Higher speed limits of 55, 60 and 70 mph are linked to "Rear end" type of collision and "Sideswipe, same direction". "Too fast for conditions" and "following too closely" are the most correlated factors for "Rear end" collisions, while "improper passing" and "improper lane change" – for "Sideswipe, same direction". In the 20 mph speed limit zones drivers are hitting parked motor vehicles, while collision with fixed object is linked to higher speed limits of 65 mph. Some of the links are trivial and self-explanatory. Others may be puzzling, which suggest one might wish to perform further investigation using drill-down. For example, the drill-down insert in Figure 4 illustrates that "Intersecting roadways" happens to be the most typical location for head-on collisions, while there are very few "highway interchanges" where head-on collisions occur. This may explain the reason for "Head-on" collisions to be correlated with lower posted speed limits.

Radar Charts

This tool allows users to compare two or more datasets on a number of different parameters at once. Instead of just looking at correlations, we can compare factors across different types of accidents. Figure 5 displays a brief comparison of relative importance of different accident characteristics for intersection-related accidents vs. those not involving any intersection.

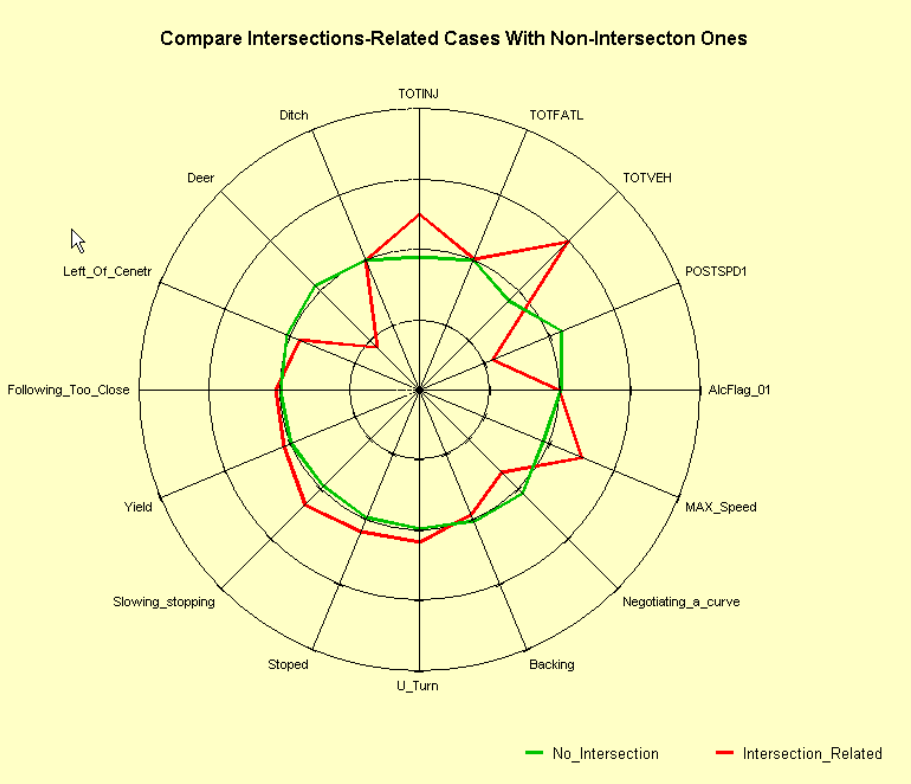

Figure 5. *Radar Chart*

The total number of vehicles and the total number of injured people are higher for intersection-related accidents, while such factors as failure to negotiate a curve or violation of posted speed limits are more important for accidents with no intersections involved.

Suppose the analyst needed to investigate what leads to "Deer" factor being more important for non-intersection accidents. Clicking on the corresponding line and drilling down reveal that in terms of location, the majority of cases are happening at the highway ramps, notification times are mostly between midnight and 3 am (about 75% of cases) and the light condition shows "Dark" (no artificial light present) for the most of them. Perhaps installing extra guard rails or fences on highway ramps and better illumination of these areas is the best way to tackle with the deer problems for the region under consideration.

This type of the analysis is also useful in comparing datasets involving different weather or light conditions, commercial vs. personal use of the vehicle and even the makes and models of cars. In the latter case some particular strong or weak points may result in safety recommendations for manufacturers.

Dimension Matrix

Browsing through large amounts of data and trying to locate the subset you need can be difficult sometimes. The Dimension Matrix allows the analyst to quickly focus on any facet of the data and be able to pull all the relevant records, without writing any SQL queries. In this case study the dimensions of accidents are time, weather conditions, location, outcomes and various contributing factors. By selecting with a mouse click any point (it is also possible to select continuous ranges or arbitrary sets of points) at any dimension, one immediately gets all the records described by the selection.

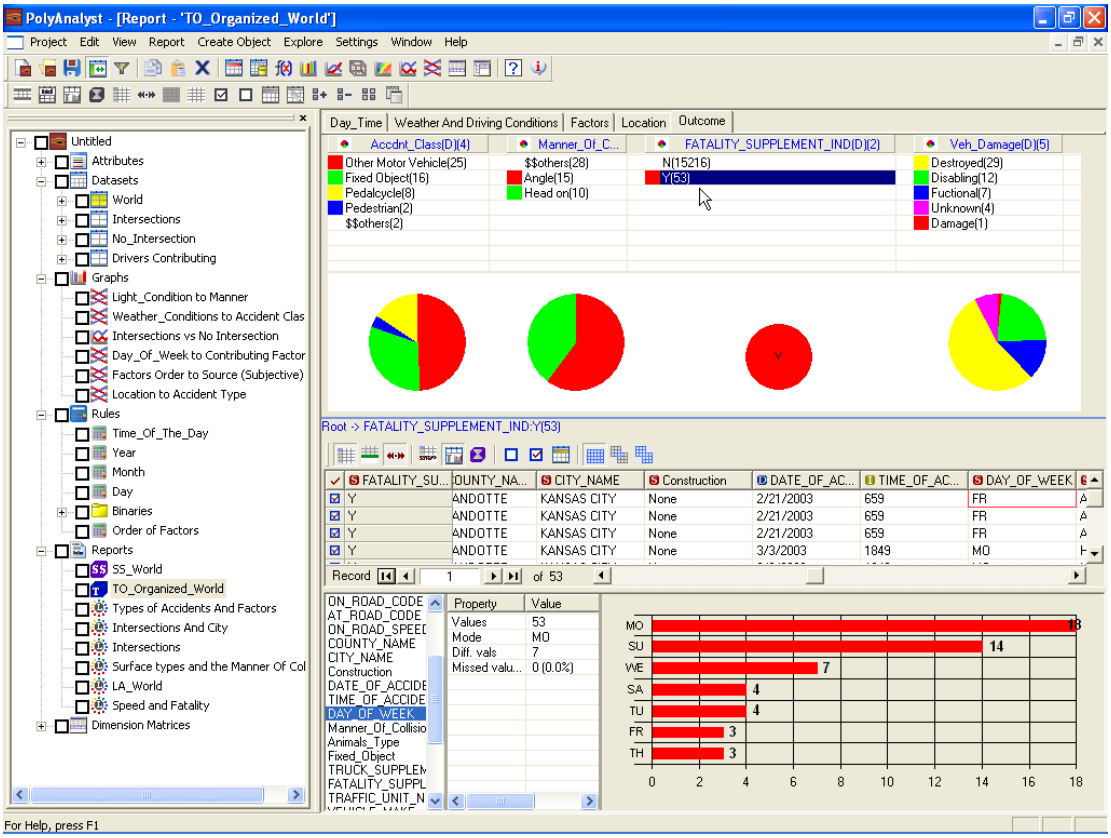

Figure 6: *Dimention Matrix.*

As illustrated by the above example, upon selection of any point(s) on any dimension(s), the relative frequencies for all other columns adjust according to the new distribution or records conditional on the selection made. The information panel allows getting all the statistics for the slice of data defined by the selection, do further drill-downs and save/export the corresponding datasets.

Conclusion

The use of a PolyAnalyst in the analysis of accident data:

- Can increase the speed and quality of analysis compared to manual processing
- Enables analysts to get objective, data-driven results
- Provides immediate visual interaction with the obtained results

The ability to quickly obtain objective, data-driven results from the analysis of raw data is particularly important in case of no prior biases existing about the dependence of a large number of attributes and their values present in the dataset. Discovering links among collected accident's characteristics and their distribution across different locations will help Departments of Transportation make infrastructure improvements at specific intersections, decide where to place new traffic lights, or eliminate insufficient or confusing signage. The ability to identify those accident factors that are most important for selected locations will help decision makers determine whether a multi-level intersection or just a traffic separation with a median barrier is needed. Finding out the most dangerous driving habits will help select most important information to include in a driver's manual, while linking particular makes and models of cars to accidents involving certain weather or surface conditions may justify the notification of the manufacturer of the car functional problems. These insights, made through the use of data mining software like PolyAnalyst, should help reduce the growing annual accident rate, diminishing both the project cost and the loss of human life.

References

1. Car Accident Statistics. http://www.car-accidents.net/car-accident-stats.html

Accident Data Analysis

Corporate and Americas Headquarters
Megaputer Intelligence, Inc.
120 West Seventh Street, Suite 310
Bloomington, IN 47404, USA
TEL: **+1.812.330.0110**; FAX: **+1.812.330.0150**
EMAIL: info@megaputer.com

Europe Headquarters
Megaputer Intelligence, Ltd.
B. Tatarskaja 38
Moscow 113184, Russia
TEL: **+7.095.951.8079**; FAX: **+7.095.953.5731**
EMAIL: info@megaputer.com

Megaputer
Your Knowledge Partner™