Analytics for Data Deconfliction

Christopher WESTPHAL
Visual Analytics, Inc
www.visualanalytics.com

Abstract
This paper introduces a refactoring of the concept called "deconfliction" as it applies to analytical and data repositories deployed throughout the law enforcement and intelligence communities. The goal is to seamlessly and automatically identify similar targets of interest (e.g., people, places, etc.) from the queries and reports being generated by the analysts, agents, and end-users operating these systems. The similarities represent the underpinnings for a number of requirements, including more efficient utilization of resources, improved analytical support mechanisms, a foundation for information sharing functions, as well as internal controls to help minimize abuses related to data privacy concerns. The application of data deconfliction is quickly becoming a mainstream capability and closely parallels the adoption of more information exchange standards. This paper discusses several methods and issues related to the emergence of data deconfliction activities.

Introduction

The content of database systems is expanding in near exponential rates and their utilization by different consumer groups is steadily increasing due to the advent of more open standards, web services, and other access protocols / interfaces. Commonplace today are wireless handheld devices which allow police officers to easily run field checks looking for potential criminals, license plate readers (mobile or fixed position) that constantly scan for stolen or unregistered vehicles, and various biometric sensors used for facial recognition, speaker identification, or even thermal body scans for

exposing disease signatures. All of these touch points are accessing various repositories looking for specific types of values or results – which then determine some type of appropriate action or follow-up.

The byproduct of these interactions, whether successful or not, are the trail of queries (i.e., the logs) and values used to access the respective systems. These logs are traditionally used for administrative functions to convey system usage statistics as well as to provide basic audit-level detail. Essentially, the log files represent their own independent data source that reflects the interests of the community they serve. Related commercial examples of how log files can be exploited include Google Trends or Yahoo Buzz List which are used to show popular search terms and help highlight new trends and domains of interest. These types of applications, albeit higher level aggregations and abstractions of the specific query instances, represent how insight into the utilization of log data can help improve the overall offerings, results, and experiences realized by their end-user communities.

The same can be said for law enforcement and intelligence applications. Consider a patrol officer running the license plate of a vehicle that was pulled over for a moving violation infraction (e.g., speeding); the queries posed to the system include the plate, which in turn results in the identification of the registered owner along with an address. This type of information is typically cross-referenced with other sources to expose outstanding warrants and criminal histories. In this case, the information returned does not indicate a high-level threat. However, several days earlier, the same plate was checked within the vicinity of a gang-related hit resulting in the shooting deaths of several persons. At this time, the correlation among the data sources has not been established because the "checks" often do not become part of a permanent record and the officer is exposed to an extended level of risk due to this type of oversight.

All of the interactions with a database create an audit trail which has not traditionally been considered a mainstream source for incorporation into the analytical process or for information sharing purposes. The use of data

deconfliction techniques to help identify similar targets can improve officer safety as well as provide more accurate results. This type of offering provides contextual awareness for analysts or agencies for information sharing purposes to help identify and coordinate analytics among seemingly disparate investigations.

Deconfliction

There is actually no recognized or official entry in the dictionary[1] for the word "deconfliction," which one could reasonably consider a type of antonym for *confliction*. As such, new approaches, concepts, and methods typically take liberties to coin new terms, especially in technology-related industries. For our purposes, the word *deconfliction* is defined as a means or a process used to identify potentially similar interests, whether they are events, actions, or data values.

Typically, within the realm of law enforcement, the term deconfliction is used in the context of "officer safety event deconfliction." Different law enforcement agencies with active operations could potentially interfere with each other, especially when they are on the same target or collocated in the same area (e.g., neighborhood or building). For these situations, there may be different types of events, including surveillance, undercover operations (e.g., drug buys), knock-and-talks (e.g., parole violation checks), or actions, such as raids or sweeps, that could potentially compromise the operations and potentially result in a dangerous condition for the officers involved.

Fortunately, there are real-world systems (e.g., RISSafe[2]) that are designed to help mitigate conflicts and notify any involved agencies about potential overlap when these types of situations exist. These systems have been instrumental in helping to reduce risk, which can potentially compromise operations, and worse, result in injury or loss of life to law

[1] http://www.merriam-webster.com/dictionary/deconfliction
[2] http://www.riss.net/rissafe.aspx

enforcement officers and / or operatives. To properly operate, the systems require each operation to be reported to a centralized monitor where it is reviewed for conflicts. If it is deemed that an operation might be in jeopardy, the concerned parties are promptly notified. Of course, this only works if there is full compliance by all participating agencies – and not all agencies or organizations necessarily contribute or document all of their ongoing activities to support these types of systems. It is not a perfect process, but it does effectively address a critical need, especially within the law enforcement community.

This same concept can also be applied to analytical and information sharing systems under an extended definition of "data deconfliction" where the utilization of the accounting logs for an analysis become a foundation of data that is integrated into the overall system thereby identifying areas of interest, common targets, and potential system abuses. Simply, if one user queries "William Washington" and another user also queries the same name or even a slight variation such as "Bill Washington," then the system would identify a probable match and generate the appropriate notifications.

There are a number of government and commercial systems that have active monitoring processes to control (via security protocols) or oversee who accesses their systems. Depending on the organization, the data can be stored as "primary entities," commonly referred to as a Master Name File. Systems that utilize such approaches, where each record correlates to a single and distinct entity (e.g., a person), include, for example, phone subscribers, driver's licenses, passports, and tax revenue databases. Access to high-profile accounts, such as celebrities and politicians, can be more closely scrutinized. A prime example of this occurred during the 2008 elections in the United States, there were several incidents[3] where unofficial access to specific candidate's passport records[4] was detected. The individuals

[3] Vijayan, Jaikumar, "FAQ: The passport breach: What exactly is in those records?" Computerworld, March 21, 2008.

[4] AP Press "Passport files of candidates breached: Records of Clinton, McCain, Obama inappropriately accessed, officials say" March 21, 2008, http://www.msnbc.msn.com/id/23736254/

involved in initiating the queries were identified and properly disciplined.

However, when such a degree of representation is not available (e.g., a master name), an agency must rely on other methods from which to track their interactions with the data sources. In a majority of agencies, the database logs are merely used to capture and encode basic operations such as invalid login attempts, expired passwords, transaction logs for loading new information, and other administrative actions. Often, due to performance reasons, the queries generated by the associated analytical software are not stored in the database logs. In these cases, the agencies rely on the internal logging and auditing capabilities of the respective analytical and reporting software to reconstruct the content of the database results.

Analytical Approaches

Law enforcement investigation techniques have traditionally been more reactive in nature and are typically in response to a crime that has already been committed. As more and more intelligence is incorporated into their business processes, law enforcement personnel are now able to act in a more proactive fashion and are therefore more capable of preempting or circumventing crimes from occurring in the first place. The process of incorporating more proactive analytics into daily operations has become the foundation for the success associated with intelligence-led policing programs[5].

The use of data deconfliction techniques for reactive analytics is fairly straight forward. Generally there are only a limited number of target entities (e.g., people, places, events etc.) related to these types of analytics. The nature of a reactive investigation typically has a starting point and expands outwards based on the volume of data related to the designated targets – thus, each query has well defined and specific values used to access the underlying data. For example, the following represents an

[5] Peterson, Marilyn, "Intelligence-Led Policing: The New Intelligence Architecture" prepared by the International Association of Chiefs of Police under cooperative agreement number 2003–DD–BX–K002 awarded by the Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice , September 2005 (NCJ 210681)

example of a standard SQL query used to access information on a specific person from a designated source:

```
SELECT
        qrysubjects.ALIAS, qrysubjects.SBJDOB,
        qrysubjects.ETHNIC,
        qrysubjects.SBJEYES, qrysubjects.GENDER,
        qrysubjects.SBJHAIR,
        qrysubjects.SBJFNAM, qrysubjects.SBJLNAM,
        qrysubjects.SBJMNAM, qrysubjects.SBJSNAM,
        qrysubjects.RACE, qrysubjects.WEAPONS,
        qrysubjects.TATTOOS, qrysubjects.NARRATIVE
FROM
        qrysubjects
WHERE
        qrysubjects.SBJLNAM = 'SMITH' AND
        qrysubjects.SBJFNAM = 'JOHN' AND
        qrysubjects.SBJDOB = {ts '1970-01-14 00:00:00'}
```

This SQL statement clearly shows that the request is for a SUBJECT with a last name of SMITH, first name of JOHN, and a date of birth equal to 01/14/1970. In this example, a single SUBJECT is returned from the source; however, one must be aware that even very specific queries may actually return multiple results, especially when dealing with a common name. For deconfliction purposes, other analysts making queries with matching (or similar) values would be flagged. Expanding the relationships associated with this SUBJECT shows there are a number of associated entities, entirely based on the source(s) accessed, and in this case include other subjects, phones, identification numbers, addresses and case references as shown in Figure 1.

Figure 1. *First Level Expansion from Target Entity*

Each one of the entities returned by this first-level expansion will also be included within the query logs and subject to their own deconfliction process. Therefore a match can occur on virtually any type of entity, as long as the values are consistent in the underlying query structures. Furthermore, some places "simulate" a deconfliction by incorporating their existing case- or records-management system as one of the primary sources queried. Therefore, whether a match was encountered in the log files or through prior case matches, the analyst can be notified by defining a special icon that differentiates the entity from a normal result. In Figure 2 below, the SUBJECT shown at the 10:00 position is marked with a red-exclamation point to signify there is an existing case (or potential inquiry by another analyst) on this particular entity.

Figure 2. *Entity Deconfliction Depicted With Special Icon*

Reactive queries are generally utilized by subscription-based services[6] where a pre-defined target entity is submitted to the system for matching purposes. There are usually some types of mandatory values or fields associated with interacting with these services that include, for example, the first and last names along with a residence state. Multiple results are possible; however, the specific details are only presented for a selected entity. If the service is made available with the analytical tool, then the query can be recorded and treated just like any other log entry. However, if the query is done manually through an alternative interface, then that level of detail will be lost unless there are other protocols or write-backs built into the analytical tool to document this process.

When the types of queries posed to the system are more proactive or strategic in nature, the process of identifying similar or like-entities can be

[6] Includes offerings by Lexis/Nexis, Acxiom, Factiva and other subscription-based services.

compounded. This is due in part to the overhead involved with transforming the query into tangible entities. Thus, a query such as "show all ADDRESSES from a specific ZIP Code" (query structure shown below) may return hundreds or thousands of results, and depending on the analytical system, can be represented as an aggregated value or as individual entities.

```
SELECT
        qrysubjects.ADDRESS,
        qrysubjects.CITY,
        qrysubjects.STATE,
        qrysubjects.ZIPCODE
FROM
        qrysubjects
WHERE
        qrysubjects.CITY = 'FREDERICK' AND
        qrysubjects.STATE = 'MD'
```

If the results are defined as entities, then the process is identical to what was described previously. An example of this output is shown in Figure 3 where addresses involved in prior queries are shown with red question marks. However, if the results are merely aggregates (e.g., counts of results), there is no way to convey that an entity has been previously observed in other queries based merely on the query submitted. Depending on the source being queried, large volumes or batch-type queries may not be supported, thereby excluding them from consideration from the data deconfliction process.

This naturally brings up the related discussion of *object-oriented* versus *record-oriented* approaches as they apply to deconfliction purposes. Some analytical tools are exclusively focused on depicting the records returned from a query where every row represents a unique instance of a result (e.g., transactional focus). These types of systems tend to cluster, or aggregate, row counts based on a common field value (e.g., a breakdown of incident types by date and region). In these instances, the application of data

deconfliction proves to be difficult because there are no target values from which to perform the matches since all rows are treated equally.

In an object-oriented approach, a series of field values uniquely defines the objects (e.g., entities). When the combination of values is encountered in the data, the object is considered to be an equivalent. Thus, objects representing, say, a SUBJECT, might be constructed from a combination of: 1) first name, last name, and middle name or 2) first name, last name, and date of birth or 3) first name, last name, and ZIP Code. Each different combination of values will affect how the system behaves and produce different results depending on the collection and content of the underlying data.

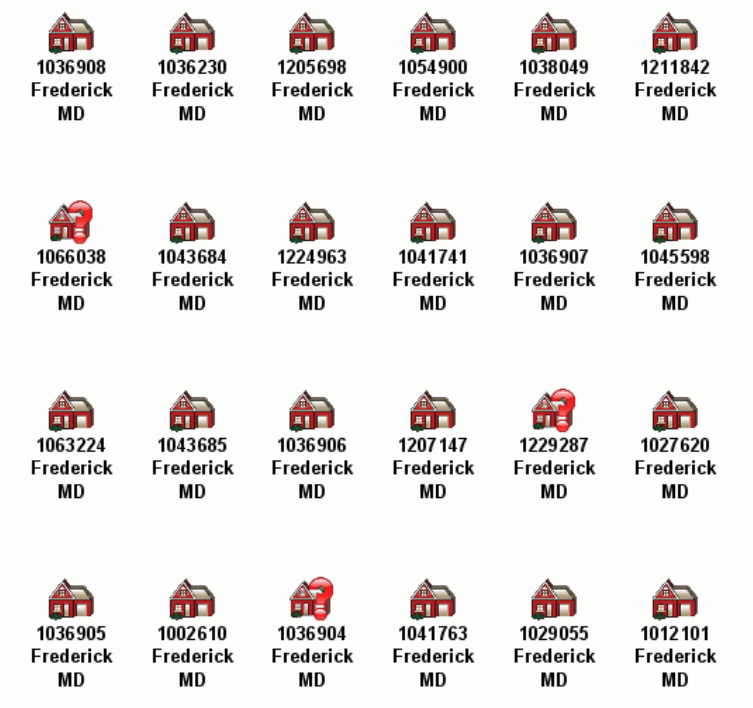

Figure 3. *Depiction of Proactive Query Results*

Additionally, how the deconfliction was defined will prove valuable to those involved, especially as it relates to the approach used to establish a

positive match. If a result is encountered based on an explicit query performed from a reactive search, then it should be viewed in a higher regard because it was of particular interest to another analyst.

Conversely, a result from a proactive search may merely be due to a wide net being cast looking for potential targets of interest. Not until the analyst has narrowed down the focus of the proactive set to a few specific entities (e.g., made it reactive) would the degree of interest be increased. Additionally, another dimension to factor into the deconfliction process is the timeframe in which the queries are / were performed in the context of the analysis. Thus, an analyst that targeted an entity from, say, three years ago might not truly have much of any impact on the deconfliction process for one that more recently queried on the same entity. The overall process should take this into consideration.

Matching Values

Using an object-oriented approach for representing data provides for the most direct method of matching values. The most simplistic is considered to be an equivalence match; where all of the components for the queried entity (e.g., first name, last name etc.) have the same values as those stored in the log files. It is assumed, by the time the results are logged, that there have been some basic value transformations, a certain degree of data clean-up, and an overall disambiguation of the values.

Additionally, the model used to represent the data as an object will have standardized the mapping to the raw table and field values associated with each source. Essentially, in its most fundamental form, the logs are merely just another data source that has been incorporated into the overall analytic process.

Needless to say, the quality of the underlying data will have an impact on the overall matching capabilities used for the data deconfliction

purposes. Many times, there will be variations, aliases, or other types of fuzzy values that won't necessarily match using equivalence. The amount of up-front emphasis placed on the overall entity resolution will determine how accurately the matches for data deconfliction will be performed. Care must be taken not to generalize too much with respect to how the matches occur; overgeneralization will result in too many false positives and diminish the value of the deconfliction process.

Of course, certain types of objects including, for example, addresses, phone numbers, identification numbers, vehicles, and accounts are fairly easy to standardize. Generally, these types of objects are considered *unique*[7] with respect to the real-world counterparts they represent because there can only be *one* of them in existence. Thus, matches within this class of objects are considered very reliable; however, care must be taken to incorporate the timeframes that the data was deemed valid as to avoid situations where re-issued values could potentially cause a situation.

The names of people and organizations are where most of the matching problems are encountered since they are not unique and can represent numerous real-world entities. Although detailing the techniques used for fuzzy matching are outside the scope of this manuscript, there are a variety of approaches that can be used to expose similar or like names using a combination of aliases, transformations, phonetics, and transpositions. To streamline the matching process and depending on the analytical tools being used, the resolved entity values could be written directly to the log files. This would require a single comparison to a single field value – speeding up the overall process. It also introduces one final point based on the anonymization of the data. Encapsulating the value of the entity into a hash value or

[7] Westphal, Christopher "Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies" CRC Press, December 2008. pp. 42-50.

encrypted format allows for matching without disclosing the raw details of the object. This can be an attractive option, especially when operating in a federated search space, where multiple agencies and users are sharing data across a wide range of activities.

Reporting

There are a number of factors to consider when detecting and reporting matches within disparate queries. Often there may be sensitivities associated with the nature of the suspects being evaluated or reviewed that could range from internal checks (e.g., inspector general / internal affairs) up to national security concerns. There are even special cases, including data related to grand jury cases, which must managed according to specific protocols. Depending on the type of data deconfliction reports that are desired, there can be different types of notifications and / or matches reported including, for example, explicit, silent, management, or special purpose(s). Other types of reports could also be defined, but for brevity, not all are covered here.

Explicit deconflictions would provide feedback directly to the analysts working on their individual cases. Thus, an alert would be generated and delivered via the analytical environment (e.g., a tool, portal or through e-mail) to notify the involved parties.

Typically, this type of unrestrictive or unbounded type of alert would be most favorable within a single agency focused on a common task, for example, Financial Intelligence Units (FIU) detecting money laundering operations or a Special Investigative Units (SIU) combating insurance fraud. Each analyst is committed to furthering the objectives of the agency and collaborative work products are encouraged to help expose larger networks of criminal behaviour.

Silent deconflictions or hits are implemented in several government systems including Treasury Enforcement Communications System (TECS),

the National Crime Information Center (NCIC), and the Automated Case Support System (ACS). These types of matches occur based on the usage and utilization of the system (based on master name lists), except that when a match is encountered, the notification is sent to a designated individual, and not the originator of the request. Usually these are assigned for more sensitive types of data, which could include notifying a case agent or senior management that access to sensitive data has occurred or was attempted (based on security levels).

Management reviews on data deconflictions can also be presented in reports or dashboards where the details can be aggregated to show the level of overlap or commonality among different requests. This can help with both strategic and tactical planning purposes, support more focused targeting objectives, and show where investigative overlaps occur across different organizations, especially for federated searches involving multiple agencies. Ultimately, it will help to bridge different cases that might have appeared unrelated – thereby helping to manage the risk associated with examining and pursuing larger case loads.

Conclusion

This paper provides a high-level overview of how data deconfliction can be incorporated into standard analytical systems to improve the general understanding of potential common targets across different investigations. The actual implementation, optimization, and accuracy of the results produced by such systems will vary depending on the underlying infrastructure and analytical approaches used by the respective agencies. Simply, the log files (or even case files) act as additional data sources that are incorporated into the overall analytical process. Detecting matches among the entities contained within these log files potentially represents a

common target or interest among the different users or agencies.

The concept of data deconfliction has long been an area of interest within the law enforcement and intelligence communities. As more and more analytical systems are incorporating the ability to provide proactive data requests, the utilization of data deconfliction methods becomes even more important.

Although multi-source data integration has made tremendous strides over the past several years, especially with the introduction of certain representation standards[8], there is still as lot of work to be done with respect to how the content or value is represented. Matching like or similar values can be challenging, especially when dealing with near real-time environments.

The outputs, reports, or notifications of potential matches are also an area that requires further development particularly when multiple sources and agencies are involved. Determining how someone should be notified, how often, and in what manner is still a very personalized dimension that is heavily based on the sensitivity of the data, the role of the analyst, and the charter of the agency. More common standards will evolve as the deconfliction capability becomes a more mainstream feature.

Ultimately, the ubiquitous capability to monitor, track, and correlate data for linking cases and investigations together will be commonplace among our analytical systems. Data deconfliction will provide support for constructing larger-sized cases, enhanced investigations using reactive analytics, and supply additional insights for exposing proactive targets.

[8] NIEM – National Information Exchange Model – is one such example of defining structural equivalence among different data sources.

Biography

Mr. Westphal is co-founder and CEO of Visual Analytics Inc. (VAI), a provider of visualization software, information sharing systems, and advanced analytical training. His clients include federal and state / local law enforcement including fusion centers, all major intelligence agencies, the US Department of Defense, and international Financial Intelligence Units (FIUs). Mr. Westphal has authored numerous publications and several books including *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies* (Westphal, CRC Press, 2008); *Data Mining Solutions: Methods and Tools for Solving Real World Problems* (Westphal / Blaxton, Wiley, 1998); and *Readings in Knowledge Acquisition: Current Practices and Trends* (McGraw/Westphal, Ellis Horwood Limited, 1990). He also authored the "Analyzing Intelligence Data: Next Generation Technologies for Connecting the Dots" chapter in *Net-Centric Approaches to Intelligence and National Security* (Ladner/ Petry, Springer 2005).