Your Knowledge Partner™

Crime Pattern Analysis

Megaputer Case Study in Text Mining

Vijay Kollepara
Sergei Ananyan

Megaputer Intelligence
120 West Seventh Street, Suite 310
Bloomington, IN 47404 USA
+1 812-330-01

www.megaputer.com

Contents

Industry Situation

Building on the success of employing the analysis of structured data to help solve and prevent crimes, Law Enforcement and Government organizations are seeking to expand the scope of their analysis to include unstructured text data. While typically over 80% of all information available to an organization resides in text form, the analysis has till date been primarily confined to only structured portion of available data. Missing four out of every five bits of useful knowledge is a very high price to pay for the lack of efficient means for text analysis. An ability to perform in-depth analysis of text data could provide both corporate and government organizations with many new insights. Yet until recently, such analysis required significant manual labor of reading and coding text narratives - the process too slow and prone to errors.

Today, new data and text mining technologies provide a next generation of tools for the analysis and visualization of both structured data and text. Such tools help increase the quality and productivity of the analysis and reduce the latency period between recording raw data and obtaining key knowledge necessary for making informed decisions.

Case Overview

A police department had a large collection of police information reports (PIR) that were filled out by officers at the time of recording incidents over a period of several years. The main portion of each PIR holds a text description of the incident. The department was seeking a capability to identify historical crime patterns from a large volume of unstructured data.

There are many questions investigators needed to get answered quickly:

- Are there correlations between the crime type and the location of the incident?
- What are the distributions of crime types involving suspects of different ethnic origin?
- How can I quickly extract reports characterized by certain parameters of interest? For example: robberies performed by white teenagers involving the knife threat.
- Are there correlations between the type of crime, weapon employed, and the location of the incident?
- What is the most typical weapon in cases when high school students are charged with weapon possession?

Traditionally, the process of finding answers to these questions involved the analysis confined only to structured portion of the data, somewhat enhanced by an officer's ability to recollect relevant past cases and repetitive keyword-based searches of the text portions of reports. Manual analysis of all PIRs was a cumbersome, time-consuming process prone to errors and biases. New automated text analysis could help the agency quickly and consistently discover important patterns in crime occurrences and empower police officers and analysts to

- Learn from historical crime patterns and enhance crime resolution rate.
- Preempt future incidents by putting in place preventive mechanisms based on observed patterns.
- Reduce the training time for officers assigned to a new location and having no prior knowledge of site-specific crime patterns.
- Increase operational efficiency by optimally redeploying limited resources (like personnel, equipment, etc.) to the right place at the right time.

Methodology

Megaputer Intelligence carried out an incident reports analysis project to demonstrate a complete analytical solution for processing a mix of structured data and text in incident reports. This case study discusses a methodology and sample results of discovering knowledge hidden in unstructured data. The project was carried out with the help of the data and text mining system PolyAnalyst™.

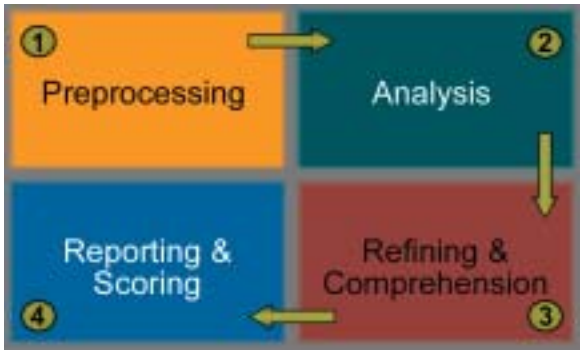

Figure 1: *Data Analysis Methodology*

An overall objective was to create an analytical solution that investigators can routinely use to identify new patterns and associations between types of incidents, locations, time and descriptive details of the incident. The developed approach consists of a series of steps:

- Preprocess data to the format suitable for further analysis
- Extract important concepts and terms through text-mining
- Analyze patterns and co-occurrences of identified concepts
- Develop an automated solution for crime pattern analysis.

Analysis & Results

Data preprocessing

The first step in creating an analytical solution involves understanding data and transforming it to a convenient format. The original PIR data was in the form of text documents and contained information entered by the investigating officer at the time of the incident. All this information was stored as unstructured text reports, singling out only the date the report was filed, the corresponding police station, and the classification of an event by a field officer filing the report. These text documents were parsed into a database format separating structured and unstructured portions of PIRs. Data was normalized.

Figure 2: *Police reports data loaded in PolyAnalyst*

Concept extraction

The first type of text analysis starts with capturing key concepts and terms in text descriptions present in PIRs with the help of a text mining engine. This engine can run in an unsupervised mode, when clusters of unusually frequently occurring terms are automatically discovered by the system, or a supervised mode when the user focuses the analysis performed by a text mining engine to only primary topics of interest to the user.

A police investigator exploring historical reports might want to capture all terms related to:

- *Weapons*
- *Narcotics*
- *Schools*

and then checking if there exist correlations between the found items and other crime characteristics (such as event type and location). This was achieved by focusing the PolyAnalyst Text Analysis engine to look for all particular instances of these broad category terms. For example, focusing the text mining engine on finding and tagging cases that involve particular instances of the term *weapon*, the investigator obtains the results shown in Fig. 3. The user does not have to manually specify all possible particular instances of *weapon*. Based on a comprehensive dictionary of English terms and semantic relations between them, PolyAnalyst text mining engine automatically expands the search and finds all possible weapon-related terms in the collection of investigated police reports.

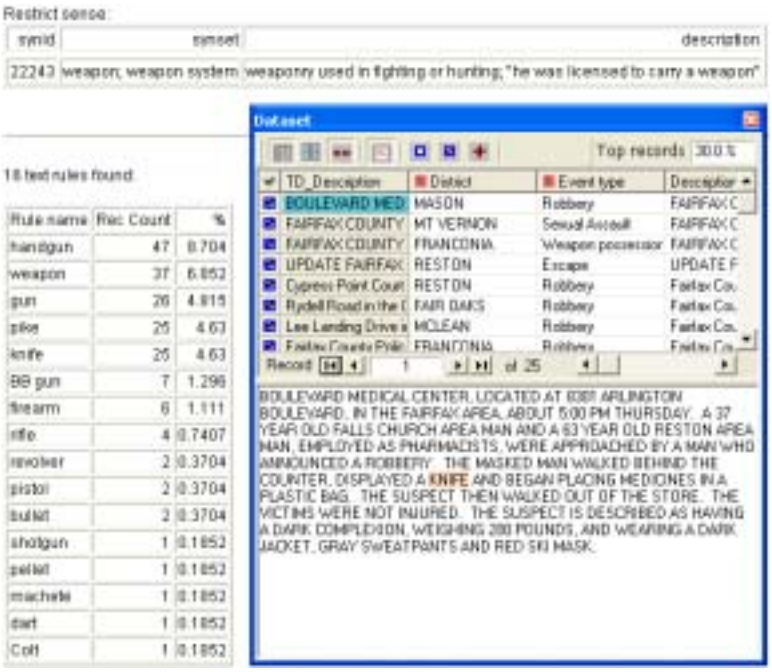

Figure 3: *Results of Text Analysis focused on* ***weapon***

It proves to be quite useful that the system supports interactive drill-down from the discovered weapon-related terms to original records in the data with the corresponding terms highlighted. For example, one can quickly learn that while the term *pike* can potentially represent a medieval weapon, in all cases recorded by police officers the term *pike* represents a certain type of a highway (Leesburg Pike, Columbia Pike, etc.). Correspondingly, the term *pike* can be excluded from the list of weapons discovered by the system in the investigated police reports.

Similarly, the investigator can run the text mining engine to extract all particular instances of *drugs* or *narcotics*, as well as particular instances of *school*.

Pattern Analysis

In the next step of the analysis, all extracted terms were used for tagging individual reports, allowing further usage of these terms as new structured attributes of individual PIRs together with the original structured data: *event type* and *location*. Then PolyAnalyst Link Analysis engine was applied to reveal associations between the extracted terms and individual values of structured attributes and displays a visual map of correlations between individual crime types, locations, weapons or narcotics involved, and school type mentioned. A sample pattern of associations extracted from text data in investigated PIRs is shown in Fig. 4.

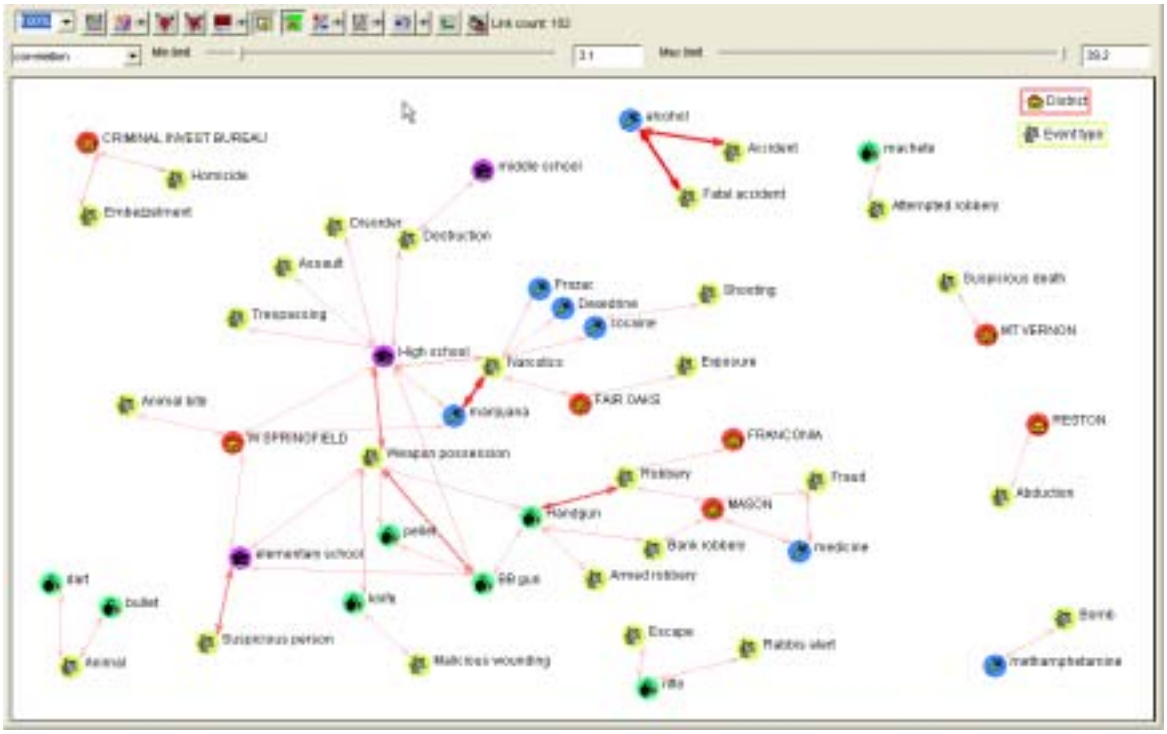

Figure 4: *PolyAnalyst Link Diagram displays correlations between* ***locations****,* ***event types****,* ***school types****, and* ***weapons*** *and* ***narcotics*** *involved*

The figure displays a number of isolated clusters of terms representing different patterns. The saturation and thickness of each line on the graph represents the strength of the correlation between the terms it connects. Upon viewing this link diagram, one can immediately draw a number of important conclusions:

1) The most typical crime occurring at *high schools* is *Weapon possession*.
2) Yet, in the majority of cases these are just *BB guns* that are classified as *Weapon possession*. This can become a serious problem for any subsequent analysis of the collected data: looking only at *Event type*, investigators will be unable to differentiate incidents involving *BB guns* from those where real weapons were involved (such as *knife* or *handgun*).
3) The second important high school problem is *Narcotics*, of which *marijuana* is the most popular one and linked back to *High school*, followed after a large gap by *Prozac*, *Dexedrine* and *cocaine*.
4) Other types of frequent events at high school include *Trespassing, Assault*, *Disorder* and *Destruction* (the latter is shared as a frequent problem with *middle schools*).

This amazing list of important conclusions drawn from a single easy-to-understand visual graph can be continued.

Drill-down and reporting

Law enforcement agencies need to be able to validate conclusions made on the basis of visual interpretation of the results of Link Analysis. A drill down to the selected patterns of interest helps in isolating the relevant records and validating our conclusions. Upon a click of a mouse, PolyAnalyst provides the user with a collection of records supporting the selected links, with terms of interest highlighted, as illustrated in Fig. 5.

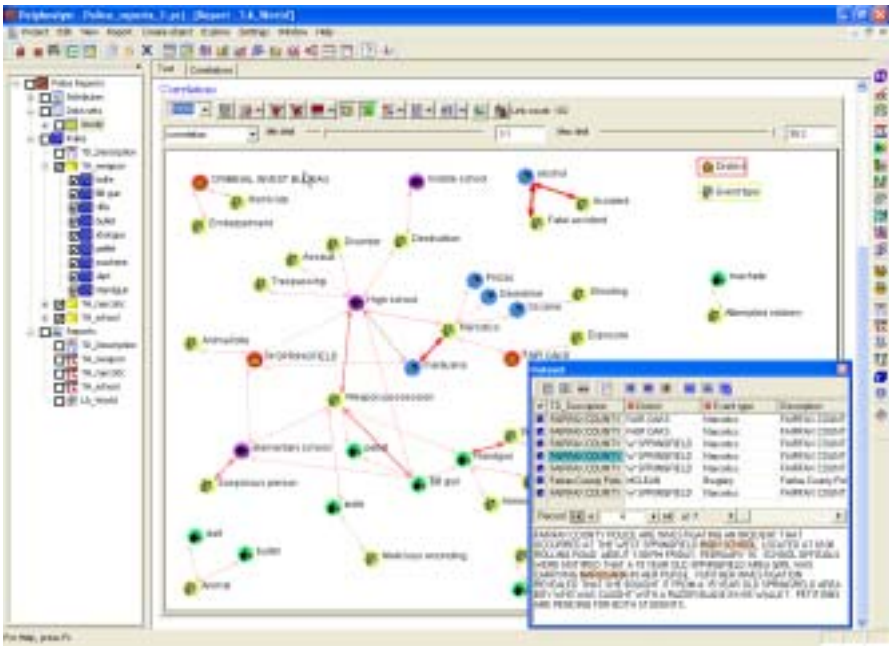

Figure 5: *Drill-down from Link Diagram to original records corresponding to the selected link*

These records can be exported in an HTML report or saved as a new data set for further analysis.

Text OLAP

Law enforcement officers are often interested in seeing distributions of criminal events by location, type, weapon used and other characteristics. This task is somewhat similar to OLAP analysis widely used in the corporate world for the analysis of structured data. However, in the case of the analysis of police reports, the main portion of information of interest is buried in unstructured text, and this causes the failure of standard OLAP in addressing this task.

PolyAnalyst provides a unique Text OLAP engine capable of organizing a mix of text and structured data in interactively manipulated multi-dimensional cubes. Text OLAP engine allows the user to define a number of dimensions of interest that include information extracted from natural language text, in addition to regular structured attributes. This makes all information hidden in the text portion of reports available for quick and efficient manipulation, analysis and reporting – a capability many investigators had dreamed of for a long time.

Fig. 6 shows a matrix defining attributes and values that will be used by PolyAnalyst Text OLAP engine for analyzing PIR data. There is one dimension defined by values of a structured attribute – *Event type* – and four more dimensions defined by values extracted from the main text portion of police reports: *Ethnicity* and *Age* of suspects, *Location* of the incident and *Weapon* used. Possible values of the suspect *Ethnicity* were defined as *White*, *Asian*, *Hispanic* and *Black*, possible *Age* groups – as suspects in their *teens*, *twenties*, *thirties*, *forties* and *fifties*, and most frequent incident locations were spelled out. It is interesting to note that the system was again instructed to search for all particular types of *weapon*, and returned back all specific types of weapon mentioned in the investigated collection of police reports automatically.

Figure 6: *Dimensional matrix defined by an analyst*

Upon applying the developed Dimension Matrix to data, the system generated an interactive Text OLAP report allowing the user to see distributions of data records across different dimensions, drill-down to subsets of records matching all selected criteria, browse through the corresponding original text records with the terms of interest highlighted, and shift around the order of dimensions of interest.

The user can derive a wealth of information from each view of the Text OLAP report. For example, from the report displayed in Fig. 7, one can first observe that out of 540 investigated reports, in 73 cases suspects were described as *Black*, in 47 – as *White*, in 31 – as *Hispanic*, and in 11 – as *Asian*. Drilling down to see the distribution of *Event Types* involving *Black* suspects, one immediately sees that the most frequent crime type for this group of suspects is *Robbery* – 52 out of 73 events!! All other event types lag *Robbery* by a wide margin. Continuing the drill-down process on *Robbery* events and selecting further branches of interest, one sees that the majority of these robberies involve suspects in their *twenties*, confronting victims when they approach their *cars*, and displaying *guns* before demanding money.

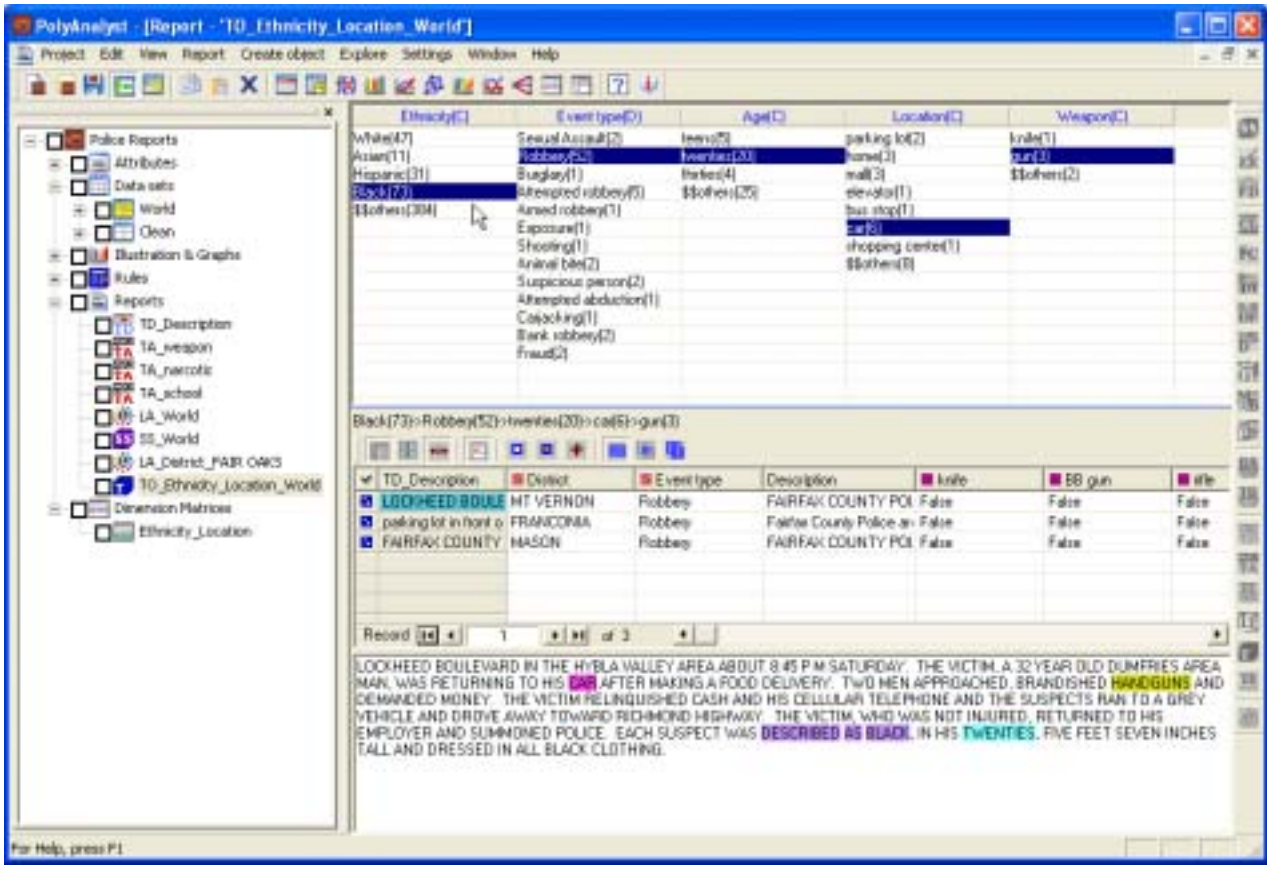

Figure 7: *PolyAnalyst Text OLAP report: Black => Robbery => twenties => car => gun*

The analyst can save all records supporting an investigated drill-down node to an HTML report maintaining the highlighting of terms found by the Text OLAP engine. Such reports listing all records of interest help quickly substantiate recommendations made for decision makers.

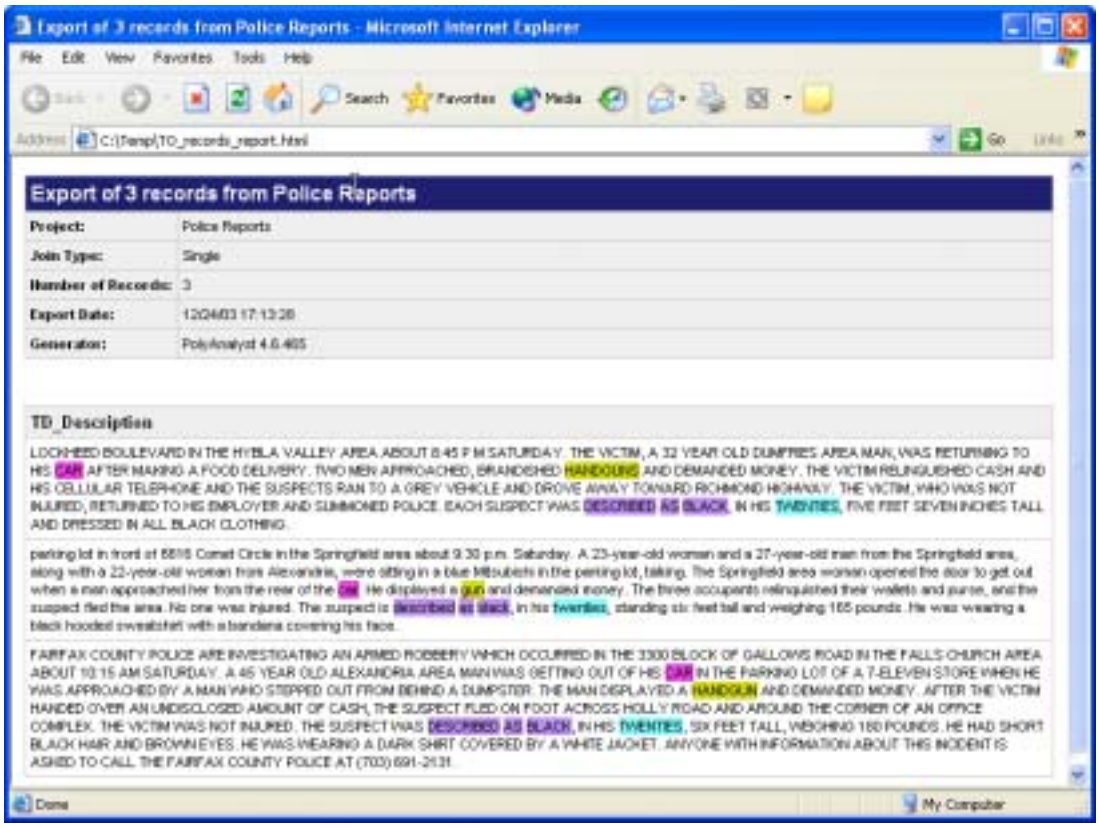

Figure 8: *HTML report with drill-down records*

Being an interactive decision support tool, PolyAnalyst Text OLAP report allows the analyst to quickly change the focus of the analysis and drill down on other branches of the report or change the order of the defined dimensions to obtain different other views of the data.

For example, Fig. 9 illustrates that while *Robbery* is still a number one crime performed by *White* suspects in the considered county (15 cases out of 47), a relative share of this crime type compared to others is quite low compared to cases involving *Black* suspects. On the other hand, for *White* suspects there appears a second widespread *Event* type – *Exposure*, which is occurring almost as frequently as *Robbery*: 9 cases out of 47. It interesting to note that in the majority of these cases the age of suspects is not being reported.

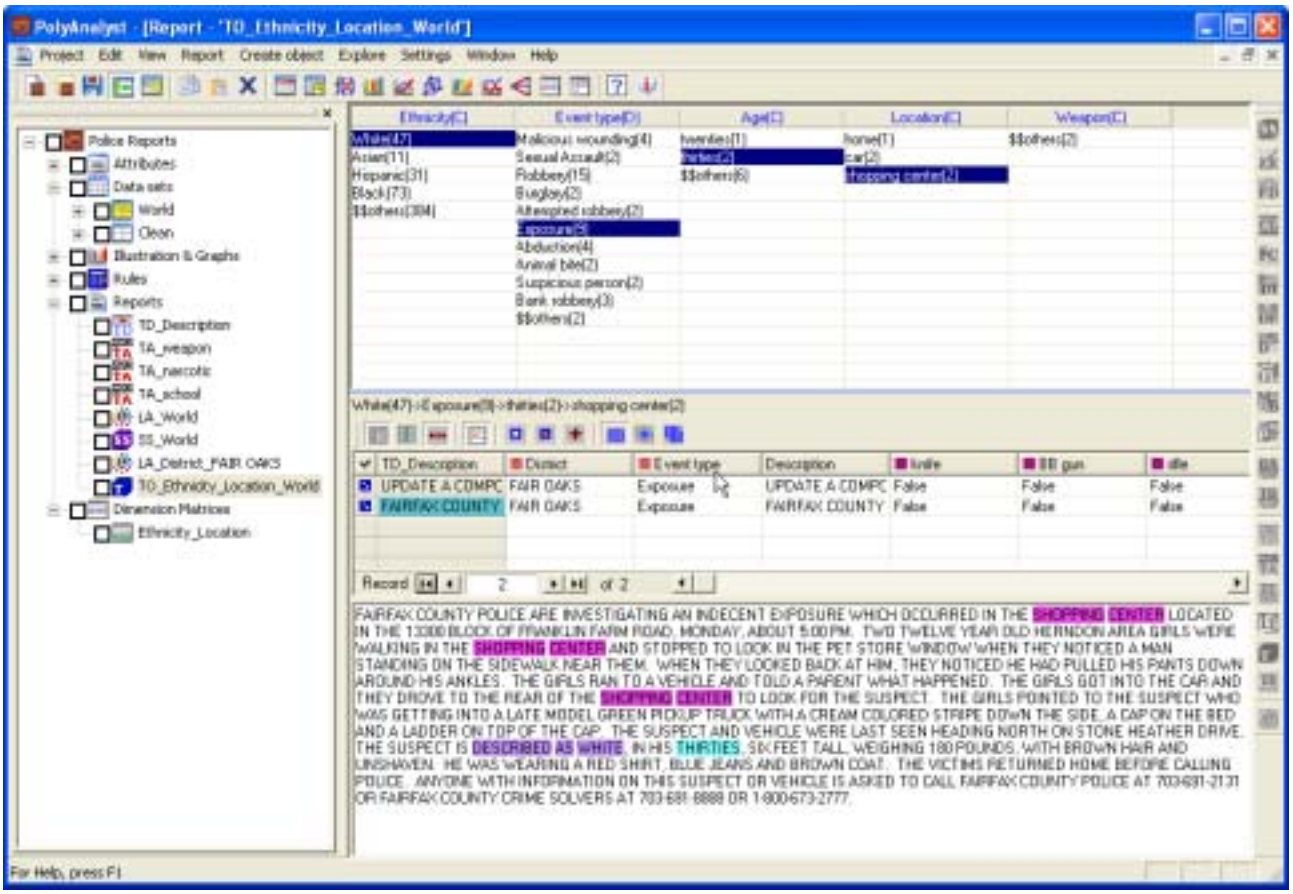

Figure 9: *PolyAnalyst Text OLAP report: White => Exposure => thirties => shopping center*

Then, by simply changing the order of defined dimensions in the report, one can read other valuable information directly from the same Text OLAP report. For example, Fig. 10 illustrates that overall the most frequent crimes in the investigated PIRs are *Robbery* (117 out of 540 cases) followed by *Malicious wounding* (30 cases).

An investigator interested in the distribution of *Weapon possession* events can immediately learn that a vast majority of them occur at *school* (16 out of 18 cases) and most frequently involve either a *knife* (6 cases) or *bb gun* (6 cases).

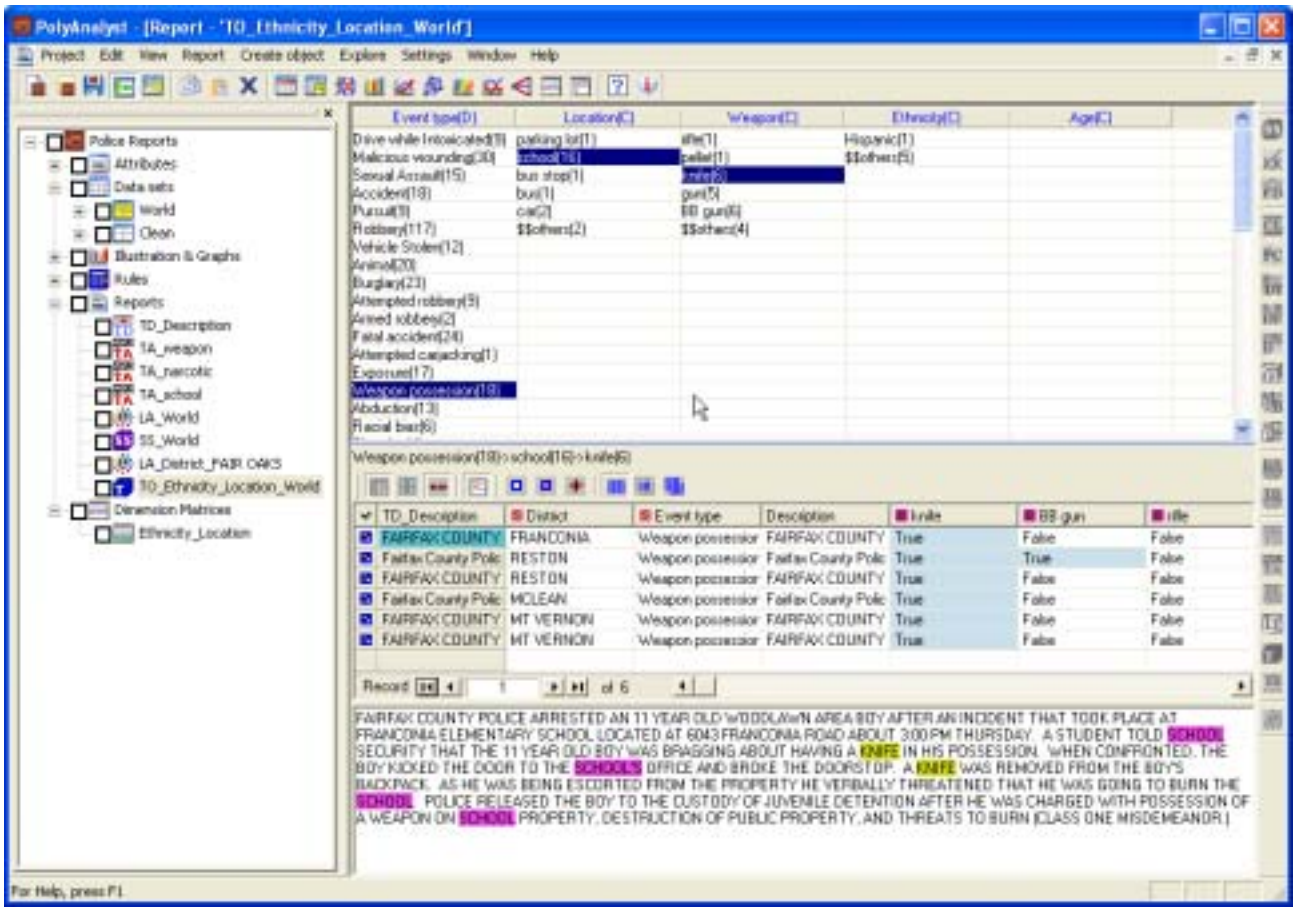

Figure 10: *PolyAnalyst Text OLAP report: Weapon possession => school => knife*

Previously, these types of observations were based only upon the experience of local investigators. Now such conclusions can, not only be made based on new observed patterns but be easily substantiated with immediate drill-down to highlighted text of relevant reports.

To sum up, we have seen how the pattern analysis can be easily performed on large volumes of data containing a mix of structured data and text collected by a law enforcement agency. The ability to quickly draw meaningful conclusions from the analysis of text data becomes an invaluable tool for educating newly coming officers about historical trends at particular locations, generating aggregate views of data for better decision making, and helping in the allocation of critical resources to appropriate areas.

Automation

The entire process of pattern extraction and visualization can be automated to a large degree, so that the results of the analysis are easily derived and shared by many users across the organization. It is also possible to interface the discussed analytical techniques into existing IT systems. An access to an easy-to-manipulate front end implementing the collection of related interactive visual reports with powerful data and text mining engines embedded in the background, empowers investigators with a new capability to quickly arrive on reliable conclusions based on objective analysis of large volumes of unstructured data.

Conclusion

The considered case illustrates an overall process for implementing a text mining solution and proves the feasibility and value of performing simultaneous analyses of both text and structured data within the same software system.

The discovered results help investigators identify hidden patterns through the automated analysis of historical police reports. Till date, this knowledge was largely dependent on local expertise (so called 'local veterans'). Moreover, the new approach to the analysis delivers a much more comprehensive and objective overall picture of the incidents as it involves evaluating both structured and textual portions of the database.

Law enforcement agencies and government organizations can benefit from this combination of text mining and pattern analysis technologies by achieving:

- Improved crime resolution rate
- Optimal resource allocation based on dynamically changing patterns
- Faster and more up to date results from raw data
- Reduced officer training time and costs
- Better crime prediction and prevention of offences.

Both government and corporate organizations are now redefining the boundaries of traditional analytical solutions that were till date centered on analyzing only structured data. Decisions made are now be based on the analysis of all available data, including the most information-rich text portion of the data, rather than on subjective analysis relying primarily on analysts' experience. As compared to the previously employed manual knowledge discovery process, modern analytical tools help accomplish superior analytical projects a hundred times faster, while consuming less than 10% of previously required resources. Text Mining is becoming the cornerstone technology in the formation of a new Intelligent Organization.

Corporate and Americas Headquarters
Megaputer Intelligence Inc.
120 West Seventh Street, Suite 310
Bloomington, IN 47404
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 **+7.095.953.5394**; FAX **+7.095.953.5382**
EMAIL info@megaputer.com