Using Data Mining for Intelligence-led Policing and Crime Analysis

Dmytro Uzlov
Head of Information and Analytical Support
Department of the Direktorate of National Police of Ukraine in Kharkiv region
Kharkiv, Ukraine
poputcik@i.ua

Volodymyr Strukov
Head of Information Technologies Department of the Kharkiv National University of Internal Affairs
Kharkiv, Ukraine
struk_vm@ukr.net

Oleksii Vlasov
Deputy head of Information and Analytical Support
Department of the Direktorate of National Police
of Ukraine in Kharkiv region
Kharkiv, Ukraine
moonreactor@gmail.com

***Abstract* — This article is devoted to a review and perspectives of using data mining methods in the work of criminal analysts in the national police by the process of developing and implementing proactive police activities for the prevention and investigation crimes. It also describes data mining tools for improving the effectiveness of information-analytical work of the law-enforcement agencies through the creation of automated intelligent technological tools. The operative part of the article outlines the basic principles, models and technologies that specialized software is used to support the analytical work of criminal analysts.**

Keywords— intelligence-led policing, crime analysis, the implicit connection, hidden patterns, spatio-temporal analysis, the visual analysis of the crime.

I. INTRODUCTION

The fight against crime requires law enforcement agencies to find new approaches to the collection, analysis and evaluation of intelligence. Intelligence-led Policing (ILP) is a model of a proactive policing activity that uses gathered data and knowledge as a fundamental basis for informed decision-making. ILP is based on a comprehensive criminal analysis of a current situation and aims at the effective use of available forces and means of law enforcement agencies.

II. CRIMINAL ANALYSIS

Criminal analysis is a consistent complex processing and detection of the interrelations between any criminal-significant data and other data potentially significant for police, judicial and criminalistics practice as well as understanding of the essence of such interrelations. Criminal analysis is based on intelligence (legal) and public information. It includes the analysis of crimes and criminals, crime victims, disorder, traffic accidents, and internal police operations.

Depending on a task and the end user of the analytical product, criminal analysis is partially based on the methodology of data mining (all possible methods of classification, modeling and prediction: decision trees, neural networks, fuzzy logic, etc.), as well as partially on statistical methods (correlation analysis, regression analysis, time series analysis, link analysis).

An analyst should study, as far as possible, all available information, and then, based solely on the facts, put forward hypotheses, make predictions and estimates. The hypothesis put forward by criminal analyst gives a theoretical premise that needs to be checked in order to see if it is correct. In particular, they should consider the following aspects of criminal activity:

- who is involved: after the formation of a list of persons and organizations, for one reason or another participating in the event or process, the relationship scheme with the weight coefficients of the relationship is modeled;
- what and how these persons do: a description of events (processes) is subjected to classification and clustering in order to form formalized models of the participants' behavioral profile;
- where and when everything happens or will happen: the analysis of time series in combination with geographical analysis makes it possible to construct a mathematical model of the observed phenomenon not only for the purpose of forecasting the development of the situation but also for verifying and justifying the hypotheses and versions developed during the investigation;
- why participants act in one way or another: on the basis of the formed behavioral profile of the participant (organization), not only the evidence base is formed, but also the justification of motivation.

Police activity of law enforcement agencies when conducting investigative activities is based on a wide range of initial data. The event (crime), which will be subject to analytical manipulation, often has three components: spatial, temporal and descriptive. Therefore, the analysis of the event will be spatio-temporal in nature.

978-1-5386-6611-1/18/$31.00 ©2018 IEEE

III. DOMAIN PROBLEMATICS

There are two fundamental questions arise in the process of spatio-temporal analysis of criminal events in the practical application Data Mining methods.

1) What should be the model that defines the space-time relations of environmental objects and how to build it?

2) How is it possible to use these models directly for analysis and what mathematical apparatus is best for this?

In a criminal analysis, the environment is fuzzy, random, has a large number of characteristics and requires the consideration of deeply structured relationships between sets of features and objects. Thus, the analyst is only able to use fragmentary knowledge about environment, which leads to the use of models, based on symbolic representations, such as Markov models, Modal Calculi or Descriptive Logics. Consequently, the automation of the criminal analysis process requires the organization of deductive inference procedures with all the ensuing consequences, as well as procedures for extracting complex knowledge about the criminal environment, evaluating the complexity of calculations and many others.

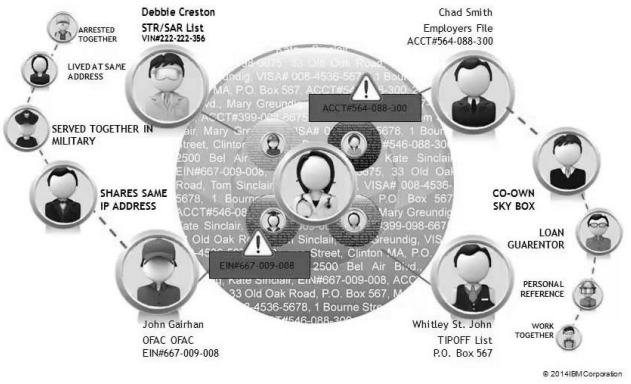

Fig. 1. Non obvious Relationships extracted from criminal environment.

Often, based on a multi-valued logic, and having a sufficient number of descriptive components, it is possible with a high probability to identify an event from any point of view. Then, the only thing left to do is to recognize the basic logic connecting all the dynamic processes. In the theory of operational-search activity, such logic is detailed from the point of view of the criminal process and the logic of predicates.

Nowadays the information about all incidents is entered in the databases of law enforcement agencies, but in fact, there were no tools for an effective analysis of such information. In some measure, this is dictated by the characteristics of typical data sets of criminal information:

- large amounts of data (up to several tens, and sometimes hundreds of millions records); and the number of these data is increasing day by day;
- large number of object features (up to hundreds characteristics);
- the various nature of features (as a rule, non-numeric);
- the possibility of the presence of omissions (the absence of values where they should be) in the data sets due to a number of subjective and objective reasons.

In practice, there is a situation of requirement to determine the proximity of multidimensional objects, taking into account the peculiarities of the formation databases of criminal information. The solution of this problem allows us to apply classical algorithms of clustering, classification and association to identify implicit and hidden links between the objects of criminal records in databases. In particular, such tasks as the search for crimes by analogy, the definition of a circle of suspects for a particular crime or a group of crimes, etc.

IV. FUNDAMENTALS OF SOFTWARE DEVELOPMENT

At the present stage of development of information and analytical technologies, a clear realization has occurred that the tools of the criminal analysis, search, link building, behavioral profile analysis and other analytical techniques, must constitute a single whole system, providing an analyst with a full range of solutions needed for construction of cause-effect relationships.

We will consider the possibility of constructing the corresponding software using the example of Real-time Intelligence crime analytics system (RICAS) is the first intelligent criminal data analysis system that has combined in single visual space the main and most advanced methods of criminal analysis, including methods of analytical search, which is very important for the possibility of crime investigation both in hot pursuit, and unsolved crimes of the past.

The system implements elements of basic criminal analytics that allow the following types of analysis: Crime Pattern Analysis, General Analysis, Methods Analysis, Case Analysis, Comparative Analysis, Offender Group Analysis, Specific Profile Analysis, Investigation Analysis. Using all these types of analysis integrally, it becomes possible to see the whole picture - predicatively and post factum. As the system is a superstructure over existing databases, it can display explicitly specified links between individuals, as well as build visual connections between persons who at the first glance, are not related to each other. The system uses several algorithms to find links.

The first algorithm is a Recursive Search for relationships on persons who participated in different events.

The second algorithm is the Visual Search for links. In the process of outputting structured information into the visual display environment, links like "place of accomplishment-accomplice-criminal", "crime-suspect-accomplices" become obvious. Due to the fact that user directly works with data presented in the form of visual objects, which he can view from different sides and from all angles, he can get additional information that will help him formulate new research goals or significantly delve into the subject of current ones. In this case, the hypotheses, in the future, are easily checked by automatic instruments (methods of statanalysis or Data Mining).

In addition, direct involvement of the analyst in visual analysis has two main advantages over automatic methods:

1) visual analysis makes it easy to work with heterogeneous and noisy data, while not all automatic methods can work with such data and give satisfactory results;

2) visual analysis is intuitive and does not require complex mathematical or statistical algorithms.

Also, the system includes a powerful core for working with semantics. Analysis of unstructured data takes place in real time. To unify search functions and build a behavioral profile, an algorithm for automated classification or "tagging" is used.

All events related to the person and his accomplices are analyzed, which makes it possible to analyze group criminality in various directions, and leads to an analysis of the behavioral profile of the group. The semantic core of the system allows you to build complex search queries, including all possible dynamic and static components - time limits, a method of committing a crime, dislocation, etc. All functions are performed instantly and allow you to visualize information and conduct analytical work quickly.

The system includes Visual Temporal Analysis. A display of the chronology of the occurred events and time delimitation allows us to reveal hidden spatial-temporal patterns between different events promptly.

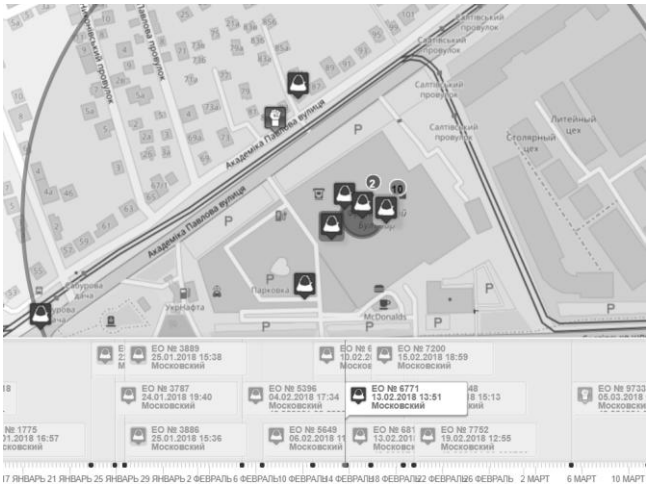

Fig. 2. Visual Temporal Analysis on chronology of the occurred events.

Let's formulate the basic principles on which the work of the intellectual system of criminal analysis in real time is built.

A. The Principle of Behavioral Profiling

The most constant and accurate, in terms of the psychology of an offender, is his behavioral profile. It displays many parameters of a criminal activity - a usual way of committing a crime, places of committal and other minor dependencies which in aggregate correspond only to a single profile. Presence of certain behavioral traits, with a certain degree of probability, may indicate that the subject might be involved in the event. Using this principle allows you to perform Group Behavioral Analysis.

Of course, a behavioral profile of an offender cannot exist without affecting other subjects. Therefore, in criminal practice, coincidences by various behavioral parameters among different subjects that have ever participated in the same events are often noticeable. It allows you to identify accomplices, associates, it would seem, without obvious links between them.

B. The Principle of Finding Hidden Patterns

Hidden patterns with a high degree of probability can always identify a link between a perpetrator and all the crimes committed by him. Certainly, some events can "stand out" of the general flow because of their spontaneity or external factors. However, based on the previous principle, such manifestations can be leveled. In RICAS, the search for hidden patterns is carried out based on the intellectual core of semantics processing. Semantic analysis is fundamental, because connections are not always constructed explicitly and they should be sought in context.

C. The Principle of Semantic Integrity

Often, all events and behavioral profiles of criminals are described verbally. Therefore, the intelligent module for semantics processing is fundamental. It reveals the wide possibilities for analysis of hidden regularities and contextual search. Undoubtedly, starting from a variety of computer "handwriting", the module is calibrated with a variety of dictionary correspondences.

D. The Principle of Visual Analysis

Making a decision by an analyst is fundamental in RICAS therefore, the visual component is very important. All links between subjects and objects are displayed visually and on a geoinformation substrate, with classification attributes and analytical data taken into account.

E. Multiplatform and simplicity

The system was developed using modern, optimized technologies in the web space. It can be used on any stationary and mobile devices in the presence of a secure communication channel.

V. AUTOMATIC SEARCHING FOR HIDDEN PATTERNS

Following semantic analysis, RICAS uses the algorithm for searching hidden patterns and generates a list of the most probable hypotheses that are difficult to detect visually or using standard statistical methods. The algorithm is based on definition of relationship between pairs of events and naturally stems from the set of fixed pairs records "time point" - "event (action)".

Criminal database contains information about time moments [1, Nt], in each of which there can be some event from the set of permissible ε. Therefore, each type of event can be associated with a set of time points TS (A):

$$TS(\mathbf{A}) = \{T_{\mathbf{A},1}, \ldots, T_{\mathbf{A},N_\mathbf{A}}\}, \quad \mathbf{A} \in \mathcal{E}, \quad 0 \leqslant T_{\mathbf{A},i} \leqslant N_t \quad (i = 1, \ldots, N_\mathbf{A})$$

During the search for regularities in the flow of events, the system is interested in relationship between distributions of individual events. Behavior pattern is characterized by appearance of its inherent components in the same order, moreover, time intervals separating components in practice are approximately same.

We say that events A and B are related by the ratio of critical interval (CI), if occurrence of event A at time t, there exists an interval [t + d1, t + d2], (d2 ≥ d1 ≥ 0) containing B more often than this is expected from assumption of events independence. This relationship is denoted as A [d1, d2] B, or briefly, (AB).

We will expand the notion "more often than expected" used by the algorithm. Let NA and NB be the number of occurrences of A and B, respectively, during [1, Nt]. P (A) = NA / Nt is probability of occurrence of event A at some point in time; P (¬A) = 1 - P (A). P (¬A) d is the probability that A appears during any interval [d1, d2], (d = d2 - d1 + 1), of length d. The probability of observing A over an interval of length d one or more times is 1-P (¬A) d. Assuming the hypothesis of an independent distribution of events, event B is contained in the interval of length d, after event A, NA × (1 - P (¬B) d) times.

$$\rho = P(\geqslant N_{\mathbf{AB}}) = 1 - P(< N_{\mathbf{AB}})$$

- is the a priori probability that NAB from NA intervals contain occurrences of B. It is obvious that P (<NAB) is distributed according to the binomial law, where NA is number of "tests", 1-P (¬B) d is probability of "success". Consequently,

$$\rho = P(\geqslant N_{\mathbf{AB}}) = 1 - \sum_{i=0}^{N_{\mathbf{AB}}-1} C_{N_{\mathbf{A}}}^{i} (1 - P(\neg \mathbf{B})^{d})^{i} P(\neg \mathbf{B})^{N_{\mathbf{A}}-i}$$

The obtained probability ρ is compared with threshold value α, which is structural parameter of the search; if ρ ≤ α, then given interval is recognized as critical, and the sequence of events corresponds to the pattern.

VI. Conclusions

The proposed model for building software based on principles that correspond to the problematics of criminal analysis is not exhaustive and undoubtedly requires further introduction of Data Mining automated methods. The given example of practical implementation searching algorithm shows the effectiveness of such process.

References

[1] Bodyanskiy Ye.V., Strukov V.M., Uzlov D.Yu. Zadacha otsenki blizosti mnogomernykh obyektov analiza dannykh [The problem of evaluating the proximity of multidimensional objects of data analysis]. USiM, 2016, No.6, pp. 67-72.

[2] Bodyanskiy Ye.V., Strukov V.M., Uzlov D.Yu. Obobshchennaya metrika v zadache analiza mnogomernykh dannykh s raznotipnymi priznakami [The generalized metric in the problem of analysis of multidimensional data with different types of characteristics]. Zbirnyk naukovykh prats Kharkivskoho natsionalnoho universytetu Povitryanykh Syl [Collection of research papers of Kharkiv National University of Air Forces], 2017, Vypusk [Issue] 3(52), pp. 98-101.

[3] Lande D. V. Internetika: Navigatsiya v slozhnykh setyakh: modeli i algoritmy [Internetics: Navigation in complex networks: models and algorithms]. D. V. Lande, A. A. Snarskiy, I. V. Bezsudnov. M.: Librokom (Editorial URSS), 2009, 264 p.

[4] Manning, P. K. (2001). Technology's ways: Information technology, crime analysis and the rationalizing of policing. Criminology and Criminal Justice, 1(1), 83-103.

[5] Manning C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. – Cambridge : Cambridge University Press, 2008. – 544 p.

[6] Taylor, B., Kowalyk, A., & Boba, R. (2007). The integration of crime analysis into law enforcement agencies: An exploratory study into the perceptions of crime analysts. Police Quarterly, 10(2), 154-169.

[7] Westphal C. Data Mining for Intelligence, Fraud and Criminal Detection. Advanced Analytic & Information Sharing Technologies / C. Westphal. – New York : CRC Press Taylor & Francis Group, 2009. – 440 p.

The paper's title, abstract and are not reflects the content and purpose of the article.	We wrote this article not for describing all range of Data Mining methods, which are used in world practice of Criminal Analysis. We use some of them in our practical everyday and this paper is our experience of using it.
What is "The operative part of the article"?	The operative part of the article is ¨IV. FUNDAMENTALS OF SOFTWARE DEVELOPMENT¨: this is the part, which ¨outlines the basic principles, models and technologies that specialized software is used to support the analytical work of criminal analysts¨.
The purpose of the paper is not formulated.	We purpose to use wide on practice ¨...methodology of data mining¨ in criminal analysis in our country as like as it's used in world practice. And we try do it so and we talk about our efforts.
The chapter III. DOMAIN PROBLEMATICS is not describe the science or practical problematic.	Data mining methods is practical realization of criminal analysis and on this stage of intelligence-led policing process we have two fundamental questions: what should be the model that defines the space-time relations and how is it possible to use these models directly for analysis and what mathematical apparatus is best for this. This is our practical problematics.
What is mean P (¬A)?	P (¬A) = 1 - P (A)
The parameters of formulas (p.4) are not defined.	Chapter ¨V. AUTOMATIC SEARCH FOR HIDDEN PATENTS¨ describes the theory of implementing a practical algorithm, so the parameters of formulas logically follow from the narrative.
It is no implementation.	We use Real-time Intelligence crime analytics system in our practice of Criminal Analysis. This system allow to use many Data Mining methods and visualize result of this analysis, therefore we think it's good example of implementation.
The English should be improve.	We do our best and promise to try harder.
Self-citation must be less 30%.	Ok, see new version of article.