Retrieval of Criminal Trajectories with an FCA-based Approach

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

[1]KU Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
[3]Amsterdam-Amstelland Police, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands

Jonas.Poelmans@gmail.com
Paul.Elzinga@amsterdam.politie.nl
Guido.Dedene@econ.kuleuven.be

Abstract. In this paper we briefly discuss the possibilities of Formal Concept Analysis for gaining insight in large amounts of unstructured police reports. We present a generic human centred knowledge discovery approach and showcase promising results obtained during empirical validation. The first case study focusses on distilling indicators for identifying domestic violence from 4814 reports with the aim of better recognizing new incoming cases. In the second case study we used FCA in combination with Temporal Concept Analysis to identify and investigate human trafficking suspects extracted from 266157 short observational reports. The third case study we present in this paper describes our application of FCA for identifying radicalising subjects from 166577 observational police reports. Finally, we conclude our paper with the case study on pedophile chat conversation analysis and the CORDIET data mining system.

Keywords. Formal Concept Analysis, Security informatics, Human trafficking, Terrorism, Pedophiles, Domestic violence

1 Introduction

During the joint Knowledge Discovery in Databases project, the Katholieke Universiteit Leuven and the Amsterdam-Amstelland Police Department have developed new special investigations techniques for gaining insight in police databases. These methods have been empirically validated and their application resulted in new actionable knowledge which helps police forces to better cope with domestic violence, human trafficking, terrorism and pedophile related data.

The implementation of the Intelligence-led policing management paradigm by the Amsterdam-Amstelland Police Department has led to an annual increase of suspicious activity reports filed in the police databases. These reports contain observations made

by police officers on the street during police patrols and were entered as unstructured text in these databases. Until now this massive amount of information was barely used to obtain actionable knowledge which may help improve the way of working by the police. The main goal of this joint research project was to develop a system which can be operationally used to extract useful knowledge from large collections of unstructured information. The methods which were developed aimed at recognizing (new) potential suspects and victims better and faster as before. In this paper we describe in detail the four major projects which were undertaken during the past five years, namely domestic violence, human trafficking (sexual exploitation), terrorism (Muslim radicalization) and pedophile chat conversations. During this investigation a knowledge discovery suite was developed, Concept Relation Discovery and Innovation Enabling Technology (CORDIET). At the basis of this knowledge discovery suite is the C-K design theory developed in Hatchuel et al. (2004) which contains four major phases and transition steps each of them focusing on an essential aspect of exploring existing and discovering and applying new knowledge. The investigator plays an important role during the knowledge discovery process. In the first step he has to assess and decide which information should be used to create the visual data analysis artifacts. During the next step multiple facilities are provided to ease the exploration of the data. Subsequently the acquired knowledge is returned to the action environment where police officers should decide where and how to act. This way of working is a corner stone for police forces who want to actively pursue an intelligent led policing approach.

2 Domestic violence

The first project started in 2007 and aimed at developing new methods to automatically detect domestic violence cases within the police databases (Poelmans et al. 2010a). The technique Formal Concept Analysis (Wille 1982, Ganter et al 1999, Poelmans et al. 2010b, 2012b) which can be used to analyze data by means of concept lattices, is used to interactively elicit the underlying concepts of the domestic violence phenomenon (van Dijk 1997).

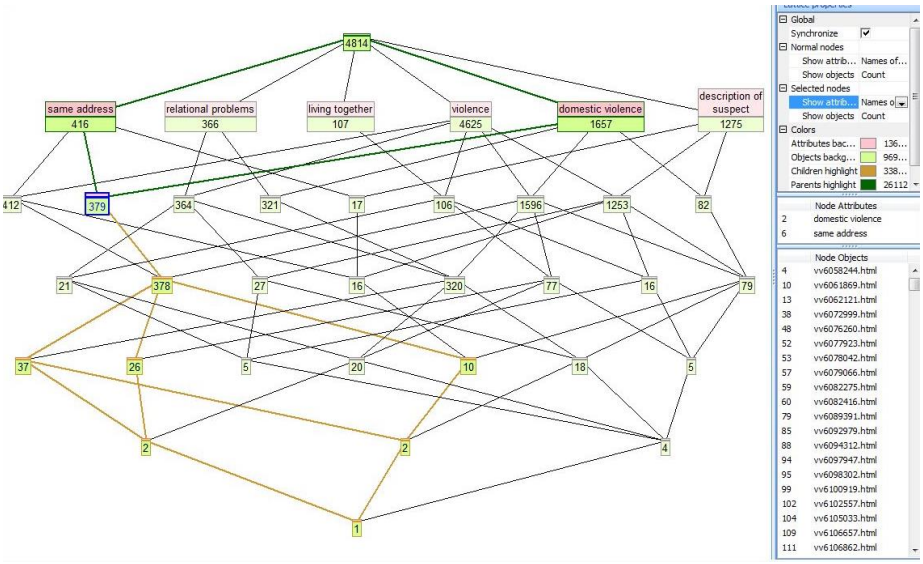

Fig. 1. Analyzing statements made by victims of a violent incident

The domestic violence definition which was employed by the Amsterdam-Amstelland police was as follows (Keus et al. 2000): "*Domestic violence can be characterized as serious acts of violence committed by someone in the domestic sphere of the victim. Violence includes all forms of physical assault. The domestic sphere includes all partners, ex-partners, family members, relatives and family friends of the victim. The notion of family friend includes persons that have a friendly relationship with the victim and (regularly) meet with the victim in his/her home.*" To identify domestic violence in police reports we make use of indicators which consist of words, phrases and / or logical formulas to compose compound attributes. The open source tool Lucene was initially used to index the unstructured textual reports using these attributes. The concept lattice visualization where reports are objects and indicators are attributes made it possible to iteratively identify valuable new knowledge. The lattice in Figure 1 contains 4814 police reports of which 1657 were labeled as domestic violence by police officers.

With CORDIET (see section 6 for details), the user can visually represent the underlying concepts in the data, gain insight in the complexity of the domain under investigation and zoom in on interesting concepts. For example we clicked on the node with 379 reports where suspect and victim lived on the same address and labeled as domestic violence by officers. Domain experts assumed that a situation where perpetrator and victim live at the same address is always a case of domestic violence, since these persons are probably family members, however this turned out not to be true. Analysis of the reports with attribute "same address" and not labeled as domestic violence revealed borderline cases such as violence in prisons, violence between a caretaker and inhabitant of an old folks home, etc. After multiple iterations of identi-

fying new concepts, composing new indicators and creating concept lattices we were able to refine the definition of domestic violence. Each of the cases were presented to the steering board of the domestic violence policy resulting in an improved definition of domestic violence and an improved handling of domestic violence cases. This investigation also resulted in a new automated case labelling system which is currently used to automatically label statements made by a victim to the police as domestic or non domestic violence (Poelmans et al. 2009, 2011a, Elzinga et al. 2009). At this moment the Amsterdam-Amstelland Police Department is using this system in combination with the national case triage system Trueblue.

3 Human trafficking

The next project focused on applying the knowledge exploration technique Formal Concept Analysis to detect (new) potential suspects and victims in suspicious activity reports and create a visual profile for each of them. The first application domain was human trafficking with a focus on sexual exploitation of the victims, a frequently occurring crime where the willingness of the victims to report is very low (Poelmans et al. 2011b, Hughes 2000).

After composing a set of early warning indicators and identifying potential suspects and victims, a detailed lattice profile of the suspect can be generated which shows the date of observation, the indicators observed and the contacts he or she had with other involved persons. In figure 2 the real names are replaced by arbitrary numbers and a number of indicators have been omitted for reasons of readability (the lattice was built using Concept Explorer). The persons (f = female and m = male) in the bottom of the figure are the most interesting potential suspects or victims because the lower a person appears in a lattice, the more indicators he or she has. For each of these persons a separate analysis can be made.

A selection of one of the men in the left bottom of figure 2 results in the concept lattice diagram in figure 3. In this figure the time stamps corresponding to each of the observations relevant for this person, together with the indicators and other persons mentioned are shown. The variant of formal concept analysis which makes use of temporal information is called temporal concept analysis (Wolff 2005). The lattice diagram shows that person D (4th left below) might be responsible for logistics, because he is driving in an expensive car (“dure auto”), and where the occupants show behavior of avoiding the police (“geen politie”). The man H (who appears in the extent of all concepts) is the possible pimp, who forced to work the possible victim woman S (1st upper right) in prostitution (“prostitutie” and “dwang”). Based on this diagram the corresponding reports can be collected and as soon as the investigators find sufficient indications a document based on section 273f of the Code of Criminal Law can be composed. This is a document that precedes any further criminal investigation against the man H.

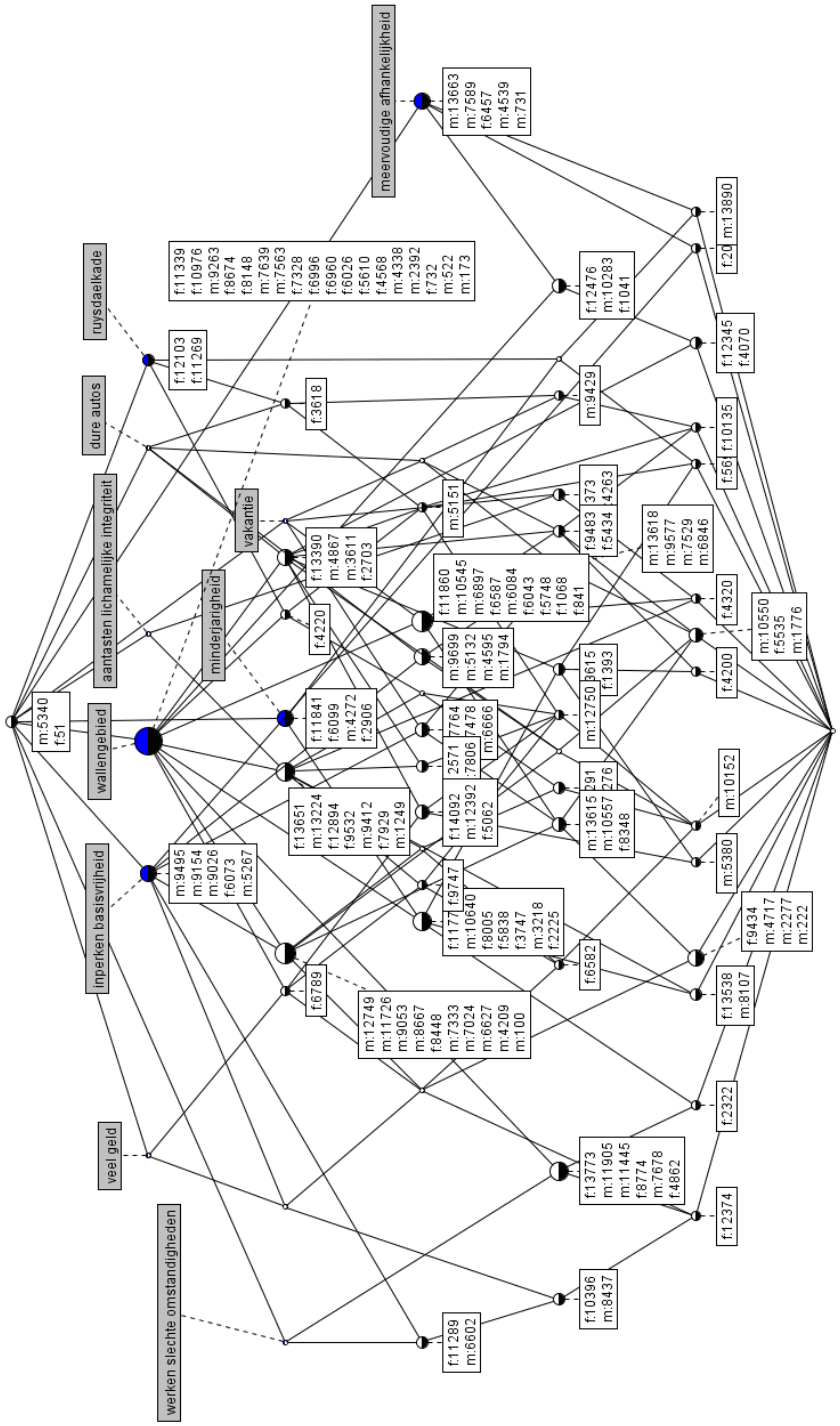

Fig. 2. Lattice of potential suspects and victims of human trafficking

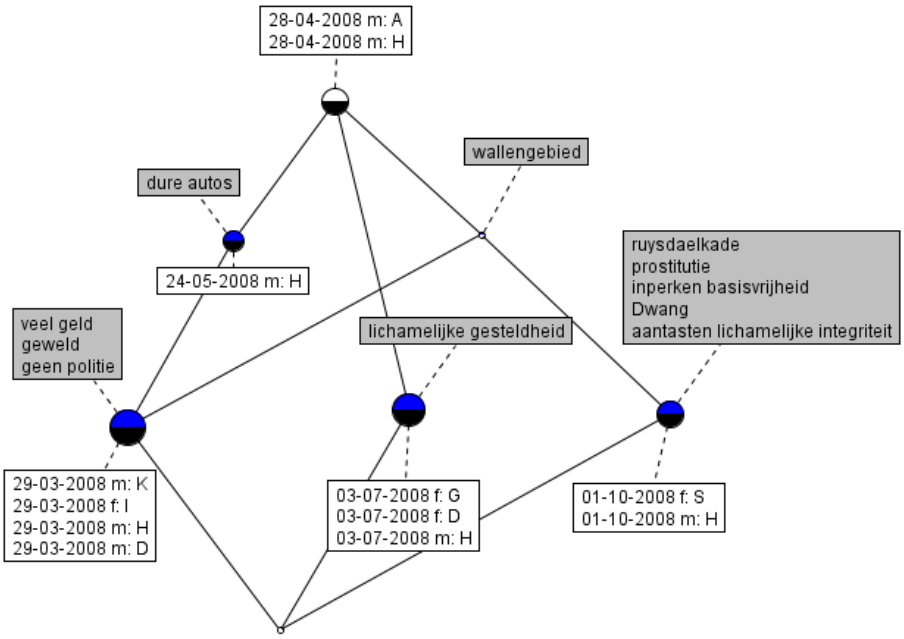

Fig. 3. Analysis of social network of suspect

4 Terrorism

During the third project we cooperated with the project team "Kennis in Modellen" (KiM, Knowledge in Models) from the National Police Service Agency in the Netherlands (KLPD). We combined formal concept analysis with the KiM model of Muslim radicalization to actively identifying potential terrorism suspects from suspicious activity reports (Elzinga et al. 2010, AIVD 2006). According to this model, a potential suspect goes through four stages of radicalization. The KiM project team has developed a set of 35 indicators based on interviews with experts on Muslim radicalism using which a person can be positioned in a certain phase. Together with the KLPD we intensively looked for characterizing words and combinations of words for each of these indicators. The difference with the previous models is that the KiM model added an extra dimension in terms of the number of different indicators which a person must have to be assigned to a radicalization phase.

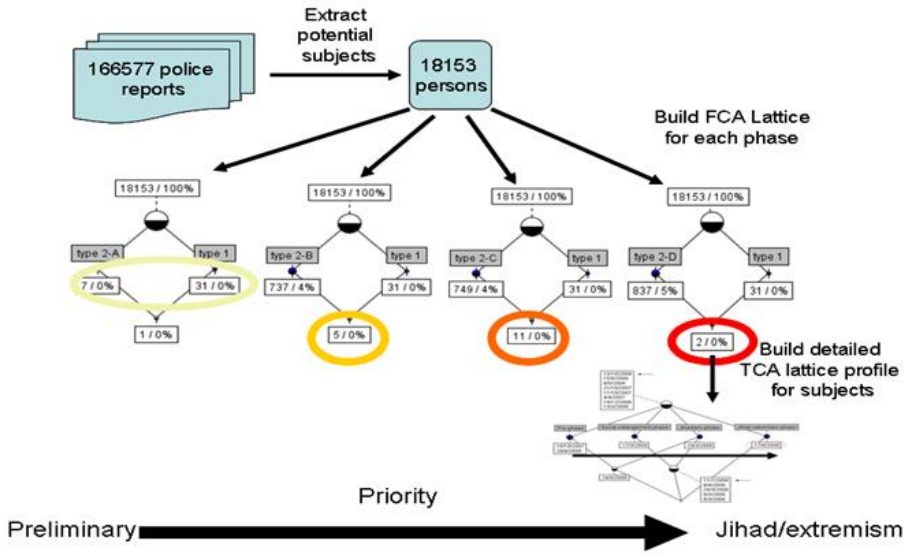

Fig. 4. The process model of extracting and profiling potential jihadists

The analysis was performed on the set of suspicious activity reports filed in the BVH database system of the Amsterdam-Amstelland Police Department during the years 2006, 2007 and 2008 resulting in 166,577 reports. From this set of observations 18,153 persons were extracted who meet at least one of the 35 indicators. From these 18.153 persons 38 persons were extracted who can be assigned to the 1st phase of radicalization, the preliminary phase ("voorfase"). Further analysis revealed that 19 were correctly identified, 3 of these persons were previously unknown by the Amsterdam-Amstelland Police Department, but known by the KLPD. From the 19 persons, 2 persons were found who met the minimal conditions of the jihad/extremism phase. For each of these persons a profile was made containing all indicators that were observed over time.

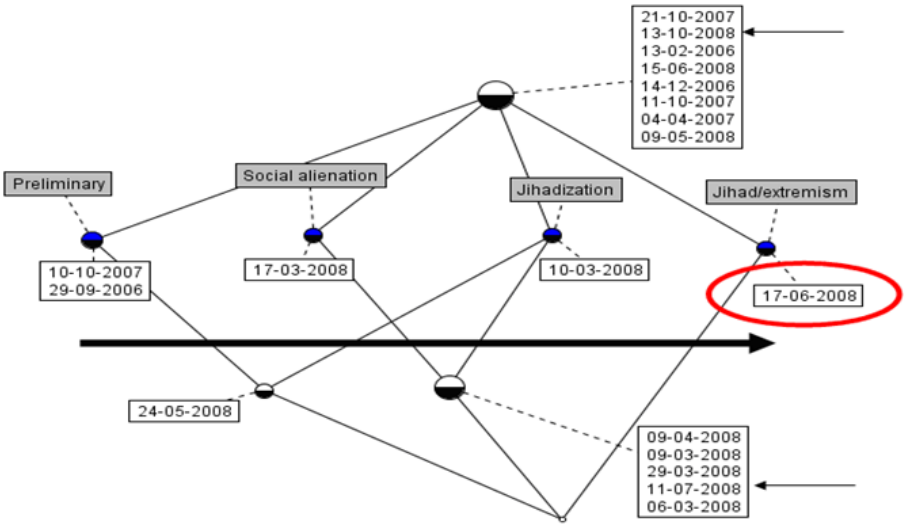

Fig. 5. Temporal lattice for subject C.

From the lattice diagram in figure 5 can be concluded that the person has reached the jihad/extremism phase on June 17, 2008 and has been observed by police officers two times afterwards (the arrows in the upper right and lower right of the figure) on July 11, 2008, and October 12, 2008.

5 Pedophile chat conversations

Chat conversations can be very long and time-consuming to read. A system which helps officers quickly identify those conversations posing a threat to a child's safety and understand what has been talked about may significantly speed up and improve the efficiency of their work (Elzinga et al. 2012).

Because original chat data collected by the Dutch police force organizations is restricted by law, results may not be made public. To demonstrate our FCA based method we use the chat data collected by a public American organization, Perverted Justice, which actively searches for pedophiles on the internet. We downloaded 533 chat files, i.e. one for each of the 533 different suspects. The victims in all chat files are adults playing the role of a young girl or boy in the age from 12 to 14. All these adults are members of the Perverted Justice organization and are trained to act as a youngster. The adults playing the victim try to lure the suspect by playing his or her role as good as possible. The behavior of the victims cannot be representative for young girls or boys, but the behavior of the suspects is realistic since they really believe to have contact with a young girl or boy and act in that way.

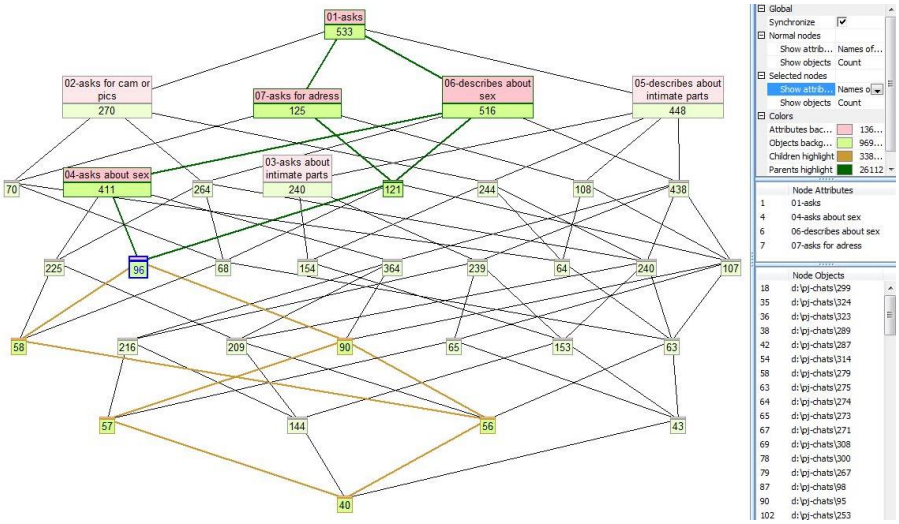

Fig. 6. Analyzing chat conversations of pedophiles with members of the perverted justice organization who pretend to be a young child

The lattice in Figure 8 shows how a set of 533 chat conversations was analyzed with FCA. We defined 7 term clusters containing keywords which were used by pedophiles in their chat conversations. We numbered these 7 attributes according to the severity of the threat to the child's safety. We clicked on a concept with 96 conversations in the extent and attributes "asks", "asks about sex", "describes about sex" and "asks for address". In the "node objects" pane the user can click on the name of a conversation to display its contents.

Figure 7 shows a transition diagram of the chat conversation 451 which was selected based on the line diagram shown in Figure 6. We explain Figure 7 intuitively in this paper, readers interested in the mathematical definitions are referred to Elzinga et al. (2012). Figure 7 is constructed by restricting the data table to the rows where chat log = 451, which are the 22 rows from 300 to 321. The chat time runs in these rows from 0 to 21. The many-valued attribute 'state' has in row 300, that is at time 0, the value '2' which means that the conversation 451 is in the state '2 Compliments'; in the next row 301, at time 1, the conversation 451 is in the state '5 Cam and pics'. This transition is graphically represented in Figure 7 by the arrow from the object concept of 300 to the object concept of 301. Clearly, the direction of the arrow is induced from the fact that time 0 is the predecessor of time 1 (in the natural ordering of integers).

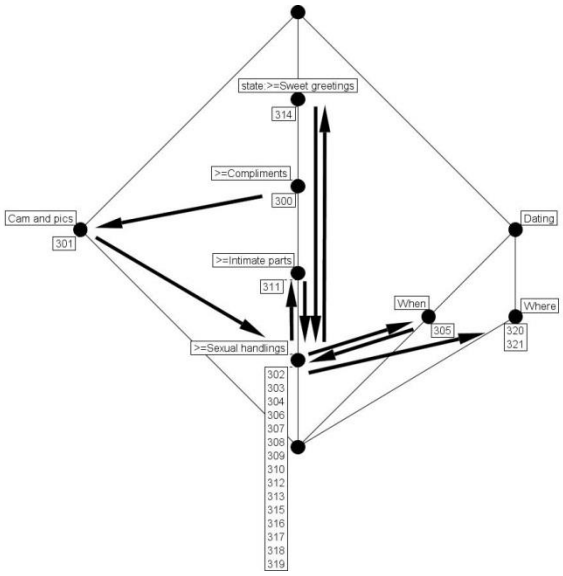

Fig. 7. A transition diagram for chat conversation 451

In Elzinga et al. (2012) we describe in detail how we selected chats from such a concept lattice and analyze them in detail with temporal relational semantic systems.

6 CORDIET

More and more companies have large amounts of unstructured data, often in textual form available. The few analytical tools that focus on this problem area offer insufficient functionality for the specific needs of many of these organizations. As part of the research work in the doctoral research of Jonas Poelmans the development of the data analysis suite Concept Relation Discovery and Innovation Enabling Technology (CORDIET, Elzinga 2011, Poelmans et al. 2012) was started in September 2010 in cooperation with the Moscow Higher School of Economics. Elzinga et al (2009) developed the first prototype where the strength of our approach with concept lattices and other visualization techniques such as Emergent Self Organizing Maps (ESOM) is demonstrated for the detection of individuals with radicalizing behavior. This toolset allows to carry out much faster and more detailed data analysis to distil relevant persons from police data.

7 Conclusions

The four projects which are carried out as part of the research chair show the potential of the knowledge exploration technique formal concept analysis. Especially the intuitively interpretable visual representation was found to be of great importance for in-

formation specialists within the police force on all levels, strategic, tactic and operational. This visualization did not only allow to explore the data interactively, but also to explore and define the underlying concepts of the investigation areas. New concepts, anomalies, confusing situations and faulty labeled cases were discovered, but also not previously known subjects were found who might be involved in human trafficking or terroristic activities. The temporal variant of formal concept analysis proved to be very useful for profiling suspects and their evolution over time. Never before unstructured information sources were retrieved in such a way that new insights, new suspects and victims became visible. That's why formal concept analysis will become an important instrument in the nearby future for information specialists within the police and will be an essential contribution to the formation of Intelligence within the Dutch police.

Among the future developents are applications of FCA-based biclustering (Ignatov et al. (2012)) and triclustering techniques (Ignatov et al. (2011)) in the Criminal Investagations domain.

References

1. AIVD (2006), Violent jihad in the Netherlands, current trends in the Islamist terrorist threat. https://www.aivd.nl/aspx/download.aspx?file=/contents/pages/65582/jihad2006en.pdf
2. Elzinga, P., Poelmans, J., Viaene, S., Dedene, G. (2009), Detecting Domestic Violence, showcasing a knowledge browser based on Formal Concept Analysis and Emerging Self Organizing Maps. 11th International Conference on Enterprise Information Systems, Milan 6-10 may 2009.
3. 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, pp.77-82.
4. Elzinga, P. (2011) Formalizing the concepts of crimes and criminals. PhD dissertation University of Amsterdam.
5. Elzinga, P., Wolff, K.E., Poelmans, J., Viaene, S., Dedene, G. (2012) Analyzing chat conversations of pedosexuals with temporal relational semantic systems. F. Domenach et al. (eds.): Contributions to 10th International Conference on Formal Concept Analysis, Leuven, Belgium, 6 – 10 May 2012. ISBN 978-9-08-140995-7, 82-97.
6. Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.
7. Hatchuel, A., Weil, B., Le Masson, P (2004) Building innovation capabilities. The development of Design-Oriented Organizations: In Hage, J.T. (Ed), Innovation, Learning and Macro-institutional Change: Patterns of knowledge changes.
8. Hughes, D.M. (2000), 'The "Natasha" Trade: The Transnational Shadow Market of Trafficking in Women,' *Journal of International Affairs*, Spring, 53, no. 2.
9. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., and Zhukov, L.E. (2011) From Triconcepts to Triclusters. In: Proc. of 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining And Granular Computing, Kuznetsov et al. (Eds.): RSFDGrC 2011, LNCS/LNAI Volume 6743/2011, Springer-Verlag Berlin Heidelberg, 257-264

10. Ignatov, D.I., Kuznetsov, S.O., Poelmans, P. (2012) Concept-Based Biclustering for Internet Advertisement. ICDM Workshops 2012, 123-130
11. Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
12. Poelmans, J., Elzinga, P., Neznanov, A., Dedene, G., Viaene, S., Kuznetsov, S. (2012) Human-Centered Text Mining: a New Software System. P. Perner (Ed.): Lecture Notes in Artificial Intelligence 7377, 258–272, 12th Industrial Conference on Data Mining. Springer
13. 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.).
14. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010a) Curbing domestic violence: Instantiating C-K theory with Formal Concept Analysis and Emergent Self Organizing Maps. Intelligent Systems in Accounting, Finance and Management 17, (3-4) 167-191. Wiley and Sons, Ltd..
15. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010b), 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.
16. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2011a) Formally Analyzing the Concepts of Domestic Violence, Expert Systems with Applications 38, (4) 3116-3130. Elsevier Ltd.
17. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011b) A concept discovery approach for fighting human trafficking and forced prostitution. Lecture Notes in Computer Science, 6828, 201-214, 19th International conference on conceptual structures, July 25-29, Derby, England. Springer.
18. Poelmans, J., Ignatov, I., Viaene, S., Dedene, G., Kuznetsov, S. (2012b) Text mining scientific papers: a survey on FCA-based information retrieval research. P. Perner (Ed.): Lecture Notes in Artificial Intelligence 7377, 273–287, 12th Industrial Conference on Data Mining, July 13-20, Berlin, Germany. Springer
19. T. van Dijk, Huiselijk geweld, aard, omvang en hulpverlening (Ministerie van Justitie, Dienst Preventie, Jeugd-bescherming en Reclassering, oktober 1997).
20. Wille, R. (1982), Restructuring lattice theory: an approach based on hierarchies of concepts, I. Rival (ed.). Ordered sets. Reidel, Dordrecht-Boston, 445-470.
21. 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.