Enhancing the Automated Analysis of Criminal Careers*

Tim K. Cocx Walter A. Kosters Jeroen F.J. Laros†

Abstract

Four enhancements have been devised, both on a semantic and efficiency level, that improve existing methods of automated criminal career analysis. A new distance measure was introduced that more closely resembles the reality of policing. Instead of previous, more rigid, comparison of career changes over time we propose an alignment of these careers. We employ a faster and better method to cluster them into a two-dimensional plane. This visualization can ultimately be used to predict the unfolding of a starting career with high confidence, by means of a new visual extrapolation technique. This paper discusses the applicability of these new methods in the field and shows some preliminary results.

1 Introduction

The amount of data being produced in modern society is growing at an accelerating pace. New problems and possibilities constantly arise from this so-called data explosion. One of the areas where information plays an important role is that of law enforcement. Obviously, the amount of criminal data gives rise to many problems in areas like data storage, data warehousing, data analysis and privacy. Already, numerous technological efforts are underway to gain insights into this information and to extract knowledge from it.

As a result from this efforts, in [10] we described research attempting to gain insights into the concept of criminal careers: the criminal activities that a single individual exhibits throughout his or her life. The resulting tool analysed the Dutch national criminal record database: an annually compiled source that extracts information from digital narrative reports stored throughout the individual departments. This method mainly addressed the extraction of the four important factors (see Section 2) in criminal careers and established an overview picture on the different existing types of criminal careers by using a stepwise or year-based approach, with the ultimate objective to prepare the data for prediction of a new offender's career. The approach centered on formalising an individual's *criminal profile* per year, representing an entire career as a string of these calculated profiles. The paper identified some difficulties in comparing these strings and provided solutions to them. The method, however, suffered from time-complexity issues and a comparison mechanism that was largely rigid and ad hoc.

In this paper we describe a number of techniques that are widely applicable but were specifically developed for the enhancement of the above mentioned analysis. We explain how these methods can be used to improve the existing paradigm by replacing the individual ad hoc methodologies and how combining them will reduce the number of steps needed to reach even better results. We supplement the existing approach by adding a preliminary prediction engine that fully employs the properties of the already realised visualisation. Experiments performed with this setup are also discussed.

The main contribution of this research lies in the novel combination of a number of separately created technologies, all very well suited for law enforcement purposes, to pursue the common goal of early warning systems that allow police agencies to prevent the unfolding of possibly dangerous criminal careers. For a detailed overview, see Section 2.2.

Four enhancements have been devised. In Section 3 a new distance measure is introduced that more closely resembles the reality of policing. Instead of previous, more rigid, comparison of career changes over time we propose an alignment of these careers in Section 4. In Section 5 we employ a faster and better method to cluster them into a two-dimensional plane. This visualization can ultimately be used to predict the unfolding of a starting career with high confidence, by means of a new visual extrapolation technique (see Section 6). Section 7 shows experiments, and Section 8 concludes.

2 Background and overview

The number of data mining projects in the law enforcement area is slowly increasing. Both inside and outside of the academic world large scale projects are underway. In this section we discuss related work and provide an overview of our approach.

*Financed by the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430

†Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands, tcocx@liacs.nl

2.1 Related work One of the larger academic projects, known as COPLINK, is a police-university collaboration in Arizona, where work has been done in the exploitation of data mining for cooperation purposes [4], the field of entity extraction from narrative reports [5], and social network analysis [6,17]. The FLINTS project and FinCEN [11] aim at revealing links between crimes and criminals and to reveal money laundering networks by comparing financial transactions. Also, Oatly et al. [16] linked burglary cases in the OVER project. Clustering techniques are widely used in the law enforcement arena as well, like for example by Adderly and Musgrove [1], who applied clustering techniques and Self Organizing Maps to model the behavior of sex-offenders, and by Cocx et al. [7, 8, 10] who made an attempt at clustering criminal investigations to reveal what offenses were committed by the same group of criminals, revealed links between crimes and demographic data and the above mentioned automated analysis of criminal careers.

Criminal careers have always been modeled through the observation of specific groups of criminals. A more individually oriented approach was suggested by Blumstein et al. [2]: little definitive knowledge had been developed that could be applied to prevent crime or to develop efficient policies for reacting to crime until the development of the criminal career paradigm. A criminal career is the characterization of a longitudinal sequence of crimes committed by an individual offender. Participation in criminal activities is obviously restricted to a subset of the population, but by focusing on the subset of citizens who do become offenders, they looked at the *frequency*, *seriousness* and *duration* of their careers.The original paper also focused criminal career information on the *nature* of such a career and employed data mining techniques to feed this information back to police analysts and criminologists.

2.2 Overview Originally the four factors were extracted from the criminal record database, recorded as numbers and treated as such. However, it appears to be beneficial to treat a collection of crimes in a single year as a *multiset*, which then describes severity, nature and frequency inherently. A multiset or *bag*, is a collection where each element can occur more than once. The set of all distinct elements in that multiset is called its *underlying set*. Figure 1 describes the relation between the two and shows how we employ it to represent a criminal's activities in a single year.

The multiset representation offers advantages, most notably the availability of standard approaches to compare multisets and calculate distances between them. Kosters and Laros [12] devised a distance function for

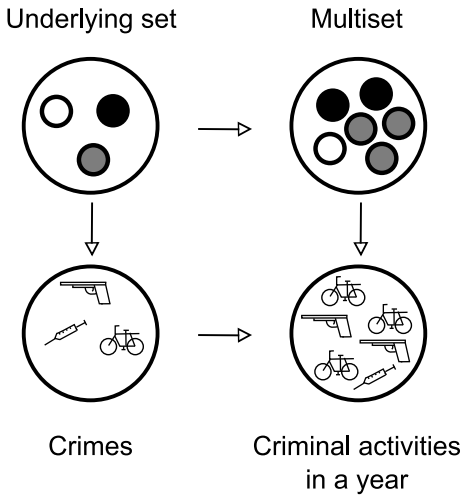

Figure 1: A multiset representation of a criminal profile in a single year

multisets that generalizes well-known distance measures like the *Jaccard distance*. This metric contains a customizable function f that can be adapted to fit specific knowledge domains. It also allows for the incorporation of weights for an element, e.g., element x counts twice as much as element y. We employ this metric for calculating the difference between two crime-multisets (see Section 3) by choosing a specific f.

Instead of the former method of a strict number-wise comparison between years (comparing the first year of criminal a with the first year of criminal b, the second year of a with the second year of b, etc.), with the possibility of stretching or shrinking careers, we propose a novel *alignment* of the mentioned multisets. This method strives for an optimal automated matching of years, using the distance measure described above, dealing penalties for every mutation needed, which enables a police analyst to better cope with situations like captivity, forced inactivity or unpenalized behaviour. Section 4 elaborates on this.

Clustering methods used before yielded very good results, mainly by incorporating direct input from police specialist into its mechanics. It was however computationally complex, taking a very long time to process the entire dataset of 1 million offenders. Within the scope of the research into criminal careers a method was developed that largely improved standard *push and pull* algorithms but eliminated the need for human input, while retaining the former strength of its output [13]. We explain this in Section 5.

Earlier results provided some important information for law enforcement personnel, but the ultimate goal, predicting if certain offenders are likely to become

(heavy) career criminals, was not realized or elaborated upon. Further investigation of this possibility led to the need of a good 2-dimensional visual *extrapolation* system, described in [9]. The conceivement of this technique paved the way for possible development of early warning systems, that we explore in Section 6.

3 A distance measure for profiling

The core of our new approach consists of the new distance measure. This metric is a special case of the *generic metric for multisets* as described in [12]:

$$d_f(X,Y) = \frac{\sum_i f(x_i, y_i)}{|S(X) \cup S(Y)|}$$

This metric calculates the distance between two finite multisets X and Y, where $S(A) \subseteq \{1, 2, \ldots, n\}$ is the underlying set of multiset A and a_i is the number of occurrences of element i in multiset A. Here $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a fixed function with finite supremum M and the following properties:

$$\begin{array}{lcll} f(x,y) & = & f(y,x) & \text{for all } x, y \in \mathbb{R}_{\geq 0} \\ f(x,x) & = & 0 & \text{for all } x \in \mathbb{R}_{\geq 0} \\ f(x,0) & \geq & M/2 & \text{for all } x \in \mathbb{R}_{>0} \\ f(x,y) & \leq & f(x,z) + f(z,y) & \text{for all } x, y, z \in \mathbb{R}_{\geq 0} \end{array}$$

These properties ensure that d_f is a valid metric [12].

Naturally, all results depend largely upon choosing the right *defining function* f for a certain knowledge domain. In the law enforcement area, it is important to realise that the relative difference between occurrences of a crime is more important than the difference alone. This is because the distance between an innocent person and a one-time offender should be much larger than for example the distance between two career criminals committing 9 and 10 crimes of the same kind respectively, thus ensuring that $f(0,1) \gg f(9,10)$ (the two arguments of f are the numbers of respective crimes of a given category, for two persons).

A good candidate for this function, that was developed in cooperation with police data analysts, seems to be the function

$$f_{crime}(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

for integer arguments x and y, both ≥ 0. This function complies with the above mentioned characteristic of criminals and yields a valid metric.

It is obviously important to still be able to incorporate *crime severity* and *nature*, next to crime frequency, into this equation. This is possible through the addition of *weights* to the generic metric described above. A suggestion was made by Kosters and Laros [12] for accomplishing this: each item i gets an assigned integer weight ≥ 1 and is multiplied by this weight in every multiset, simulating the situation where all weights were equal but there were simply more of these items in each set. An impression of how this affects the calculation of distances between criminal activities is given in Figure 2.

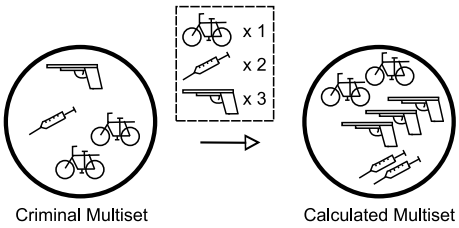

Figure 2: Adding weights to criminal activities

By using this specifically tailored distance measure with weighing possibilities, we are able to calculate distances between criminals per single time-unit. This distance can serve as basis to discover distances between careers, that can be seen as strings of these time-units.

4 Alignment of criminal careers

Since the temporal ordering of crimes is of high importance in the definition of a criminal career, a tool occupied with its analysis should provide a way to calculate distances between these ordered series based upon distances between their elements. Sequence alignment could be a valuable tool in this process.

In contrast with the strict year-by-year comparison used before [10], the alignment paradigm [15] tries to find a best match between two ordered series, allowing a small number of *edits* to be made in such an ordering: insertions, deletions and other simple manipulations. The penalty for a deletion or insertion is governed by a so-called *gap-penalty function*. Alignment is an algorithm typically used within the area of computational biology. Famous instances include those of Needleman-Wunsch and Smith-Waterman. Each alignment is based upon a valid metric for elements (a year of accumulated crimes in this case), as for example the metric discussed in Section 3. Figure 3 describes a typical alignment situation, showing two such edits. The treatment of gaps, however, may somewhat differ from biological applications; also note that empty multisets (an "innocent" year) have a special relation with the gap penalty. Next to these differences, careers may or may not be "complete"; some of them are still unfolding. These issues and their relation to classical alignment will be addressed in a forthcoming paper.

One of the rationales for using alignment instead

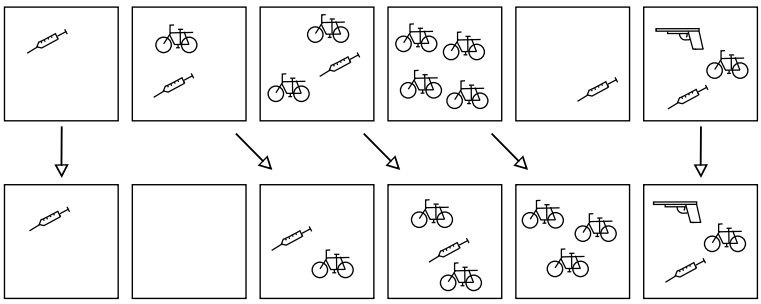

Figure 3: Two criminal careers who's similarity is revealed by alignment

of strict sequence matching is the frequent occurrence of gaps or stretches in criminal data. One could think of, for example, prison time, forced inactivity, judicial struggles consuming time between an offense and its penalization and, most importantly, the frequent occurrence of unpenalized criminal activity due to undiscovered crime. Treating this naturally occurring plethora of phenomena without any descriptive value on ones intentions in the criminal area as non-existent or not important will reveal a deformed image of compared criminal careers. Therefore the usage of an alignment mechanism is strongly favored over a strict comparison on a semantic level. Another reason for this might be the occurrence of randomly appearing year changes within ones career, e.g., two crimes occurring either in June and July, or in January and December, are in essence the same, although strict separation between years will place the latter in two different time-units. Altough data is often gathered in such broad time units, leading to the inevitable occurence of this problem, the alignment paradigm could potentially be used to reduce the subseqent effects.

Careful consideration needs to be given to the gap-penalty function in the field of crime analysis. While each gap in standard (biological) alignment procedures represents a constant penalty, a gap in a criminal career can be considered less or more important based upon the amount of time passed within this gap. We therefore assign a time stamp $t(a)$ to each element a in the criminal career. The gap-penalty is then computed by applying the gap-penalty function to the different $\Delta t(u,i) = t(u_{i+1}) - t(u_i)$ involved, where i is an element in u, an ordered, time stamped series. The known algorithms to compute distances have to be adapted, causing an increase in time ($O(n^2) \rightarrow O(n^3)$) and space complexity ($O(n) \rightarrow O(n^2)$), where n is the length of the series.

Using this alignment we compare all possible couples of criminal careers and construct a standard distance matrix of the results.

5 Toward a new clustering

In the original setup, a *push and pull algorithm* [3] was used to create a (sub)optimal 2-dimensional clustering of the produced distances matrix from a randomized starting point. Initially, points get random positions in the unit square of the plane;after which they are repeatedly pulled together or pushed apart, depending on the relation between current and desired distance. This method, a type of Multi Dimensional Scaling, relied upon domain knowledge provided by the analyst using the tool. This domain input realised a significant gain in results compared with simpler, unsupervised algorithms of the same kind [14].

The complexity of the previous algorithm, however, prohibited a time efficient usage of the tool, costing near days on standard equipment to analyse the target database and thus posing a serious bottleneck in the original setup that needed to be overcome, preserving the property that elements can be easily added and traced.

One of the major problems of the simpler variants of the push and pull algorithm was the fact that a large number of randomly positioned individuals tended to create subgroups that pushed a certain item harder away from its desired position in the plane, than comparable items pulled this point toward that position (see Figure 4, standard).

Especially in the case of criminal careers, where analysis tends to result in vastly different, very large clusters, this effect appears to present. Addressing this issue could therefore lead to a representative clustering and good response times, eliminating the need for human input. Kosters and Laros [13] suggested the usage of a *torus* rather than the standard bounded flat surface. Within such a torus, all boundaries are identified with their respective opposite boundaries, enabling the "movement" of an individual from one end of the surface to the opposite end, thus overcoming the above mentioned problem (Figure 4, torus).

When using this torus to construct a 2-dimensional

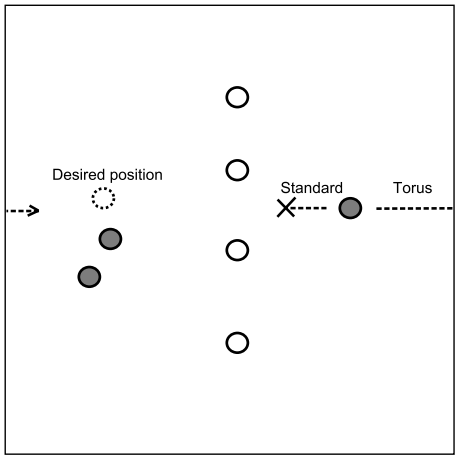

Figure 4: Advantage of torus clustering versus standard clustering

clustering of our distance matrix, time complexity was reduced to the level of simple push and pull algorithms, but the visualisation is expected to be of the same quality as the image produced by the supervised algorithm. Other clustering techniques are also potential candidates for adoption, if they adhere to the demand of incremental addition of single items.

6 Prediction of unfolding careers

A lot of information is inherently present in the visualization of mutual distances. In [9] we propose to utilise the power of a 2-dimensional clustering for temporal extrapolation purposes. This algorithm plots the first, already known, time units of criminal activity of a certain person in the correct place (compared to other fully grown careers already displayed in the image) and extrapolates these points in the plane to automatically discover careers that are likely to become very similar in the future. A system employing this should be able to provide police operatives with a warning when such a starting career can easily develop into that of a heavy career criminal.

According to [9], the best way to extrapolate within the static clustering is to interpolate the plotted time units with a cubic spline and extrapolate with a straight line, having the same derivative as the last part of that spline. The r closest careers to the extrapolation line are then used as reference careers. Naturally, the reference careers closer to the last plotted time unit have higher reliability and will therefore receive a higher weight factor. The selection and weighing process is displayed in Figure 5 (where $r = 9$).

Using the selected and weighted careers, an average amount of all crimes is calculated which in turn assigns

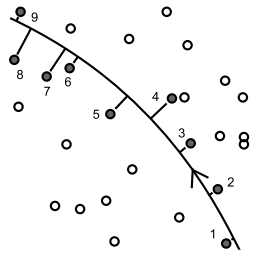

Figure 5: Locating future comparable careers for curved spline extrapolation

a predicted class of criminal activity to the career under evaluation. If this class falls within the severe categories (minor and heavy career criminals), a software alarm could potentially be sounded at police stations.

7 Experiments

We tested our new approach on the actual Dutch National Criminal Record Database. This database contains approximately one million offenders and the crimes they committed. Our tool analyzed the entire set of criminals and presented the user with a clear 2-dimensional clustering of criminal careers as can be seen in Figure 6.

The database used within the experiments is available for scientific research at CBS (Statistics Netherlands) in an anonimized version. It contains approximately one million rows, representing offenders, and 250 columns representing both demographic data and categorized criminal activities. Only the latter part of columns was used during experimentation, resulting in approximately 150 columns, representing the number of crimes per category. Linking this database to an also available database of offenses allowed us to add a date to each offense. From this, a multiset string of offenses was added to each anonimized offender, where each bucket contained crimes committed within a certain time frame.

The image in Figure 6 gives an impression of the center of the outputted torus produced by our tool when analyzing the before mentioned database. This image shows the identification that could easily be coupled to the appearing clusters after examination of its members.

This image appears to be describing reality better than the clustering resulting from [10]. Just like the image constructed by the previous approach (Figure 7) the new image features a large “cloud” of one-time offenders. However a clear distinction can now be noticed between the left, dark, part of that cloud, which represent minor crimes, and the right, lighter side of the

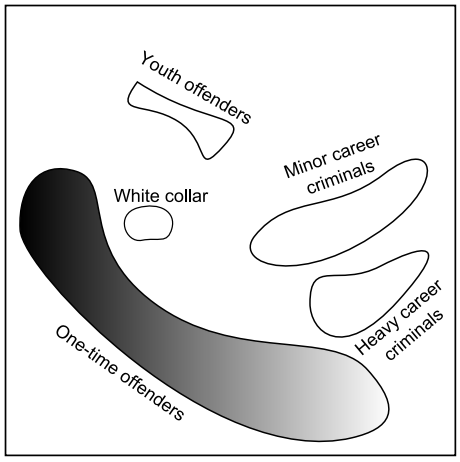

Figure 6: Impression of a clustering of criminal careers

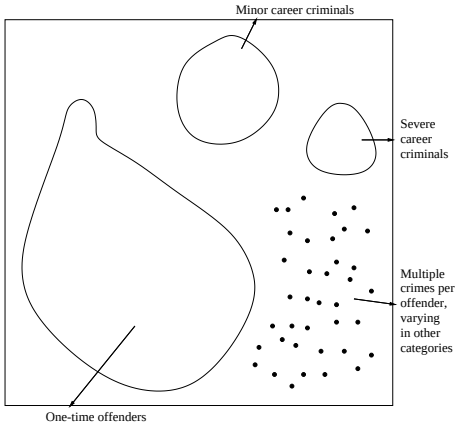

Figure 7: Earlier results

cluster that contains the heavier felonies. Next to this, the group of miscellaneous offenders was split into two, more specific, groups; the white border criminals and youth offenders. The remaining unclustered individuals were left out of the image for clarity reasons. Analysis of this group can however also reveal interesting results, answering why these individuals do not belong to one of the larger clusters and what kind of crimes these people commit. Next to the better results provided by this approach, far better computation times were realized that outperform the previous method by a factor 100.

Getting more insights into the possible existence of subgroups in any cluster remains a desirable functionality of this approach. Future research will focus on getting this improvement realized (cf. Section 8).

A prototype of a criminal career predictor was also constructed. Preliminary results of this tool were discussed in [9]. The suggested method appears to deliver results fast, reaching an accuracy level of 88.7%. The predictor has been submitted to police analysts to reveal possible improvements to the technique employed. It is noteworthy that this prediction method leaves the area of the analysis of large numbers and strives to impose a newly discovered truth on individual cases. Future research should focus on the statistical properties of this shift in paradigm as well as on its inherent privacy issues.

8 Conclusion and Future directions

In this paper we demonstrated the applicability of some newly developed methods in the field of automated criminal career analysis. Each of the enhancements provided a significant gain in computational complexity or improved the analysis on a semantic level. An integral part of the new setup consists of the multiset metric that was specifically tuned toward the comparison of criminal activities. It leaned strongly upon knowledge provided by domain experts. This distance measure was used to search for optimal alignments between criminal careers. This edit distance for sorted or temporal multiset data improved greatly upon the original year-by-year comparison, dealing with problematic situations like captivity or forced inactivity. Using this alignment a distance matrix can be constructed and clustered on a flat surface using a new unsupervised method, yielding comparable results with the previous setup, but largely reducing the time complexity. The power of this visual representation was employed to extrapolate a starting career within the plane of the clustering, predicting its unfolding with some accuracy. The end report consists of a visual 2-dimensional visualization of the results and a prototype of an early warning system that predicts newly developing criminal careers, which is ready to be used by police experts.

This paper describes research in progress, striving to improve on a previous, methodology for the analysis of criminal careers. Building upon a series of successfully introduced algorithms this novel approach is subject to a high number of parameters and leans heavily on the way these techniques work together. It is to be expected that the current setup can also be improved by tweaking its parameters and elaborating on the internal cooperation between segments. Future work will therefore focus on a larger experimental setup to find and verify optimal settings and retrieve results that are even more descriptive and usable. We also hope to equip our tool with a sub-cluster detection algorithm to provide even better insights into the comparability of criminal careers.

It may be of interest to set fuzzy borders between the different years. Crimes within months ending or beginning such a time unit can be (partly) assigned

to the next or previous year respectively as well, thus eliminating the problems arising with strict coherence to the change of calendar year.

Special attention will be directed toward the predictability of criminal careers, and the eventual suitability of this approach for early warning systems running at police headquarters throughout the districts. Future research could focus on modifications of the temporal extrapolation method, more closely meeting the demands of real-life criminal career prediction. Incorporation of this new tool in a data mining framework for automatic police analysis of their data sources is also a future topic of interest.

References

[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, New York, 2001.

[2] A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher. *Criminal Careers and "Career Criminals"*. The National Academies Press, 1986.

[3] J. Broekens, T. Cocx, and W.A. Kosters. Object-centered interactive multi-dimensional scaling: Let's ask the expert. In *Proceedings of the Eighteenth Belgium-Netherlands Conference on Artificial Intelligence (BNAIC2006)*, pages 59–66, 2006.

[4] 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.

[5] 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.

[6] 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.

[7] T. K. Cocx and W.A. Kosters. Adapting and visualizing association rule mining systems for law enforcement purposes. In *Proceedings of the Nineteenth Belgium-Netherlands Conference on Artificial Intelligence (BNAIC2007)*, pages 88–95, 2007.

[8] T.K. Cocx and W.A. Kosters. A distance measure for determining similarity between criminal investigations. In *Proceedings of the Industrial Conference on Data Mining 2006 (ICDM2006)*, volume 4065 of *LNAI*, pages 511–525. Springer, 2006.

[9] T.K. Cocx, W.A. Kosters, and J.F.J. Laros. Temporal extrapolation within a static clustering. Accepted for ISMIS 2008, Toronto.

[10] J.S. de Bruin, T.K. Cocx, W.A. Kosters, J.F.J. Laros, and J.N. Kok. Data mining approaches to criminal career analysis. In *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM 2006)*, pages 171–177, 2006.

[11] 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.

[12] W. A. Kosters and J. F. J. Laros. Metrics for mining multisets. In *Proceedings of the Twenty-seventh SGAI International Conference on Artificial Intelligence SGAI2007*, pages 293–303, 2007.

[13] W.A. Kosters and J.F.J. Laros. Visualization on a closed surface. In *Proceedings of the Nineteenth Belgium-Netherlands Conference on Artificial Intelligence (BNAIC2007)*, pages 189–195, 2007.

[14] W.A. Kosters and M.C. van Wezel. Competitive neural networks for customer choice models. In *E-Commerce and Intelligent Methods, Studies in Fuzziness and Soft Computing 105*, pages 41–60. Physica-Verlag, Springer, 2002.

[15] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965.

[16] 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.

[17] 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.