Universiteit Leiden
The Netherlands

Algorithmic tools for data-oriented law enforcement

Cocx, T.K.

Citation

Cocx, T. K. (2009, December 2). *Algorithmic tools for data-oriented law enforcement*. Retrieved from https://hdl.handle.net/1887/14450

Version:	Corrected Publisher’s Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/14450

Note: To cite this publication please use the final published version (if applicable).

Chapter 1

Introduction

In the wake of the data explosion of the late 1990s a research area has evolved from statistics and computer science. The main goal of this form of computer guided data analysis, known as *Data Mining* [19] or *Knowledge Discovery in Databases* (KDD), is to extract knowledge from an, often large, collection of data, combining elements of statistics [8], database technology, machine learning [2], artificial intelligence [15, 9, 10] and visualization, within its varying approaches.

The increase in capabilities of information technology of the last decade has led to a large increase in the creation of data [11], as a by-product of corporate and governmental administration or resulting from scientific analyses. Although most of this data has a purpose of its own, for example customer management, tax refund declaration archives and DNA analysis, data mining software aims to aggregate a higher level of information, knowledge, from this data, through the automatic discovery of (underlying) patterns, behavioral classes or (communication) networks. Respectively, convenient knowledge can be gained about customer behavior, tax evasion patterns or discriminating DNA strings that define human biological disorders. Naturally, the algorithms compiled for such purposes are often easily transferable to other domains with the purpose of performing similar tasks.

One of these potential application domains is that of law enforcement. As a an authority that is strictly governed by judicial constraints, the adaptation from paper based archives to a digitally oriented data-infrastructure has been relatively slow, but in recent years, especially since the tragic events of 9/11, law enforcement agencies have begun to bridge this gap by investing more resources in suitable and uniform information systems [12]. Just like in other areas, this change has led to an abundance of data, that is probably suitable for data mining purposes. This thesis describes a number of efforts in this direction and reports on the results reached on the application of its resulting algorithms on actual police data. The usage of specifically tailored data mining algorithms is shown to have a great potential in this area, which forebodes a future where algorithmic assistance in “combating” crime will be a valuable asset.

1.1 Data Mining

Within the area of computer guided statistical data analysis a clear distinction is made between data and knowledge, but the relation between these two types of information is subtle. Although it is clear that data in itself has great value within almost all aspects of modern society, most data collected remains unused as it is never called upon by its creator, its usability depending on independent and very specific cases. For example customer records are only relevant if this particular customer has a complaint about one of the transactions and since the majority of transactions happen without any problems, these records can lie dormant forever. Also, data in itself is sometimes irrelevant if no lessons can be learned from it. For example, obtaining the entire human genome can be quite useless if researchers were unable to use this data for the identification or prevention of disorders. Knowledge, being a result of learning and reasoning on certain perception, has an inherit value and is often usable for progression. Taking this into account the relation can be described as data being the building blocks from which knowledge can be derived.

Obtaining knowledge from perception has been an activity performed by humans alone, for a long time, but the ever growing amount of data, beyond the limits of human perception, and the improvements in information technology have opened the door for computer involvement in this process as both a possibility and a necessity. This process of computer guided data analysis consists of several sub-steps:

1. Data collection
2. Data preparation
3. Data analysis
4. Result visualization

There is some general confusion about the usage of the term data mining to describe either this entire process [19] or the third sub-step alone [6], however, this distinction is only of interest in theoretical discourse, rather than being relevant in practice: If wrong or incomplete data is collected, it is prepared poorly or the results are not or counter intuitively visualized to the user, the data analysis efforts are hardly relevant. Therefore, data mining is usually seen as the process that “extracts implicit, previously unknown and potentially useful information from data” [19].

Naturally, data exists in different forms, but in the scope of data mining, one is limited to data stored in digital form. However, a distinction can still be made, within this subcategory, between structured data and data that is stored without any constraints in an unstructured way. Usually and preferably, corporate data is structured into databases that allow for easy manipulation and searching through the association of an object with attribute-value pairs. Data stored in such a way is highly structured and allows for easy data collection and preparation. A drawback to this approach is that data that does not adhere to this predefined structure is stored with difficulties, including empty attributes or mismatches in attribute types. Next to relational database oriented storage, a lot of information is stored in an unstructured way, like, for example, websites, policy documents, meeting transcripts, etc. Data mining methods that deal with this kind of information first attempt to structure this data in some way before any analysis can be successfully employed. Naturally, mining unstructured data is a harder endeavor than mining structured data. Information stored in video, i.e., video and photographs, is also considered to be unstructured data, but its analysis falls out of the scope of this thesis.

There are two main goals that can be distinguished within the area of computer guided data analysis, that of *descriptive* and that of *predictive* data mining. An analyst using a predictive data mining system usually strives for a model that is able to predict certain attributes of a yet unknown "item" based upon some classifier that was constructed through analysis of items that were known at that time. Such a classifier decides on the value of the attribute to be predicted based upon other attributes that are known, usually employing some kind of *divide-and-conquer* principle [14]. Often, such classifiers are represented by decision trees. In contrast with the predictive approach, a descriptive algorithm tries to discover covering knowledge about a certain (application) domain, dealing with commonalities between items or attributes, rather than their distinguishing properties. Such algorithms typically yield correlations that occur often (have a high *support*) within a certain data set that is representative for a certain domain and hopefully have meaning in the real world, e.g., suburban children are often over-weight. The results of descriptive data mining are diverse and have various visualization possibilities, the most common being a clustering, that visually separates certain subgroups based upon some mutual distance, a list of relations between attribute values, that signifies which attributes appear *frequently* together, and a link or network analysis that draws a network of people based upon, for example, a list of phone calls or emails, to identify central figures in a social environment. Given these subdivisions, Figure 1.1 provides an rough impression of the research area of data mining.

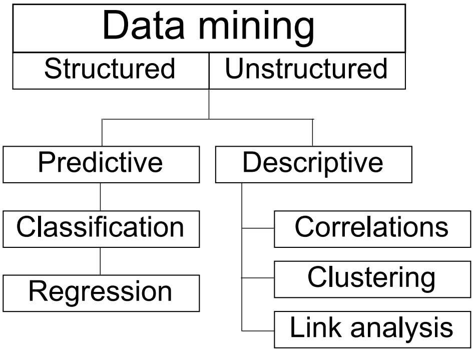

Figure 1.1: Data mining ontology

Most of the different types of data mining have been used throughout this thesis and are therefore relevant to the area of law enforcement to some extent.

1.2 Law Enforcement

The area of law enforcement is very broad in origin and involves all kinds of agencies that assist governing bodies at any level of society to enforce their laws on its populace. Examples of these are among others: military police, religious police, customs, secret police, tax authorities, police, etc. Some of these organizations are subject of law enforcement themselves through "quality control" agencies that check them on internal compliance with the law. However, these agencies are most often placed outside the law enforcement definition and are referred to as "Internal Affairs" or "Professional Standards". Also, the justice department and its courts are usually seen as a separate entity in this area.

One division that can be made in this rather large collection of governmental activity is the type of people they deal with; while some of the mentioned organizations concern themselves with all people in a certain society, e.g.: tax authorities, most of the organizations only deal with people who actively break the law, e.g.: police. Therefore, a division can be made between administrative, more passive agencies and investigative, more active agencies, although most administrative organizations also have an investigative branch. This thesis focuses on the investigative agencies or branches.

Within this subsection of law enforcement, another division can be made, also on the level of activity. The most active divisions are occupied with *tactical law enforcement*, the effectuation of agency policies in the direct combat of crime, dealing with crime investigations, forensic research and the preparation of criminal cases through its litigation branch, i.e.: the District Attorney. Part of these organization, however, deal with the formation of such policies, like, investigative focus points, division of agency personnel to specific tasks, etc., based upon knowledge of criminal activity within a legislative entity, These activities are seen as *strategic law enforcement*. This view of law enforcement is shown in in Figure 1.2.

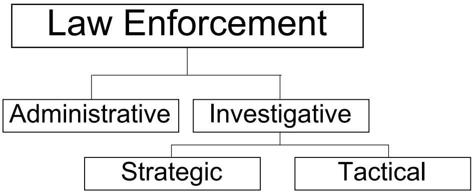

Figure 1.2: The law enforcement area

This thesis is split in two parts, Part I dealing with strategic operations and Part II dealing with the tactical area of policing.

1.2.1 Data on Criminal Activities

Naturally, the agencies mentioned above create a lot of data, exercising their tasks, for example, telephone tapping records, narrative reports, criminal records, and case data. It is clear that most of this data is unstructured, being either recorded (partly) in plain text or not recorded digitally at all. Also all data gathered in the process of law enforcement is governed by very strict preservation and privacy laws. This effectively means, within most legislations, that data can be saved for a very limited time, usually only for the duration of a certain investigation, it can only be used for purposes within that investigation and can only be handled by individuals who are involved in that investigation. Naturally, both issues negatively affect any data mining effort greatly.

Next to that fact, knowledge gained from statistical analysis, of which data mining is a part, has no actual power as evidence in most courts of law, limiting the usefulness of any knowledge from these efforts to the investigation itself, where police officers are still required to gather more “true” evidence.

Also, it is debatable if data mining can play a role as predictive tool in any investigation. Most legislations allow for search warrants and arrests only if certain locations or people can undeniably be linked to a certain crime. The nature of predictive data mining, however, is to transfer knowledge learned from the many to the one, accompanied with a certain reliability percentage, but never in relation to a specific crime. For example, the chance that each of 50 individuals are drug dealers may be 95%, which is considered to be highly reliable, but law enforcers are not allowed to search all these people’s homes every time a drug deal goes down. Because of this discrepancy, predictive analysis is not always allowed. leading to “tunnel vision” and illegitimate warrants or arrests. Some of these issues, relevant in the Netherlands, are discussed in [16, 17]. Due to these existing limits on the adoption of data mining in investigative law enforcement, a larger part of all efforts is being directed toward strategic rather than to tactical law enforcement purposes.

1.2.2 Related Work on Data Mining for Law Enforcement

Despite these drawbacks, the number of data mining projects in the law enforcement area is now slowly increasing, both in- and outside of the academic world. Commercial players vary from very small to very large multi-nationals, like statistical software producer SPSS. One of the first large-scale academic projects is the COPLINK project in Arizona where some excellent work has been done in the field of entity extraction from narrative reports [4], the exploitation of data mining for cooperation purposes [3] and social network analysis [20, 5]. In the often mentioned FLINTS project, soft (behavioral) and hard (fingerprints, DNA) forensic evidence was combined to give analysts the ability to build a graphical image of (previously unthought-of) relations between crimes and criminals. Another link-analysis program, FinCEN [7], aimed to reveal money laundering networks by comparing financial transactions. Also, Oatly et al. did some link analysis work on burglary cases in the OVER project [13]. Clustering techniques have also been employed in the law enforcement area. Skillicorn [18] did some work on the detection of clusters within clusters to filter the surplus of information on possible terrorist networks and present law enforcement personnel with a viable subset of suspects to work with. Adderly and Musgrove [1] applied clustering techniques and Self Organizing Maps to model the behavior of sex-offenders. This thesis aims to augment the existing palette of algorithms through the incorporation of new insights into this area.

1.2.3 Motivation

This thesis results from a cooperation between the Dutch National Police (KLPD) and the academic world and the belief that a number of goals set in the area of combating crime can be reached with the assistance of computer guided data analysis, especially where it concerns unstructured or semi-structured data sources. It is believed that fundamental research driven by questions arising from law enforcement practice can lead to a long-term data mining framework centered around the themes knowledge engineering and learning, which is demonstrated by the research described in this thesis.

1.2.4 Overview

This thesis is divided into two parts: one part about algorithms that can be employed for strategic purposes and one part about applications in the tactical domain.

Chapter 2, being the first chapter of the strategic part, describes a first analysis of a large database of criminal records (cf. Appendix B), that aims to relate different crimes to each other or to demographic data, based upon frequent co-occurrence in a record. To accomplish this, the existing and well-known Apriori algorithm was adapted to search for this type of connections in this specific database, incorporating solutions to a variety of problems native to data on criminal activities. This file, that was available in an anonymized version, was also used for more fundamental analyses.

Because this file contains an enormous amount of "raw" data, standard methodologies to visualize data are often not very well suited to this task. Chapter 3 therefore describes a method that optimizes the visualization of relations between criminals in this database by using the domain knowledge of the analyst as a vital part of the clustering phase. Because this expert is able to directly manipulate a physical representation of this data, a high quality of visualization can be reached with a minimum amount of computational effort.

An important concept that can be evaluated using the criminal record database is that of the *criminal career*, which can be seen as a temporally ordered series of crimes committed by an individual throughout his or her life. An ad-hoc method is suggested in Chapter 4 that uses four important factors of such a career to calculate distances between different careers. These distances can then be visualized in a two-dimensional clustering. Chapter 5 proposes a number of enhancements of this method, that are proven to be functional in another domain. Next to that, some methods are discussed that could eventually lead to a prediction of new careers, further examined in Part II.

After the completion of a clustering and classifying system, a search for subcareers that occur often can be performed. An even more noteworthy endeavor is to find specific subcareers that are common in one class and are not in all others, taking the role of defining subcareers that can be used to identify certain classes. These opportunities are researched in Chapter 6, where an existing method for market basket analysis is adapted to suit the demands put forward by the search for common subcareers.

In the second part about tactical applications, the possibilities for predicting criminal careers are discussed in Chapter 7, where a method is described that employs the power of a visualization to create reliable predictions through simple mathematical calculations.

This method is effectuated, expanded en tailored towards criminal data in Chapter 8, where the different variables of this method are tested on the actual data. Under certain conditions, this method can predict criminal careers with a high accuracy.

In Chapter 9, an investigation is described that strives to answer the question if files from confiscated computers from crime scenes can be an indication of which of these scenes are related to the same criminal organizations. For this purpose, a specific distance measure was developed that determines the chance that two computers were owned by the same organization. For this project, text mining software was employed that extracted special entities from computers retrieved from synthetical drugs laboratories.

Chapter 10 describes how “online predators”, child sexual abusers on the internet, can be recognized automatically on social networking sites, like the Dutch “Hyves”. A genetic algorithm was designed that automatically selects groups that show a significant difference between predators and regular users in the amount of under-aged “friends” on their respective profiles. It turns out that in some specific cases this variable can be a strong indicator for danger classification of certain user groups.

This thesis ends in Appendix A with some considerations about statistics, law and privacy that play a pivotal role for everybody using, or intending to use (parts of) our work in the daily practice of police matters. It discusses the applicability, statistical relevance and insightfulness of and reservations to our methods in general and police usage in specific, focusing mostly on the methods in Part II, that deal with tactical policing. As an assurance our methods are viewed in the correct context and our tools are used in a concise way, a deliberation on their possibilities and limitations for use within society is both important and natural.

Bibliography

[1] R. Adderley and P. B. Musgrove. Data mining case study: Modeling the behavior of offenders who commit serious sexual assaults. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 215–220, 2001.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science, 2006.

[3] M. Chau, H. Atabakhsh, D. Zeng, and H. Chen. Building an infrastructure for law enforcement information sharing and collaboration: Design issues and challenges. In *Proceedings of The National Conference on Digital Government Research*, 2001.

[4] M. Chau, J. Xu, and H. Chen. Extracting meaningful entities from police narrative reports. In *Proceedings of The National Conference on Digital Government Research*, pages 1–5, 2002.

[5] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey, and Z. Huang. COPLINK: Visualization for crime analysis. In *Proceedings of The National Conference on Digital Government Research*, pages 1–6, 2003.

[6] M. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice-Hall, 2003.

[7] H.G. Goldberg and R.W.H. Wong. Restructuring transactional data for link analysis in the FinCEN AI system. In *Papers from the AAAI Fall Symposium*, pages 38–46, 1998.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer, 2001.

[9] M. Hutter. *Universal Artificial Intelligence; Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

[10] G.F. Luger. *Artificial Intelligence; Structures and Strategies for Complex Problem Solving*. Pearson Education, 6th edition, 2009.

[11] P. Lymand and H.R. Varian. How much information? Technical report, Berkeley, 2004.

[12] J. Mena. *Homeland Security; Techniques and Technologies*. Charles River Media, 2004.

[13] G.C. Oatley, J. Zeleznikow, and B.W. Ewart. Matching and predicting crimes. In *Proceedings of the Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence (SGAI2004)*, pages 19–32, 2004.

[14] J.R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan-Kaufmann, 1993.

[15] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2003.

[16] B. Schermer. *Software Agents, Surveillance, and the Right to Privacy: A Legislative Framework for Agent-enabled Surveillance*. PhD thesis, Leiden University, 2007.

[17] R. Sietsma. *Gegevensverwerking in het kader van de opsporing; toepassing van datamining ten behoeve van de opsporingstaak: Afweging tussen het opsporingsbelang en het recht op privacy*. PhD thesis, Dutch, Leiden University, 2007.

[18] D.B. Skillicorn. Clusters within clusters: SVD and counterterrorism. In *Proceedings of the Workshop on Data Mining for Counter Terrorism and Security*, 2003.

[19] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2nd edition, 2005.

[20] Y. Xiang, M. Chau, H. Atabakhsh, and H. Chen. Visualizing criminal relationships: Comparison of a hyperbolic tree and a hierarchical list. *Decision Support Systems*, 41(1):69–83, 2005.