Concept Discovery Innovations in Law Enforcement

A Perspective

Jonas Poelmans[1], Paul Elzinga[3], Stijn Viaene[1,2], Guido Dedene[1,4]

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
[3]Amsterdam-Amstelland Police, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl

Abstract— In the past decades, the amount of information available to law enforcement agencies has increased significantly. Most of this information is in textual form, however analyses have mainly focused on the structured data. In this paper, we give an overview of the concept discovery projects at the Amsterdam-Amstelland police where Formal Concept Analysis (FCA) is being used as text mining instrument. FCA is combined with statistical techniques such as Hidden Markov Models (HMM) and Emergent Self Organizing Maps (ESOM). The combination of this concept discovery and refinement technique with statistical techniques for analyzing high-dimensional data not only resulted in new insights but often in actual improvements of the investigation procedures.

Keywords-Formal Concept Analysis; Intelligence Led Policing; knowledge discovery

I. INTRODUCTION

In many law enforcement organizations, more than 80 % of available data is in textual form. In the Netherlands and in particular the police region Amsterdam-Amstelland the majority of these documents are observational reports describing observations made by police officers on the street, during motor vehicle inspections, police patrols, interventions, etc. Intelligence Led Policing (ILP) aims at making the shift from a traditional reactive intuition-led style of policing to a proactive intelligence led approach [2]. Whereas traditional ILP projects are typically based on statistical analysis of structured data, e.g. geographical profiling of street robberies, we go further by uncovering the underexploited potential of unstructured textual data.

Formal Concept Analysis (FCA), a mathematical unsupervised clustering technique originally invented by Wille [6] offers a formalization of conceptual thinking. The intuitive visualization of concept lattices derived from formal contexts has had many applications in the knowledge discovery field [9, 17]. Concept discovery is an emerging discipline in which FCA based methods are used to gain insight into the underlying concepts of the data. In contrast to standard black-box data mining techniques, concept discovery allows to analyze and refine these underlying concepts and strongly engages the human expert in the data discovery exercise. The main goal is to make previously inaccessible information available in a for practitioners easy to interpret visual display. In this paper we report on our recently finished and ongoing research projects on concept discovery in law enforcement.

The remainder of this paper is composed as follows. In section 2 we give a short introduction on some of the techniques we used such as FCA, Hidden Markov Models (HMM), Emergent Self Organizing Maps (ESOM) and Temporal Concept Analysis (TCA). Section 3 discusses the recently finished domestic violence, human trafficking and terrorism case studies. Section 4 gives an overview of some projects in initial stages of development. Section 5 concludes the paper.

II. BACKGROUNDER

A. *Formal Concept Analysis*

Formal Concept Analysis [7] is a data analysis technique that supports the user in analyzing the data and discovering unknown dependencies between data elements. In particular, the visualization capabilities are of interest to the domain expert who wants to explore the information available, but at the same time has not much experience in mathematics or computer science. The details of FCA theory and how we used it for KDD can be found in [15]. Traditional FCA is mainly using data attributes for concept analysis. We also used process activities (events) as attributes [16]. Typically, coherent data attributes were clustered to reduce the computational complexity of FCA.

B. *Temporal Concept Analysis*

Temporal Concept Analysis (TCA) is an extension of traditional FCA that was introduced in scientific literature about nine years ago [8]. TCA addresses the problem of conceptually representing time and is particularly suited for the visual representation of discrete temporal phenomena.

The pivotal notion of TCA theory is that of a conceptual time system [8]. In the visualization of the data, we express the "natural temporal ordering" of the observations using a time relation R on the set G of time granules of a conceptual time system. We also use the notions of transitions and life tracks. The basic idea of a transition is a "step from one point to another" and a life track is a sequence of transitions.

C. *Emergent Self Organising Maps*

Emergent Self Organizing Maps (ESOM) [10] are a special class of topographic maps. ESOM is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of its structure [12]. Topographic maps perform a non-linear mapping of the high-dimensional data space to a low-dimensional one, usually a two-dimensional space, which enables the visualization and exploration of the data. ESOM is a more recent type of topographic map. According to Ultsch, "emergence is the ability of a system to produce a phenomenon on a new, higher level". In order to achieve emergence, the existence and cooperation of a large number of elementary processes is necessary. An emergent SOM differs from a traditional SOM in that a very large number of neurons (at least a few thousands) are used [11]. In the traditional SOM, the number of nodes is too small to show emergence.

D. *Hidden Markov Models*

A Hidden Markov Model (HMM) is a statistical technique that can be used to classify and generate time series. A HMM [20] can be described as a quintuplet $I = (A, B, T, N, M)$, where N is the number of hidden states and A defines the probabilities of making a transition from one hidden state to another. M is the number of observation symbols, which in our case are the activities that have been performed to the patients. B defines a probability distribution over all observation symbols for each state. T is the initial state distribution accounting for the probability of being in one state at time $t = 0$. For process discovery purposes, HMMs can be used with one observation symbol per state. Since the same symbol may appear in several states, the Markov model is indeed "hidden".

We visualize HMMs by using a graph, where nodes represent the hidden states and the edges represent the transition probabilities. The nodes are labelled according to the observation symbol probability.

III. Finished Concept Discovery Projects at the Amsterdam-Amstelland Police

A. *Domestic violence*

In 1997, the Ministry of Justice of the Netherlands made its first inquiry into the nature and scope of domestic violence [14]. It turned out that 45% of the population once fell victim to non-incidental domestic violence. For 27% of the population, the incidents even occurred on a weekly or daily basis. These gloomy statistics brought this topic to the centre of the political agenda.

In the domestic violence case study we found that FCA concept lattices were particularly useful for analyzing and refining the underlying concepts of the data [15]. Some previous approaches tried to develop black box neural network classification models to automatically label incoming cases as domestic or non-domestic violence but never made it into operational policing practice. One of the fundamental flaws of these approaches is that they assume that the underlying concepts of the data are clearly defined. As a consequence the concept of domestic violence itself had never been challenged. We combined FCA with ESOM (Figure 1) for doing the text mining analyses. The neural network technique ESOM helped us gaining insight in the overall distribution of the high-dimensional data. We can see three main clusters of domestic violence cases in Figure 1, one in the middle and two on the left of the map.

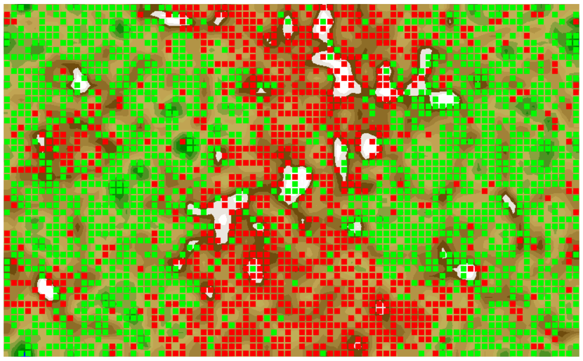


Figure 1. ESOM map. Domestic violence cases are shown as red dots and non-domestic violence cases as green dots.

ESOM functioned as a catalyst for distilling new concepts from the data and feed them into the FCA based discovery process. We uncovered multiple issues with the domestic violence definition, the training of police officers, etc. These issues include but are not limited to:

- niche cases and confusing situations: what if the perpetrator is a caretaker and the victim an inhabitant of an institution such as an old folks home? They have no family ties with each other, however there is a clear dependency relationship between them.
- faulty case labelings: we found police officers regularly misclassified burglary cases as domestic violence.
- data quality issues: multiple domestic violence cases lacked a formally labeled suspect.
- highly accurate and comprehensible classification rules: A comprehensible rule-based labeling system has been developed based on the FCA analyses for automatically labeling incoming cases. Currently, 75 % of incoming cases can be labeled correctly and automatically whereas in the past all cases had to be dealt with manually (Figure 3).

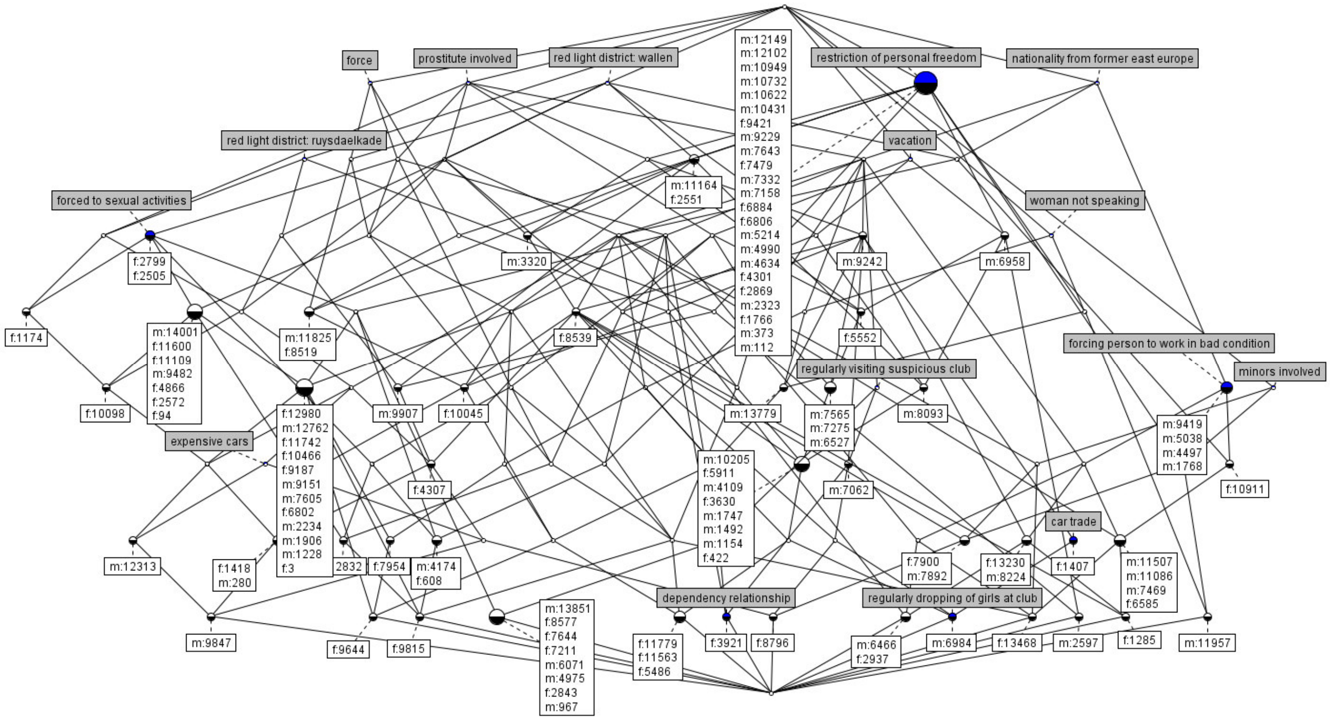
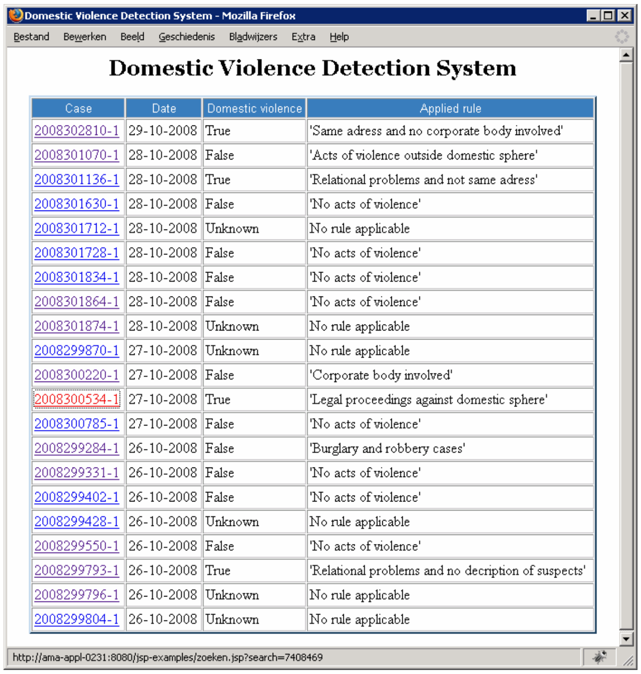


Figure 2. Human trafficking suspect detection lattice


Domestic Violence Detection System - Mozilla Firefox
Bestand
Bewerken
Beeld
Geschiedenis
Bladwijzers
Extra
Help
http://ama-appl-0231:8080/jsp-examples/zoeken.jsp?search=7408469

Domestic Violence Detection System

Case	Date	Domestic violence	Applied rule
2008302810-1	29-10-2008	True	'Same adress and no corporate body involved'
2008301070-1	28-10-2008	False	'Acts of violence outside domestic sphere'
2008301136-1	28-10-2008	True	'Relational problems and not same adress'
2008301630-1	28-10-2008	False	'No acts of violence'
2008301712-1	28-10-2008	Unknown	No rule applicable
2008301728-1	28-10-2008	False	'No acts of violence'
2008301834-1	28-10-2008	False	'No acts of violence'
2008301864-1	28-10-2008	False	'No acts of violence'
2008301874-1	28-10-2008	Unknown	No rule applicable
2008299870-1	27-10-2008	Unknown	No rule applicable
2008300220-1	27-10-2008	False	'Corporate body involved'
2008300534-1	27-10-2008	True	'Legal proceedings against domestic sphere'
2008300785-1	27-10-2008	False	'No acts of violence'
2008299284-1	26-10-2008	False	'Burglary and robbery cases'
2008299331-1	26-10-2008	False	'No acts of violence'
2008299402-1	26-10-2008	False	'No acts of violence'
2008299428-1	26-10-2008	Unknown	No rule applicable
2008299550-1	26-10-2008	False	'No acts of violence'
2008299793-1	26-10-2008	True	'Relational problems and no decription of suspects'
2008299796-1	26-10-2008	Unknown	No rule applicable
2008299804-1	26-10-2008	Unknown	No rule applicable

Figure 3. Domestic violence detection system

B. *Human trafficking*

Human trafficking is the fastest growing criminal industry in the world, with the total annual revenue for trafficking in persons estimated to be between $5 billion and $9 billion [19]. The council of Europe states that "people trafficking has reached epidemic proportions over the past decade, with a global annual market of about $42.5 billion" [22].

In the past, police officers had to manually search multiple databases regularly for signals of human trafficking. This was a very labor intensive approach and probably many signals remained undetected given the large amount of textual data available. In the project on human trafficking FCA was used to detect potential human trafficking suspects from unstructured observational police reports [21]. First FCA was used to iteratively build a domain specific thesaurus containing terms and phrases referring to human trafficking indicators. Then these indicators and the police reports were used to build FCA lattices from which potential suspects were distilled. An example of such a lattice is displayed in Figure 2. Persons lower in the lattice have more indicators and are more likely to be involved in human trafficking.

Temporal Concept Analysis, the FCA variant particularly suited for representing discrete temporal phenomena, was used to build visually appealing suspect profiles collecting all available information about these suspects in one picture. These lattices gave interesting insights into the criminal careers of the suspect and its evolution over time. This allows police officers to quickly determine if a subject should be monitored or not. The TCA lattices were finally used to investigate the evolution of the social network surrounding a suspect over time. This lattice also gave insights in the role of certain suspects in the network.

C. *Terrorist threat assessment*

In the terrorist threat assessment case study [18], FCA was again used to detect subjects from observational reports. Since the brute murder on the Dutch film maker Theo van Gogh, proactively searching for terrorists and signals of radicalizing behavior became more and more important to the police and intelligence agencies [5]. Investigators have to face the challenge of finding a few potentially interesting subjects in millions of text documents. The National Police Service Agency of the Netherlands (KLPD) developed a four-phase model of radicalization. According to this model, each subject passes through 4 phases before committing attacks: the preliminary, social alienation, jihadisation and jihad/extremism phase. With each phase, a combination of indicators is associated which should be available if the subject belongs to the phase. We used this model for the first time as text mining instrument and built a thesaurus with search terms for these indicators.

The goal of the analyses was to detect subjects as early as possible in their criminal careers to prevent them from committing attacks and increase chances of re-embedding them successfully in Dutch society. TCA lattices were found to give interesting insights into the radicalization process over time of a subject. The transition points from one phase to another and the points in time where the police should (have) intervene(d) are clearly visible. Figure 4 shows an example of a TCA lattice for a newly found suspect who went through all 4 phases.

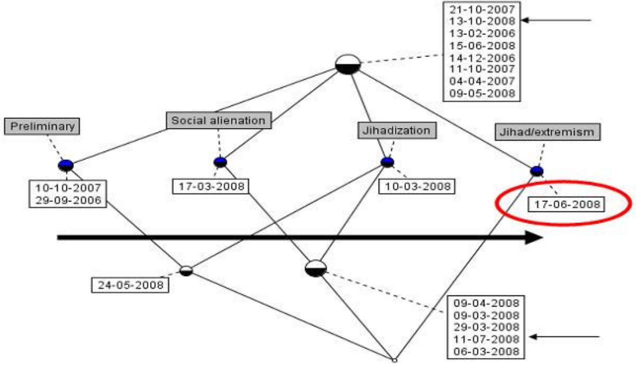


Figure 4. TCA radicalization lattice

The date of each observation of the suspect by the police and the severity of the indicators found are shown. At 17-06-2008 (red oval) the suspect reached the jihad/extremism phase and was spotted twice by the police afterwards (arrows).

IV. Innovative Concept Discovery Projects in Initial Stages of Development

A. *Predicting criminal careers of suspects*

In a project with the GZA hospital group in Antwerp (Belgium), we used FCA in combination with HMMs to gain insight into the breast cancer care process [16]. Activities performed to patients were turned into event sequences that were used as input for the HMM algorithm. We exposed multiple quality of care issues, process variations and inefficiencies after analyzing there data and models with FCA.

We are currently exploring the possibilities of using these techniques to predict the evolvement of criminal careers over time. At the Amsterdam-Amstelland police there is a list of repeat offenders and professional criminals. For each of these suspects there are multiple documents contained in police databases. Criminals typically go through successive phases with certain characteristics in their criminal careers and the indicators observed in the police reports related to a suspect can be turned into event sequences that can be fed into the HMM algorithm. Standard FCA analyses can be performed with the suspects as objects and the indicators observed as attributes. We believe that the combination of TCA and HMMs may be of considerable interest. Whereas TCA models as-is realties and is ideally suited for post-factum analysis, HMMs offer the advantage of being probabilistic models that can be used to predict the future evolvement of criminal careers and make risk assessment of certain situations occurring. FCA plays a pivotal role in analyzing the characteristics of suspicious groups distilled from the HMM models.

B. *Concept Discovery and Inovation Enabling Technology (CODIET) project*

Currently in its early stages of development is the software program Concept Discovery and Innovation Enabling Technology (CODIET). The package is based on the following 7 main modules:

- Data preprocessing:
  - o Data preparation step will allow the user to introduce different kinds of attributes including but not limited to text mining, temporal and compound attributes.
  - o Text mining attributes will consist of search terms for indexing the textual data sources and clusters grouping together semantically coherent search terms.
  - o The temporal attributes are based on temporal logic rules and make use of timestamps available in the data.
  - o Compound attributes can be composed from text mining and temporal attributes using first order logic.
  - o Segmentation rules can be defined based on these attributes to chop the data in pieces.

  - Textual documents are indexed using open source engine Lucene
  - A general XML data input format is defined and connectors to some popular data sources such as FCABedrock [1] and FCAStone [4] are integrated.
  - Generation of FCA, ESOM, HMM, TCA artifacts.
- FCA module
  - Concept Explorer [3] will be extended with some functionality that makes data browsing and exploration easier. For example, police report names in the extent of a concept will be made clickable and the report with relevant terms highlighted will be opened.
- HMM module
  - Currently makes use of Matlab but should preferably be reprogrammed in R or any other open source package or language
- ESOM module
  - A more efficient and flexible implementation should be made in R
- PRIDIT module
  - A method for unsupervised data classification [13]
- Temporal data analysis module
  - A temporal concept lattice can be generated in which concepts are ordered along a horizontal time axis
- User interface:
  - Graphically appealing and easy to use

One of the key elements in successfully embedding concept discovery methods in real life practice is making them easily available to police officers, medical specialists, etc. The main focus of this project lies on making the system intuitive to use for non-experts while at the same time offering strong data analysis capabilities. In particular data preparation should be performed semi-automatically since police officers are typically no experts in statistics or data analysis.

V. Conclusions

Each case study revealed the benefit of FCA as a human-centered instrument for data analysis that made domains previously inaccessible to analysts because of the overload of information, available for human reasoning and knowledge creation. In our study on domestic violence we used FCA for exploring and refining the underlying concepts of police data. Traditional machine learning and classification techniques build a model on the data without challenging the underlying concepts of the domain. We proposed FCA as a human-centered KDD instrument, that truly engages the analyst in the knowledge acquisition process. Terms are clustered in term clusters and the concept lattice shows the relationships between these term clusters and the police reports. We combined FCA with Emergent Self Organizing Maps to discover emergent structures in the high-dimensional data space. The KDD was a continuous process of iterating back and forth between analyzing the FCA and ESOM artifacts, selecting reports for in-depth manual inspection, gaining new knowledge and beginning a new knowledge creation cycle. Using FCA and ESOM we analyzed a large set of unstructured text reports from 2007 indicating incidents in the Amsterdam-Amstelland police region. We not only uncovered the true nature of domestic violence but also found multiple anomalies, faulty case labelings, confusing situations for police officers, niche cases, concept gaps, etc. This resulted in a refinement of the domestic violence definition, improvement of police training, reopening and relabeling filed reports and an automated domestic violence detection system. This system is based on 37 classification rules that were discovered during the successive knowledge discovery iterations. Each of these rules consist of a combination of early warning indicators which flag the nature of the case. If a domestic violence incident is detected, a red flag is raised. 75% of the incoming cases can be labeled correctly with this system.

We also analyzed FCA's applicability to data with an inherent time dimension. We twice made a combination of FCA and Temporal Concept Analysis. In our first case study we used FCA to distill potential human trafficking suspects from observational police reports and for suspicious persons a detailed profile was constructed with TCA. These profiles aided police officers in deciding which subjects should be monitored or further investigated. In a next step, we analyzed the social network of a suspicious person with TCA and used it to gain insight into the network's structure. We then repeated this exercise for terrorism subjects and based our text analysis method on the early warning indicators of the four phase model developed by the KLPD. The results were the discovery of several persons who were radicalizing or reached a critical radicalization phase but were not known by the police of Amsterdam- Amstelland. These subjects are currently being monitored by police authorities. For analyzing care processes, TCA's representation turned out to be too complex. Therefore we chose to use Hidden Markov Models for distilling process models from patient treatment data. FCA was used to expose quality of care issues and identifying characteristics of best practice process variations. These insights are currently being used to improve the quality of the provided care and will be applied to criminal career analysis.

In cooperation with the Amsterdam-Amstelland police we will develop and implement a toolset based on FCA theory for analyzing police data. Functionality will include text mining support such as indexing police reports using Lucene with a thesaurus, FCA lattice visualization and making the lattices easier to browse, connectors to database systems, etc.

ACKNOWLEDGMENT

The authors would like to thank the police of Amsterdam-Amstelland for granting them the liberty to conduct and publish this research. In particular, we are most grateful to Deputy Police Chief Reinder Doeleman and Police Chief Hans Schönfeld for their continued support. Jonas Poelmans is aspirant of the Research Foundation – Flanders.

REFERENCES

[1] Andrews, S., Orphanides, C. (2010) FcaBedrock, a Formal Context Creator. M. Croitoru, S. Ferr´e, and D. Lukose (Eds.): ICCS, LNAI 6208, pp. 181–184, 2010. Springer.

[2] Collier, P.M. (2006) Policing and the intelligent application of knowledge. Public money & management. Vol. 26, No. 2, pp. 109-116.

[3] Yevtushenko, S.A. (2000). System of data analysis "Concept Explorer." Proceedings of the 7th national conference or Artificial Intelligence. KII-2000. 127-134, Russia

[4] Priss, U. (2008) FcaStone - FCA file format conversion and interoperability software. Conceptual Structures Tool Interoperability Workshop (CS-TIW).

[5] AIVD (2006), Violent jihad in the Netherlands, current trends in the Ilamist terrorist threat. https://www.aivd.nl/aspx/download.aspx?file= /contents/pages/65582/jihad2006en.pdf

[6] Wille, R. (1982), Restructuring lattice theory: an approach based on hierarchies of concepts, I. Rival (ed.). Ordered sets. Reidel, Dordrecht-Boston, 445-470.

[7] Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.

[8] Wolff, K.E. (2005) States, transitions and life tracks in Temporal Concept Analysis. In: B. Ganter et al. (Eds.): Formal Concept Analysis, LNAI 3626, pp. 127-148. Springer, Heidelberg.

[9] Stumme, G., Wille, R., Wille, U. (1998), Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods, In: J.M. Zytkow, M. Quafofou (eds.): Principles of Data Mining and Knowledge Discovery, Proc. 2 nd European Symposium on PKDD '98, LNAI 1510, Springer, Heidelberg, 1998, 450-458.

[10] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230.

[11] Ultsch, A., Hermann, L. (2005) Architecture of emergent self-organizing maps to reduce projection errors. In Proc. ESANN 2005, pp. 1-6.

[12] Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp. 232-239.

[13] Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert, (2002) Fraud Classification Using Principal Component Analysis of RIDITs, Journal of Risk and Insurance, 69(3): 341-372.

[14] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.

[15] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. In : Lecture Notes in Artificial Intelligence, Vol. 5633(XI), (Perner, P. (Eds.)). Industrial conference on data mining ICDM 2009. Leipzig (Germany), 20-22 July 2009 (pp. 402 p.).

[16] Poelmans, J., Dedene, G., Verheyden, G., Van der Mussele, H., Viaene, S., Peters, E. (2010). Combining business process and data discovery techniques for analyzing and improving integrated care pathways. Lecture Notes in Computer Science, Advances in Data Mining. Applications and Theoretical Aspects, 10th Industrial Conference (ICDM), Leipzig, Germany, July 12-14, 2010. Springer

[17] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010), Formal Concept Analysis in knowledge discovery: a survey. Lecture Notes in Computer Science, 6208, 139-153, 18th international conference on conceptual structures (ICCS 2010): from information to intelligence. 26 - 30 July, Kuching, Sarawak, Malaysia**.** Springer**.**

[18] Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., Morsing, S. (2010) Terrorist threat assessment with Formal Concept Analysis. Proc. IEEE International Conference on Intelligence and Security Informatics. May 23-26, 2010 Vancouver, Canada. ISBN 978-1-42446460-9/10, 77-82.

[19] Highes, D.M. (2000) The "Natasha" Trade: The transnational shadow market of trafficking in women. Journal of international affairs, Spring 2000, 53, no. 2. The trustees of Colombia University in the City of new York.

[20] Rabiner, L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings IEEE 77 (2): 257-286.

[21] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010). A method based on Temporal Concept Analysis for detecting and profiling human trafficking suspects. Proc. *IASTED International Conference on Artificial Intelligence* (AIA 2010). Innsbruck, Austria, 15-17 february. Acta Press ISBN 978-0788986-817-5, pp. 330-338.

[22] Equality Division, Directorate General of Human Rights of the Council of Europe (2006) Action against trafficking in human beings: prevention, protection and prosecution. Proceedings of the regional seminar, Bucharest, Romania, 4-5 April.