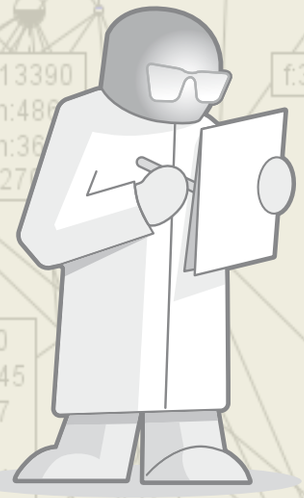
Formalizing the concepts of crimes and criminals

Elzinga, P.G.

Formalizing the Concepts of Crimes and Criminals

Paul Elzinga

Formalizing the Concepts of Crimes and Criminals



Paul Elzinga

Formalizing the concepts of crimes and criminals

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. Dr. D.C. van den Boom
ten overstaan van een door het college van promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 11 oktober 2011, te 14:00 uur
door Paul Godfried Elzinga
geboren te Alkmaar

Promotiecommissie

Promotor	Prof. Dr. G.G.M Dedene
Co-promotores	Prof. Dr. S. Viaene
	Prof. Dr. Ir. R. Maes
Overige leden	Prof. Dr. P.W. Adriaans
	Prof. Dr. A. Heene
	Dr. J. Poelmans
	Dr. E.J. de Vries

Faculteit der Economie en Bedrijfskunde

Aan Mirjam en Maria,
Jan en Jannie
en Jennifer

CONTENTS

SUMMARY I
CHAPTER 1 1
INTRODUCTION 1
1.1 Concept Discovery 1
1.2 Intelligent-Led Policing, a historical overview 2
1.3 Intelligence-led policing and C-K modeling 3
1.3.1 3-i model of Ratcliffe 4
1.3.2 Concept Knowledge theory 4
1.4 Intelligence-led Policing and text mining 7
CHAPTER 2 9
Formal concept analysis in the literature 9
2.1 Introduction 9
2.2 Formal Concept Analysis 10
2.2.1. FCA essentials 10
2.2.2. FCA software 13
2.2.3. Web portal 13
2.3 Dataset 14
2.4 Studying the literature using FCA 14
2.4.1 Knowledge discovery and data mining 15
2.4.2 Information retrieval 17
2.4.3 Scalability 19
2.4.4 Ontologies 19
2.5 Conclusions 20
CHAPTER 3 23
Curbing domestic violence: Instantiating C-K theory with Formal Concept Analysis and Self Organizing Maps 23
3.1 Introduction 23
3.2 Intelligence Led Policing 26
3.2.1 Domestic violence 26
3.2.2 Motivation 28
3.3 FCA, ESOM and C-K theory 29
3.3.1 Formal Concept Analysis 29
3.3.2 Emergent Self Organizing Map 32
3.3.2.1 Emergent SOM 32
3.3.2.2 ESOM parameter settings 33
3.3.3 C-K theory 34
3.4 Instantiating C-K theory with FCA and ESOM 35
3.5 Dataset 38
3.5.1 Data pre-processing and feature selection 40
3.5.2 Initial classification performance 41
3.6 Iterative knowledge discovery with FCA and ESOM 42
3.6.1 Transforming existing knowledge into concepts 44
3.6.2 Expanding the space of concepts 51

3.6.3 Transforming concepts into knowledge....53
3.6.4 Expanding the space of knowledge....56
3.7 Actionable results....58
3.8 Comparative study of ESOM and multi-dimensional scaling....65
3.9 Conclusions....69
CHAPTER 4....71
Formal concept analysis of temporal data.....71
4.1 Terrorist threat assessment with Temporal Concept Analysis....71
4.1.1 Introduction....71
4.1.2 Backgrounder....72
4.1.2.1 Home-grown terrorism....72
4.1.2.2 The four phase model of radicalism....73
4.1.2.3 Current situation....74
4.1.3 Dataset....75
4.1.4 Temporal Concept Analysis....76
4.1.4.1 FCA essentials....76
4.1.4.2 TCA essentials....77
4.1.5 Research method....79
4.1.5.1 Extracting potential jihadists with FCA....79
4.1.5.2 Constructing Jihadism phases with FCA....81
4.1.5.3 Build detailed TCA lattice profiles for subjects....81
4.1.6 Conclusions....82
4.2 Identifying and profiling human trafficking and loverboy suspects....83
4.2.1 Introduction....83
4.2.2 Human trafficking and forced prostitution....84
4.2.2.1 Human trafficking model....84
4.2.2.2 Loverboy model....85
4.2.3 Dataset....86
4.2.4 Method....86
4.2.4.1 FCA analysis....87
4.2.4.2 Thesaurus....88
4.2.5 Analysis and results....89
4.2.5.1 Detection of suspects of human trafficking and forced prostitution....90
4.2.5.2 Case 1: Turkish human trafficking network....90
4.2.5.3 Case 2: Bulgarian male suspect....92
4.2.5.4 Case 3: Hungarian woman both victim and suspect....94
4.2.5.5 Case 4: Loverboy suspect....96
4.2.6 Discussion....97
4.2.7 Conclusions....100
CHAPTER 5....101
Concept Relation Discovery and Innovation Enabling Technology (CORDIET)101
5.1 Introduction....101
5.2 Data analysis artefacts....102
5.2.1 Formal Concept Analysis....102
5.2.2 Temporal Concept Analysis....102

5.2.3 Emergent Self Organising Maps....103
5.2.4 Hidden Markov Models....103
5.3 Data sources....103
5.3.1 Data source BVH....104
5.3.2 Data source scientific articles....104
5.3.3 Data source clinical pathways....105
5.4 Application domains....107
5.4.1 Domestic violence....107
5.4.2 Human trafficking....107
5.4.3 Terrorist threat assessment....108
5.4.4 Predicting criminal careers of suspects....109
5.5 CORDIET system architecture and business use case diagram....110
5.5.1 Business use case diagram....110
5.5.2 The software lifecycles of CORDIET....111
5.5.3 The development of an operational version of CORDIET....112
5.5.3.1 Presentation layer....112
5.5.3.2 Service....113
5.5.3.3 Business layer....113
5.5.3.4 Data access layer....113
5.5.3.5 Data....113
5.5.3.6 User interface....113
5.5.3.7 Language module....113
5.6 CORDIET functionality....113
5.6.1 K->C phase: start investigation....113
5.6.1.1 Load data sources....114
5.6.1.2 PostgreSQL database:....114
5.6.1.3 Lucene:....116
5.6.1.4 Create, load or modify ontology....116
5.6.1.5 Text mining attributes....118
5.6.1.6 Temporal attributes....118
5.6.1.7 Compound attributes....118
5.6.2 C->C phase: compose artefact....119
5.6.2.1 Select ontology....119
5.6.2.2 Define rules....119
5.6.2.2.1 Segmentation rules....120
5.6.2.2.2 Object cluster rules....120
5.6.2.2.3 Classifier rules....120
5.6.3 Choose and create artefact....121
5.6.3.1 C->K phase: analyze artefact....121
5.6.3.1.1 Detect object of interest....121
5.6.3.1.2 Detect anomaly....122
5.6.3.1.3 Detect knowledge concept....122
5.6.3.2 K->K phase: deploy knowledge product....122
5.7 Data and domain analysis scenarios....123
5.7.1 The functionality of the CORDIET toolbox....124
5.7.1.1 Knowledge space options....125

5.7.1.1.1 Ontology 125
5.7.1.1.2 Rule base 125
5.7.1.1.3 Summary report 126
5.7.1.1.4 Concept space options 126
5.7.1.1.5 TuProlog 126
5.7.1.1.6 ConExp 126
5.7.1.1.7 ESOM 126
5.7.1.1.8 Venn Diagramm 126
5.7.1.1.9 Tool menu options 127
5.7.1.1.10 Lucene index 127
5.7.1.1.11 Export RDBMS 128
5.7.1.1.12 Export Topicview 128
5.7.1.1.13 Export Topicmap 128
5.7.1.1.14 Export to HTML 128
5.7.2 Data analysis scenario “Create an ontology and a rule base for Domestic Violence” 129
5.7.2.1 K->C, prepare the datasets and create the ontology 129
5.7.2.1.1 Prepare the datasets 129
5.7.2.1.2 Create a new ontology. 130
5.7.2.2 C->C: compose artefact 134
5.7.2.2.1 Select the ontology and rules 134
5.7.2.3 C->K analyze the artefacts 135
5.7.2.3.1 Analyze the initial results with a Venn diagram 135
5.7.2.3.2 Analyze the initial results with FCA lattices 136
5.7.2.3.3 Validate the ontology using FCA lattice 137
5.7.2.4 K->K: deploy new knowledge 139
5.7.2.5 Start a new C/K iteration 139
5.7.2.6 Validate the ontology using ESOM toroid map 141
5.7.2.7 C->C: compose the ESOM input files 143
5.7.2.8 C->C: Analyze the results of the ESOM map 145
5.7.2.9 K->K and K->C: update the ontology 146
5.7.2.10 C->C and C->K: compose new FCA input files and analyze the FCA lattices 147
5.7.2.11 K->K: deploy new knowledge 148
5.7.3 Domain analysis of human trafficking. 148
5.7.3.1 Identify possible suspects and or victims 149
5.7.3.1.1 K->C: Create the signals ontology 149
5.7.3.1.2 C->C: compose the FCA lattices 150
5.7.3.1.3 C->K: analyze the FCA lattices 150
5.7.3.1.4 K->K: Creating a 27-construction report 157
5.7.4 Analyze the workforce intelligence of clinical pathways 158
5.7.4.1 Data sources 158
5.7.4.2 Ontology for workflow intelligence 159
5.7.4.3 Process variations 161
5.7.4.4 Analyzing the workflow intelligence 164
5.8 Conclusions 167

CHAPTER 6 169
Thesis conclusions 169
6.1 Thesis conclusions 169
6.2 Future work 171
6.2.1 Terrorist threat assessment. 171
6.2.2 Soloist threateners threat assessment. 171
6.2.3 Human trafficking. 172
6.2.4 Domestic violence. 172
6.2.5 Improve the information quality of the BVH system. 172
6.2.6 Financial Crime Analysis. 172
6.2.7 Predicting crime careers 172
6.2.8 Supporting Large-scale investigation Teams 173
6.2.9 Intelligence Led Policing and Concept Discovery Toolset 173
SAMENVATTING 175
DANKWOORD 185
PUBLICATIONS 187
APPENDIX A 191
Literature survey thesaurus 191
APPENDIX B 193
Domestic violence case thesaurus 193
APPENDIX C 197
Human trafficking thesaurus 197
APPENDIX D 201
Simulating the Trueblue Domestic Violence rule 201
APPENDIX E 205
The rule based application 205
APPENDIX F 211
Topicmap with FCA literature ontology examples 211
APPENDIX H 215
Human trafficking and Loverboy indicators 215
APPENDIX I 219
Excerpts of ESOM input files 219
BIBLIOGRAPHY 221

SUMMARY

1. Introduction

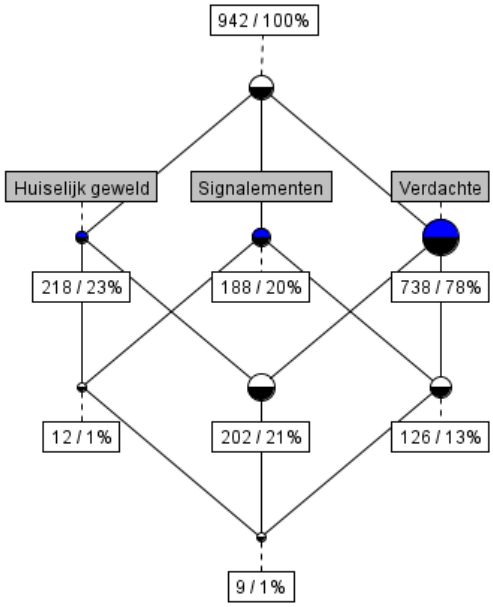
During the joint Knowledge Discovery in Databases project, the Katholieke Universiteit Leuven and the Amsterdam-Amstelland Police Department have developed new special investigations techniques for gaining insight in police databases. These methods have been empirically validated and their application resulted in new actionable knowledge which helps police forces to better cope with domestic violence, human trafficking and terrorism 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 thesis we describe in detail the three major projects which were undertaken during the past three years, namely domestic violence, human trafficking (sexual exploitation) and terrorism (Muslim radicalization). 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 Hatchuell et al. (1999, 2002 and 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 database. The technique Formal Concept Analysis (Wille 1982, Ganter et al 1999) 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). 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 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. After multiple iterations of identifying new concepts, composing new indicators and creating concept lattices we were able to refine the definition of domestic violence. During this process, multiple situations were found which were confusing to police officers. Also many faulty labels assigned to domestic and non domestic violence cases were detected. This investigation 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, Elzinga et al. 2009). At this moment the Amsterdam-Amstelland Police Department is using this system in combination the national case triage system Trueblue. An example of a concept lattice diagram showing cases which are potentially faulty labeled as domestic violence is shown below. The nodes in the lattice are the concepts. Each concept consists of two parts, a set of objects and a set of attributes. The figures in the white rectangle are the number of objects belonging to the concept. The gray rectangles are the attributes. A concept has an attribute when it is possible to navigate from the corresponding node by only following upwards lines towards the attribute. The lattice in the figure below can be read in the following way. Starting from the lowest node, following the lines upwards results in the attributes “Huiselijk geweld” (domestic violence), “Signalementen” (description of the suspect) and “Verdachte” (formally labeled suspect).

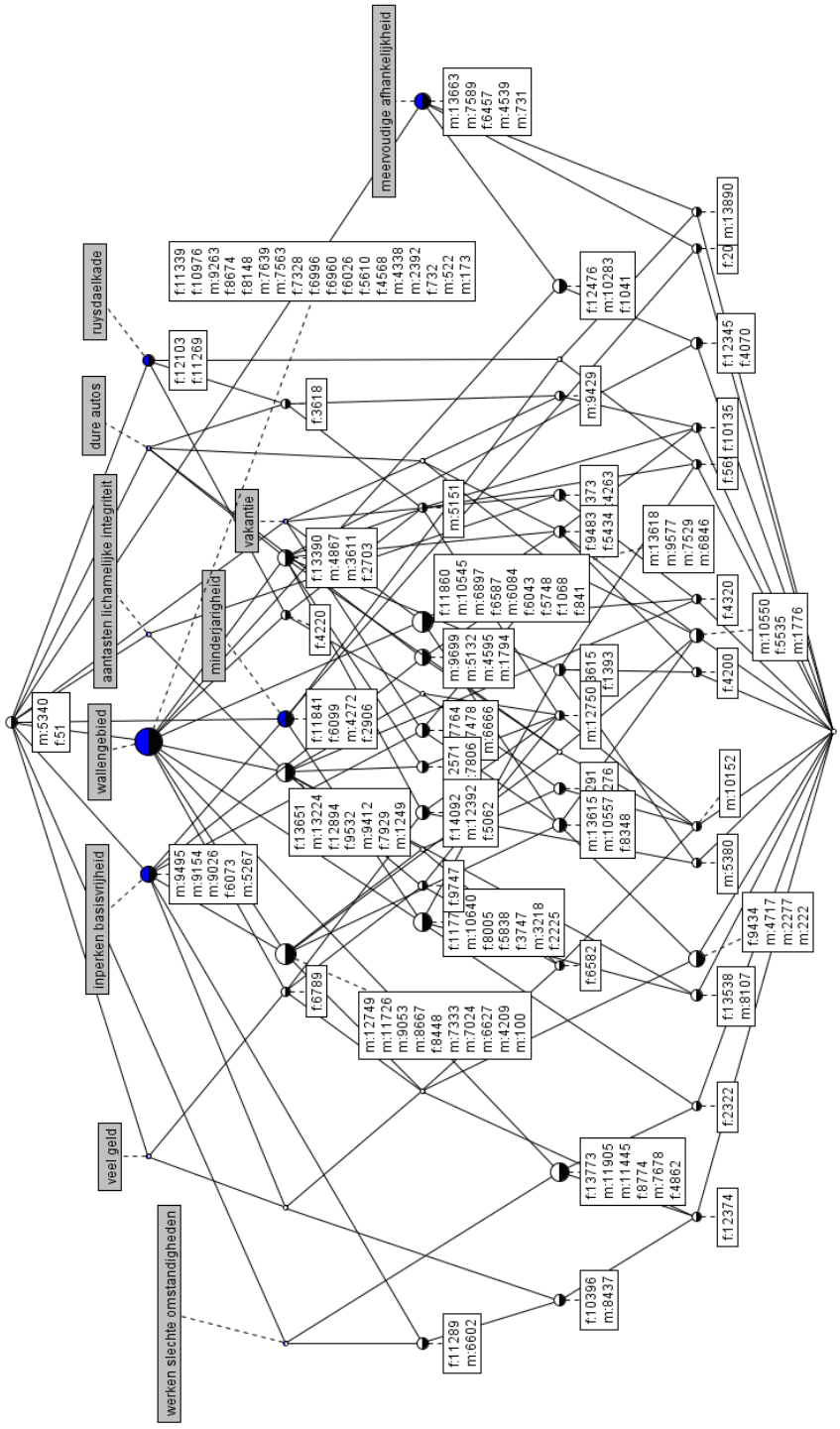


218 cases have been labeled as domestic violence by police officers. A subset of 202 cases has been labeled as domestic violence and mention a formally labeled

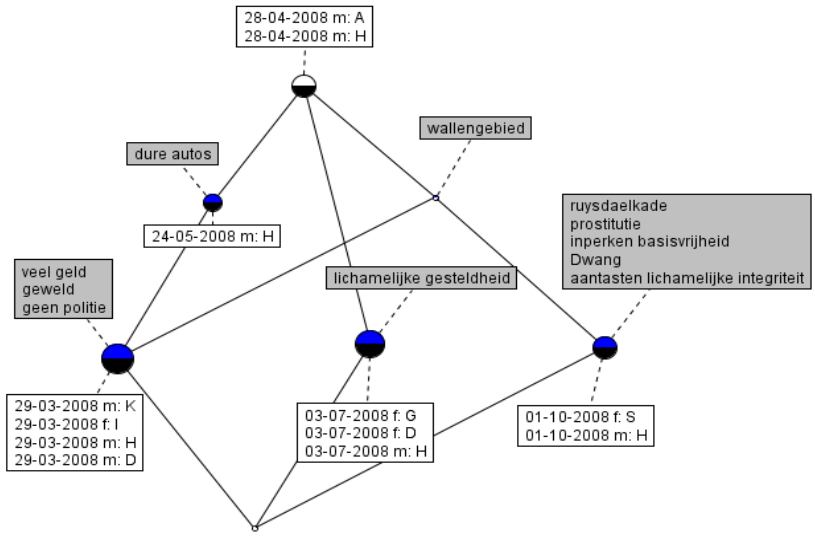
suspect. The lattice shows 9 domestic violence cases which mention both a formally labeled suspect and a description of a suspect. After in depth investigation it turned out that the 9 suspects do not have an official living address and an arrest warrant has been issued. We also observe 3 cases labeled as domestic violence which contain a description of the suspect but no formally labeled suspect is mentioned. It turned out that all 3 cases were faulty classified as domestic violence. From this analysis a knowledge rule can be obtained which can be used to classify with an accuracy of almost 100% violence cases with a description of the suspect but not mentioning a formally labeled suspect as non-domestic violence.

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. 2010, Highs 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 this figure the real names are replaced by arbitrary numbers and a number of indicators have been omitted for reasons of readability.


meervoudige afhankelijkheid
m:13663
m:7589
f:6457
m:4539
m:731
ruysdaelkade
dure autos
aantasten lichamelijke integriteit
minderjarigheid
vakantie
wallengebied
inperken basisvrijheid
veel geld
werken slechte omstandigheden
f:11339
f:10976
m:9263
f:8674
f:8148
m:7639
m:7563
f:7328
f:6996
f:6960
f:6026
f:5610
f:4568
m:4338
m:2392
f:732
m:522
m:173
f:12103
f:11269
f:3618
m:5151
m:5340
f:51
f:13390
m:4867
m:3611
f:2703
f:4220
f:11860
m:10545
m:6897
f:6587
m:6084
f:6043
f:5748
f:1068
f:841
m:9699
m:5132
m:4595
m:1794
f:11841
f:6099
m:4272
f:2906
f:13651
m:13224
f:12894
f:9532
m:9412
f:7929
m:1249
f:14092
2571
7764
m:12392
7806
7478
f:5062
m:6666
m:9495
m:9154
m:9026
f:6073
m:5267
f:1177
f:9747
m:10640
f:8005
f:5838
f:3747
m:3218
f:2225
f:6789
f:6582
m:12749
m:11726
m:9053
m:8667
f:8448
m:7333
m:7024
m:6627
m:4209
m:100
f:13773
m:11905
m:11445
f:8774
m:7678
f:4862
f:10396
m:8437
f:11289
m:6602
f:12476
m:10283
f:1041
m:9429
f:9483
373
f:5434
4263
m:13618
m:9577
m:7529
m:6846
f:20
m:13890
f:12345
f:4070
f:561
f:10135
f:4320
f:4200
m:10550
f:5535
m:1776
f:1393
m:12750
3615
m:13615
291
m:10557
276
f:8348
m:10152
m:5380
f:9434
m:4717
m:2277
m:222
f:13538
m:8107
f:2322
f:12374

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 the figure results in the following concept lattice diagram:

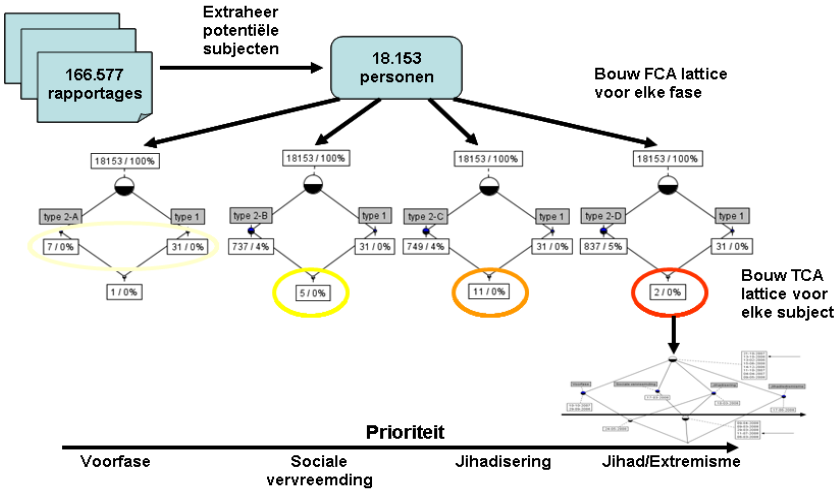


In this figure the time stamps corresponding to each of the observations relevant for this person, together with the indicators and other persons mentioned are shown. The variant of formal concept analysis which makes use of temporal information is called temporal concept analysis (Wolff 2005). The lattice 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 (Staatscourant 2006, 58) can be composed. This is a document that precedes any further criminal investigation against the man H.

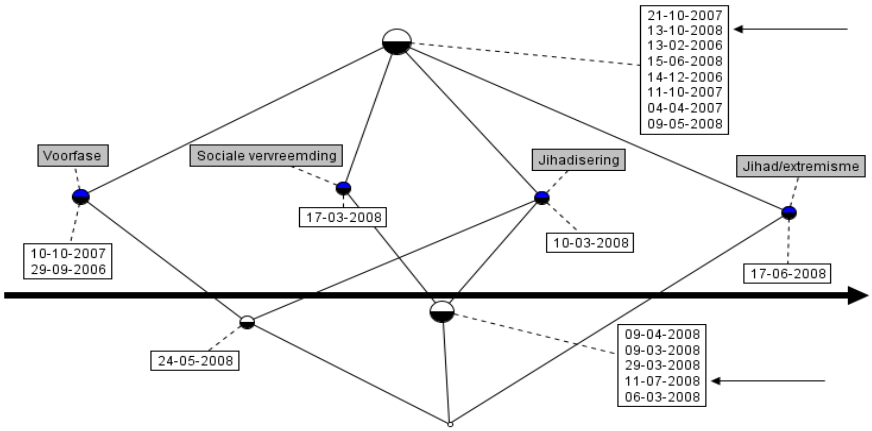
4. Terrorism

During the last 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.



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.



From this lattice diagram 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. 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 (Aspirant FWO[1]) the development of the data analysis suite Concept Relation Discovery and Innovation Enabling Technology (CORDIET) was started in September 2010 in cooperation with the Moscow Higher School of Economics. A project plan has been composed under supervision of Prof. Sergei Kuznetsov PhD, drs. Paul Elzinga and Jonas Poelmans PhD, where 20 master students, 2 doctoral researchers, 2 post doctoral researchers and 2 professors, all from Russia, are involved. The result of the cooperation will be the complete data analysis suite CORDIET, including the successful application of this toolset on the unstructured reports of the Amsterdam-Amstelland Police Department and the medical reports of the GZA hospitals. This toolset will be used in ongoing projects for the proactive detection of possible potential suspects of terrorism and human trafficking in the region of Amsterdam-Amstelland. Elzinga et al (2010) has conducted a proof of concept 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. During this PhD study, a number of possible suspects and victims of human trafficking are analyzed and profiled (Poelmans et al. 2010c). This toolset allows to carry out much faster and more detailed data analysis to distil relevant persons from police data. The methodology

[1] FWO: Fonds voor Wetenschappelijk Onderzoek - Vlaanderen

of the toolset does not only fit within the philosophy of Intelligent-led policing, but also fits within the context of hospitals where data of breast cancer patients were analyzed to improve the care provided (Poelmans et al. 2010d). In the hospital group GZA, the toolset will be used in a project to improve the 75 care processes with over 45 active care pathways. On this topic the Katholieke Universiteit Leuven and the Moscow Higher School of Economics have organized in the summer of 2011 a workshop with title "Concept discovery in Unstructured Data"[2]. Together with the Amsterdam-Amstelland Police Department will be considered whether CORDIET can be used to predict criminal careers of potential professional criminals.
The architecture of CORDIET includes 3 layers. The database layer consists of both the data storage as the ontology. The unstructured texts from the documents are indexed with Lucene[3] and the ontology elements in XML are translated to Lucene syntax. In the middle layer the FCA, ESOM, HMM and text analysis components are used to generate visual models based on the data and ontology. The third layer is the presentation layer with the graphical user interface. The graphical user interface will be developed in a way to perform complex analysis by users with little knowledge of statistics and data analysis. In the ontology, text mining attributes can be defined to analyze the documents. Temporal attributes can help to discover relationships over time. Compound attributes allow creating complex attributes composed of text mining attributes and temporal attributes using first order logic. For this specific ontological structures and the associated persistence (data storage), a new XML format will be defined. Parsers need to be developed to connect the working environment with the traditional data storage (SQL databases) and data warehouse systems. The generated models with the components from the middle layer will be used as follows:

- FCA concept lattices: detect human trafficking, terrorism, domestic violence, etc.
- TCA concept lattices: creation of visual profile of potential suspects and interesting patients.
- HMM: visualize care pathways and criminal careers.
- ESOM: used in combination with FCA to explore the data.

We want to mention that each of the four techniques are applied separately in one of more statistical environments like Matlab and SPSS, but have never been combined and implemented in one environment before. The consequence is that analysis with CORDIET can be applied on a larger scale, much faster and more efficient. The user interface allows to change the ontology elements by using a graph, tree structure and data display. The models can easily be generated and analyzed. Moreover, different extensions of FCA will be included, especially metrics like concept stability, etc.

[2] Concept Discovery in Unstructured Data 2011:
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-757/
[3] http://lucene.apache.org

6. Conclusions.

The three 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 information 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.

CHAPTER 1
INTRODUCTION

Formal Concept Analysis was originally introduced as a mathematical theory by Rudolf Wille in 1982. We performed a semantic text mining analysis on papers in which FCA was used by the authors from 2003 to 2009 and revealed FCA has found its way in numerous publications in knowledge discovery and information retrieval. We found a gap in the existing literature, today 80% to 90% of the information available in the police resides in textual form. We investigated the possibilities of FCA as a human-centered instrument for distilling new knowledge from these data. In 2005 the Amsterdam-Amstelland Police Department introduced Intelligence-led Policing, which has resulted in an increasing number of general reports every year. Until now, the general reports are hardly used by the criminal intelligence departments. Intelligence-led policing, as is defined by Ratcliffe (2008), does not show the dynamics of the Intelligence-led policing process. We introduce the Concept-Knowledge design theory to map the 3-i model of Ratcliffe on the design square of Hatchuell (2003). The design square is also used to illustrate the process of knowledge discovery of large amounts of unstructured police reports.

1.1 Concept Discovery

Concept discovery is a relatively new approach for discovering knowledge from textual information (Poelmans et al, 2010a). At the core of the method is the visualization of the underlying concepts of the data by means of Formal Concept Analysis (FCA) lattices (Ganter 1999, Wille 1982, 2005) which are interpreted, analyzed and discussed by domain experts. FCA arose twenty-five years ago as a mathematical theory (Stumme, 2002) and has over the years grown into a powerful framework for data analysis, data visualization (Priss 2000), information retrieval and text mining (Godin 1989, Carpinetto 2005, Priss 1997). In this thesis FCA is for the first time used as an exploratory data analysis and knowledge enrichment technique for police data. Compared to traditional black-box data mining techniques, this human-centered approach has the advantage of actively engaging expert knowledge in the discovery process.

Formal Concept Analysis was originally introduced as a mathematical theory by Rudolf Wille in 1982. Between the beginning of 2003 and the end of 2009, over 700 papers have been published in which FCA was used by the authors. We performed a semantic text mining analysis of these papers. We downloaded these 702 pdf-files and built a thesaurus containing terms related to FCA research. We used Lucene to index the abstract, title and keywords of these papers with this thesaurus. After clustering the terms, we obtained several lattices summarizing the most notorious FCA-related research topics. While exploring the literature, we found FCA to be an interesting meta-technique for clustering and categorizing papers in different research topics.

Over the years FCA has found its way from mathematics to computer science,

resulting in numerous applications in knowledge discovery (20% of papers), information retrieval (15% of papers), ontology engineering (13 % of papers) and software engineering (15% of papers). 18 % of the papers described extensions of traditional FCA such as fuzzy FCA and rough FCA.

In this thesis we filled in some of the gaps in the existing literature. During the past 20 years, the amount of unstructured data available for analysis has been ever-increasing. Today, 80% to 90% of the information available to police organizations resides in textual form. We investigated the possibilities of FCA as a human-centered instrument for distilling new knowledge from these data. FCA was found to be particularly useful for exploring and refining the underlying concepts of the data. To cope with scalability issues, we combined its use with Emergent Self Organising Maps. This neural network technique helped us gain insight in the overall distribution of the data and the combination with FCA was found to have significant synergistic results. The knowledge extraction process was framed in the C-K design theory. At the basis of the method are multiple successive iterations through the design square consisting of a concept and knowledge space. The knowledge space consists of the information used to steer the action environment, while this information is put under scrutiny in the concept space.

1.2 Intelligent-Led Policing, a historical overview

For the three past generations policing were overwhelming reactive in nature. Tilley (2003) calls this 'fire brigade' policing, where once

> "*the fire is put out, the case is dealt with and then the police withdraw to await the next incident that requires attention. There is nothing strategic about response policing. There are no long term objectives. There is no purpose beyond coping with the here and now"*.

During the 1970s groups of offenders bond together for mutual support and mutual protection, and their tentacles spread across different types of criminal endeavor. While organized crime has been discussed and perceived as a problem since the 1920s, the explosion in drug and people trafficking has propelled transnational organized crime into a problem that has been taken seriously only since the 1990s (Gill 2000). The recent change in complexity of modern criminality has had local implications. Local police are now unable to isolate themselves and fixate on local issues. As offenders learn and adapt, as their mobility increases and they cross jurisdictional boundaries to a greater extent now then at any time in history, the policing environment has become more complex and challenging.

Since the 1980s, the rapid digitalization of the rest of the world has not gone unnoticed within the sphere of policing. Computerized intelligence databases are now available to cross-reference information across numerous databases, search by name or keywords, and perform fuzzy searches of partial information, and new software can disseminate the results in a range of output formats such as link diagrams and maps. This has dramatically changed the nature of police intelligence practice by raising the volume of what can be accessed and integrated into an intelligence package.

Police services and departments around the world have all been affected to a greater or lesser degree by an environment that is more complex and accountability oriented, where demand outpaces resource availability, and where emerging threats to community safety present challenges for the traditional order of policing. The rising of the community policing and problem-oriented policing turns out to be the key drivers towards Compstat (Weisburd 2004) and Information-led Policing (Ratcliffe 2008).

Compstat began in the Crime Controle Strategy meetings of the New York City Police Department (NYPD) in January 1994. William Bratton, newly hired from the city's Transit Police by Mayor Rudy Giuliani, created Compstat with the primary aim of establishing accountability among the city's 76 police commanders (Magers 2004). The much published crime drop in New York around this time cemented the popular view that Compstat was responsible for making the city saver: major crime in the city fell by half from 1993 to 1998 (Walsh 2001). Compstat coincided with the digital explosion that reduced computing costs; and, finally police leaders were becoming more comfortable with professional management concepts.

The goal of Intelligence-led Policing (ILP) is to complement intuition led policing actions with information coming from analyses on aggregated operational data, such as crime figures and criminal characteristics (Collier 2004, 2006, Viaene et al 2009). Despite the fact ILP found its way in law enforcement organizations in different countries, there are as many definitions of ILP. Ratcliffe (2008) proposes a definition for Intelligence-led policing:

> "*Intelligent-led policing is a business model and managerial philosophy where data analysis and crime intelligence are pivotal to an objective, decision-making framework that facilitates crime and problem reduction and prevention through both strategic management and effective enforcement strategies that target profilic and serious offenders.*"

The pivotal subjects of the definition are data analysis and crime intelligence. Criminal intelligence is referred by the International Association of Law Enforcement Intelligence Analysts (IALEIA) as "*information compiled, analyzed, and/or disseminated in an effort to anticipate, prevent, or monitor criminal activity"* (IALEIA 2004:32). The definition of intelligence is later expanded to "*the product of gathering, evaluation, and synthesis of raw data on individuals or activities suspected of being, or known to be, criminal in nature. Intelligence is information that has been analyzed to determine its meaning and relevance*" (IALEIA 2004:33).

1.3 Intelligence-led policing and C-K modeling

In this section we propose a new modeling technique which can describe the process of Intelligence-led Policing. We first describe the 3-i model used by Ratcliffe (2008) and then describe how the intelligence led policing process fits in the Concept/Knowledge design theory.

1.3.1 3-i model of Ratcliffe

Ratcliffe introduced the 3i model which is shown in Figure 1.1.

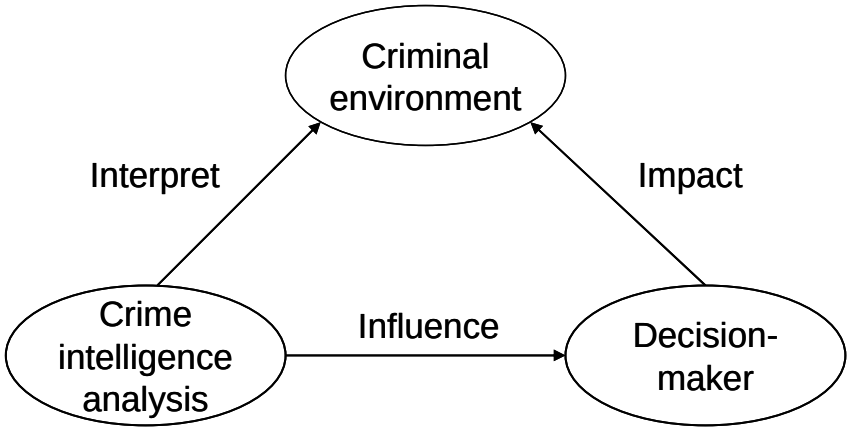


Fig. 1. 1 3-i model from Ratcliffe (2008)

The criminal environment is interpreted by the police analysts and results in several reports with crime figures and criminal characteristics. The reports are used by the police analysts to influence the decision makers to force an impact on the criminal environment. This does not only demands a well structured information architecture and tooling for the analysts, but also demands analysts to work closely with the decision makers, like police chiefs and both national and local government, who are able to control and direct resources. Many police organizations, like the Dutch police, share the view of Ratcliffe with Intelligence-led policing that the aim of Intelligence-led policing for police executives is *"to have a strategic overview of crime problems in their jurisdiction so that they can have better allocate resources to the most important crime priorities"* (Ratcliffe 2008).

Crucial is the link between of the crime intelligence analysis and the criminal environment. The idea of making knowledge actionable, which is the result of the interpretation and analysis of the criminal environment, and the basis concept of intelligence, is the main reason to introduce the Concept-Knowledge theory as process model for intelligence led policing.

1.3.2 Concept Knowledge theory

The Concept-Knowledge theory (C-K theory) was initially proposed by Hatchuel et al. (1999), Hatchuel et al. (2002) and further developed by Hatchuel et al. (2004). C-K theory is a unified design theory that defines design reasoning dynamics as a joint expansion of the Concept (C) and Knowledge (K) spaces through a series of continuous transformations within and between the two spaces (Hatchuel 2003). C-K theory makes a formal distinction between Concepts and Knowledge: the knowledge space consists of propositions with logical status (i.e. either true or false) for a designer, whereas the concept space consists of propositions without logical

status in the knowledge space. According to Hatchuel et al. (2003), concepts have the potential to be transformed into propositions of K but are not themselves elements of K. The transformations within and between the concept and knowledge spaces are realized by the application of four operators:

Concept → Knowledge, the conceptualization
Knowledge → Concept, the concept expansion
Concept → Concept, the concept activation and
Knowledge → Knowledge, the knowledge expansion.

These transformations form what Hatchuel calls the design square, which represents the fundamental structure of the design process. The last two operators remain within the concept and knowledge spaces. The first two operators cross the boundary between the Concept and Knowledge domains and reflect a change in the logical status of the propositions under consideration by the designer (from no logical status to true or false, and vice versa).

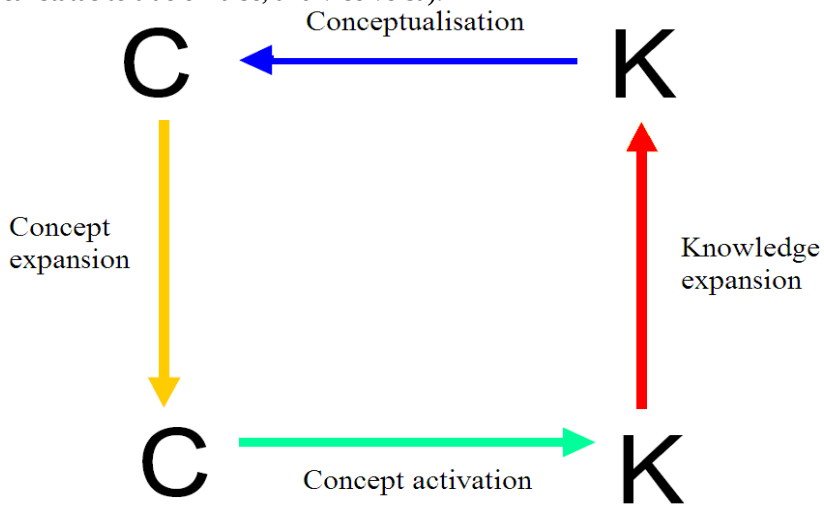


Fig. 1. 2 Design square (adapted from (Hatchuel 2003))

Design reasoning is modeled as the co-evolution of C and K. Proceeding from K to C, new concepts are formed with existing knowledge. A concept can be expanded by adding, removing or varying some attributes (a “partition” of the concept). Conversely, moving from C to K, designers create new knowledge either to validate a concept or to test a hypothesis, for instance through experimentation or by combining expertise. The iterative interaction between the two spaces is illustrated in Figure 1.2. The beauty of C-K theory is that it offers a better understanding of an expansive process. The combination of existing knowledge creates new concepts (i.e. conceptualization), but the activation and validation of these concepts may also generate new knowledge from which once again new concepts can arise.

Figure 1.3 demonstrates how the 3-i model of Ratcliffe can be framed in the C/K theory. The design reasoning process becomes an equivalence of the knowledge discovery process.

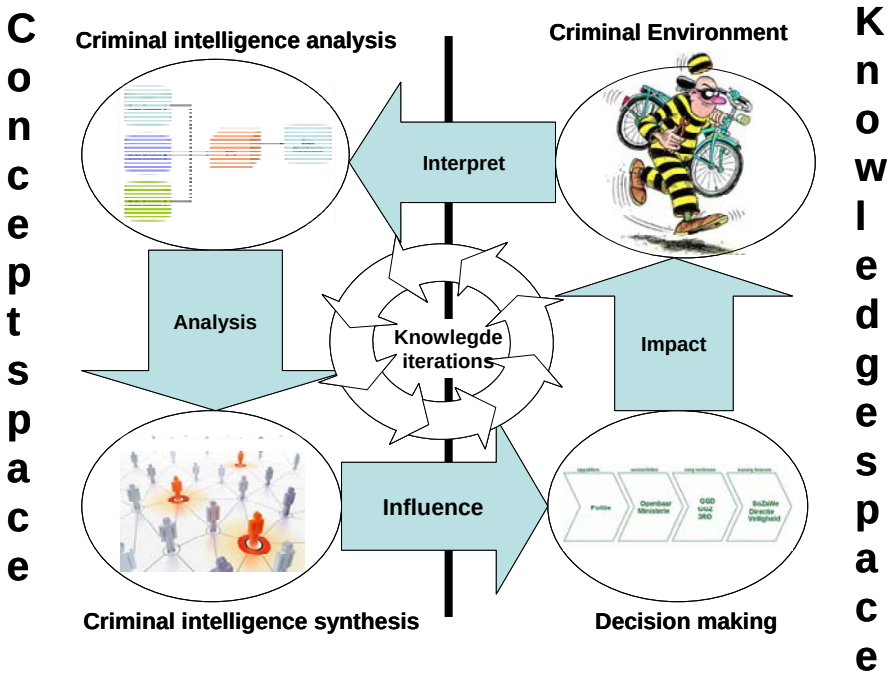


Fig. 1. 3 C/K modeling and Intelligence-led Policing

The first step is interpreting the criminal environment. The information about the criminal environment is transformed into information products, like ontologies, social media, data warehouses, law enforcement rules, etc. This is the conceptualization process, transforming knowledge into concepts. The next phase is analyzing the concepts and produce new concepts aiming at influencing the decision makers and getting impact on the criminal environment.

In fig 1.1 the criminal intelligence analysis process is shown as a single black box. To implement the C/K model, we divide the criminal intelligence analysis into two separate intelligence processes, the criminal intelligence analysis and the crime intelligence synthesis. The main motivation of this division is the fact that analysts synthesize new information from existing information (generating new concepts from existing concepts). The result of the synthesis is used to influence the decision makers. Producing new information (new concepts) can be seen as expanding the concept space. If the decision makers are influenced by the new information, this can be seen as concept activation. If the new information is used by the decision makers and gets impact on the criminal environment, then the new information has become actionable and this can be seen as knowledge expansion. Knowledge expansion is the equivalent of making knowledge actionable or creating intelligence.

In this thesis we will demonstrate how well the concept-design theory from Figure 1.3 fits in the overall knowledge discovery process, from data and domain analysis (chapter 3 and 4) to the design and implementation of the intelligence software (chapter 5).

1.4 Intelligence-led Policing and text mining

The change from reactive to proactive policing has led to an explosion of information. Officers are stimulated to report as many suspicious situations as possible. This information is stored in general reports, with the aim to inform other officers if it happens again, to collect new information to get a better picture. Opposed to general reports, there are incident reports such as a woman who come to the police and states that she was robbed in the red light district. Incident reports demands reactive policing. Information about incidents has more structure of what, when and how it has happened. These reports have incident labels like burglary, theft, fraud, and so on. General reports are lacking this specific information and are labeled as “attention reports”, “common reports” or “other reports”. A general report can be labeled with a project label like “domestic violence”, “prostitution” or “terrorism”, but this project label is not mandatory. 15% or less of the general reports do have a project label. Unknown is how many reports actually should have a project label. An example is the domestic violence case in chapter 3, where we developed an application to detect possible domestic violence cases. Having a project label or not depends on how well officers have been instructed, how much experience they have and most important of all, how well they are able to interpret and describe the suspicious situations.

Because most general reports lack a label about the suspicious event, the officers need to read the unstructured information to get a picture every time when needed. The unstructured information can not properly be used for data analysis and data mining by the Amsterdam-Amstelland Police Department (van der Veer 2009). This is really an issue, because the number of general information is growing year by year. Since 2005, the year when the ILP program was introduced at the Amsterdam-Amsterdam Police Department, the total number of general reports grows from 34,818 in 2005 to 40,703 in 2006, 53,583 in 2007, 69,470 in 2008 and 67,584 in 2009. Despite the increasing number of unstructured reports, there is no structured approach within the Dutch police to refine the information from the general reports into structured information and make it available to data analysis and data mining. It turns out to be very difficult to apply an automated text mining technique. Attempts were made with classifying, clustering and feature extraction with scientifically and commercial applications, but none of them had been successful and implemented into production.

This was the main motivation to start a pilot project in 2006, “textmining by fingerprints”. The first real life case study described in chapter 3 of this thesis zoomed in on the problem of domestic violence at the Amsterdam-Amstelland Police Department with FCA. This project has led to new insights how text could be structured. The human interaction in this process turns out to be crucial. Starting from the knowledge of an investigation domain, a thesaurus was built. The thesaurus has a structure of term clusters with search terms. A term cluster could be a family,

consisting of a collection with search teams of all family members (father, mother, sister, brother, etc). Another term cluster could be acts of violence, consisting of all violence terms. The next step was using a search engine which returns for each document a vector with the term clusters and search terms. We did the discovery that combinations of the term clusters with the collected reports gave interesting insights in the investigation area, like whether a case was a domestic violence case or not. Formal Concept Analysis is an unsupervised technique which clusters police reports based on the terms and term clusters they contain. We exposed multiple anomalies and inconsistencies in the data and were able to improve the employed definition of domestic violence. An important spin-off of this KDD exercise was the development of a highly accurate and comprehensible rule-based case labelling system. This system can be used to automatically assign a label to 75% of incoming cases whereas in the past all cases had to be dealt with manually.

Formal Concept analysis also has solved the problem of maintaining the thesaurus, because new emerging concepts can be found from the lattice. Another discovery made is the process of enriching and refining the thesaurus. This process has a cyclic nature of interacting with domain knowledge and domain concepts. After our domestic violence case study, we adapted the Concept space/Knowledge space design theory to structure our knowledge discovery process. We will show in this thesis, the combination of FCA and C/K is a very powerful methodology for criminal investigations.

For the analysis of other phenomena such as human trafficking and terrorism threat, a complicating factor is the inherent time dimension in the data. We applied the temporal variant of FCA, namely Temporal Concept Analysis (TCA), to the unstructured text in a large set of police reports. The aim was to distill potential subjects for further investigation. In both case studies, TCA was found to give interesting insights into the evolution of subjects over time. Amongst other things, several (to the police unknown) persons involved in human trafficking or the recruitment of future potential jihadists were distilled from the data. The intuitive visual interface allowed for an effective interaction between the police officer who used to be numbed by the overload of information, and the data.

Each of these projects helped us define the essential requirements of a generic text mining tool named CORDIET that would help in dealing with the challenges encountered by 21st century police organizations. CORDIET is currently under development by the Katholieke Universiteit Leuven, the Moscow Higher School of Economics and the Amsterdam-Amstelland Police Department and takes as input unstructured text documents and some additional structured information. The user can compose an ontology consisting of text mining attributes containing keywords to search and index these texts. Temporal attributes allow the user to work with the timestamps of the documents. Compound attributes are formulas that use first order logic to compose multiple ontology elements that should or should not be available in the texts. Using segmentation rules the data can be chopped in pieces and object-cluster rules are used to cluster individual documents. Then the user may compose an artifact such as an FCA lattice, ESOM map or HMM to browse through the data and gain new knowledge. CORDIET is described in detail in chapter 5.

CHAPTER 2

Formal concept analysis in the literature

In this chapter, we analyze the literature on Formal Concept Analysis (FCA) and some closely related disciplines using FCA[4]. We collected 702 papers published between 2003-2009 mentioning Formal Concept Analysis in the abstract. The toolset, a knowledge browsing environment which was initially developed to explore police reports and in detail described in chapter 5, was for this purpose extended to support our FCA literature analysis process. The pdf-files containing the papers were converted to plain text and indexed by Lucene using a thesaurus containing terms related to FCA research. We use the visualization capabilities of FCA to explore the literature, to discover and conceptually represent the main research topics in the FCA community. We zoom in on the papers published between 2003 and 2009 on using FCA in knowledge discovery and data mining, information retrieval, ontology engineering and scalability.

2.1 Introduction

Formal Concept Analysis (FCA) was invented in the early 1980s by Rudolf Wille as a mathematical theory (Wille 1982). FCA is concerned with the formalization of concepts and conceptual thinking and has been applied in many disciplines such as software engineering, knowledge discovery and information retrieved during the last 15 years. The mathematical foundation of FCA is described by Ganter et al. (1999) and introductory courses were written by Wolff (1994) and Wille (1997).

A textual overview of part of the literature published until the year 2004 on the mathematical and philosophical background of FCA, some of the applications of FCA in the information retrieval and knowledge discovery field and in logic and AI is given by Priss (2006). An overview of available FCA software is provided by Tilley (2004). Carpineto et al. (2004) present an overview of FCA applications in information retrieval. In Tilley et al. (2007), an overview of 47 FCA-based software engineering papers is given. The authors categorized these papers according to the 10 categories as defined in the ISO 12207 software engineering standard and visualized them in a concept lattice. In Lakhal et al. (2005), a survey on FCA-based association rule mining techniques is given.

In this chapter, we describe how we used FCA to create a visual overview of the existing literature on concept analysis published between the years 2003 and 2009. The core contributions of this chapter are as follows. We visually represent the literature on FCA using concept lattices, in which the objects are the scientific papers and the attributes are the relevant terms available in the title, keywords and

[4] Part of this chapter has been published in Poelmans, J, Elzinga, P., Viaene, S., Dedene, G. (2010) Formal Concept Analysis in Knowledge Discovery: s Survey. LNCS 6208, 139-153, 18th International Conference on Conceptual Structures.

abstract of the papers. The toolset of chapter 5 is used to generate the lattices. We zoom in on the papers published between 2003 and 2009 on using FCA in knowledge discovery and data mining, information retrieval, ontology engineering and scalability.

The remainder of this chapter is composed as follows. In section 2.2 we introduce the essentials of FCA theory and the knowledge browsing environment we developed to support this literature analysis. In section 2.3 we describe the dataset used. In section 2.4 we visualize the FCA literature on knowledge discovery, information retrieval, ontology engineering and scalability using FCA lattices. Section 2.5 concludes the chapter.

2.2 Formal Concept Analysis

2.2.1. FCA essentials

Formal Concept Analysis is a recent mathematical technique that can be used as an unsupervised clustering technique (Ganter et al. 1999, Wille 1982). Scientific papers containing terms from the same term-clusters are grouped in concepts. The starting point of the analysis is a database table consisting of rows M (i.e. objects), columns F (i.e. attributes) and crosses $T \subseteq M \times F$ (i.e. relationships between objects and attributes). The mathematical structure used to reference such a cross table is called a formal context (*T, M, F*). An example of a cross table is displayed in Table 2.1. In the latter, scientific papers (i.e. the objects) are related (i.e. the crosses) to a number of terms (i.e. the attributes); here a paper is related to a term if the title or abstract of the paper contains this term. The dataset in Table 2.1 is an excerpt of the one we used in our research. Given a formal context, FCA then derives all concepts from this context and orders them according to a subconcept-superconcept relation. This results in a line diagram (a.k.a. lattice).

Table. 2.1. Example of a formal context

	browsing	mining	software	web services	FCA	information retrieval
Paper 1	X	X	X		X	
Paper 2			X		X	X
Paper 3		X		X	X	
Paper 4	X		X		X	
Paper 5				X	X	X

The notion of concept is central to FCA. The way FCA looks at concepts is in line with the international standard ISO 704, that formulates the following definition: A concept is considered to be a unit of thought constituted of two parts: its extension and its intension (Ganter et al. 1999, Wille 1982). The extension consists of all objects belonging to the concept, while the intension comprises all attributes shared by those objects. Let us illustrate the notion of concept of a formal context using the data in Table 2.1. For a set of objects $O \subseteq M$, the common features can be identified, written $\sigma(O)$, via:

$$A = \sigma(O) = \{ f \in F \mid \forall o \in O : (o, f) \in T \}$$

Take the attributes that describe paper 4 in Table 2.1, for example. By collecting all reports of this context that share these attributes, we get to a set $O \subseteq M$ consisting of papers 1 and 4. This set *O* of objects is closely connected to set A consisting of the attributes "browsing", "software" and "FCA."

$$O = \tau(A) = \{ i \in M \mid \forall f \in A : (i, f) \in T \}$$

That is, *O* is the set of all objects sharing all attributes of *A*, and *A* is the set of all attributes that are valid descriptions for all the objects contained in *O*. Each such pair (*O, A*) is called a formal concept (or concept) of the given context. The set $A = \sigma(O)$ is called the intent, while $O = \tau(A)$ is called the extent of the concept (*O, A*). There is a natural hierarchical ordering relation between the concepts of a given context that is called the subconcept-superconcept relation.

$$(O_1, A_1) \subseteq (O_2, A_2) \Leftrightarrow (O_1 \subseteq O_2 \Leftrightarrow A_2 \subseteq A_1)$$

A concept $d = (O_1, A_1)$ is called a subconcept of a concept $e = (O_2, A_2)$ (or equivalently, *e* is called a superconcept of a concept *d*) if the extent of *d* is a subset of the extent of *e* (or equivalently, if the intent of *d* is a superset of the intent of *e*). For example, the concept with intent "browsing," "software", "mining" and "FCA" is a subconcept of a concept with intent "browsing", "software" and "FCA." With

reference to Table 2.1, the extent of the latter is composed of papers 1 and 4, while the extent of the former is composed of paper 1.

The set of all concepts of a formal context combined with the subconcept-superconcept relation defined for these concepts gives rise to the mathematical structure of a complete lattice, called the concept lattice of the context. The latter is made accessible to human reasoning by using the representation of a (labeled) line diagram. The line diagram in Figure 2.1, for example, is a compact representation of the concept lattice of the formal context abstracted from Table 2.1. The circles or nodes in this line diagram represent the formal concepts. It displays only concepts that describe objects and is therefore a subpart of the concept lattice. The shaded boxes (upward) linked to a node represent the attributes used to name the concept. The non-shaded boxes (downward) linked to the node represent the objects used to name the concept. The information contained in the formal context of Table 2.1 can be distilled from the line diagram in Figure 2.1 by applying the following reading rule: An object “g” is described by an attribute “m” if and only if there is an ascending path from the node named by “g” to the node named by “m.” For example, paper 1 is described by the attributes “browsing”, “software”, “mining” and “FCA.”

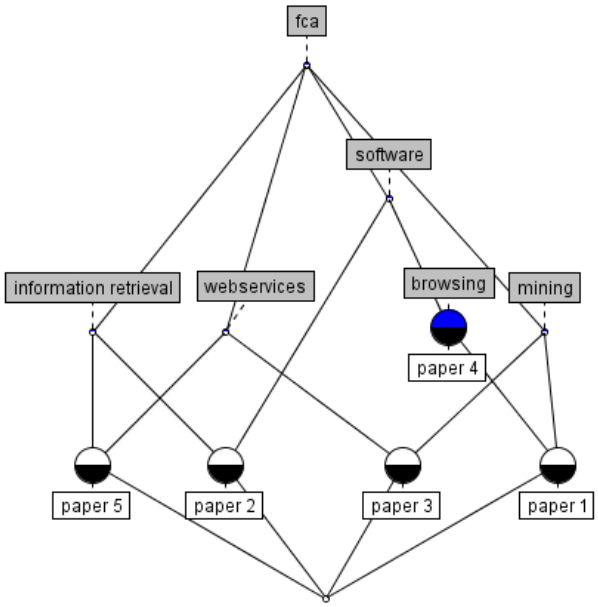


Fig. 2.1 Line diagram corresponding to the context from Table 2. 1

Retrieving the extension of a formal concept from a line diagram such as the one in Figure 2.1 implies collecting all objects on all paths leading down from the corresponding node. To retrieve the intension of a formal concept one traces all paths leading up from the corresponding node in order to collect all attributes. The

top and bottom concepts in the lattice are special. The top concept contains all objects in its extension. The bottom concept contains all attributes in its intension. A concept is a subconcept of all concepts that can be reached by travelling upward. This concept will inherit all attributes associated with these superconcepts.

2.2.2. FCA software

We developed a knowledge browsing environment to support our literature analysis process. One of the central components of our text analysis environment is the thesaurus containing the collection of terms describing the different research topics. The initial thesaurus was constructed based on expert prior knowledge and was incrementally improved by analyzing the concept gaps and anomalies in the resulting lattices. The thesaurus is a layered thesaurus containing multiple abstraction levels. The first and finest level of granularity contains the search terms of which most are grouped together based on their semantical meaning to form the term clusters at the second level of granularity.

An excerpt of this thesaurus is shown in Appendix A, which shows amongst others the termcluster “Knowledge discovery”. This term cluster contains search terms “data mining”, “KDD”, “data exploration”, etc. which can be used to automatically detect the presence or absence of the “Knowledge discovery” concept in the papers. Each of these search terms were thoroughly analyzed for being sufficiently specific. For example, we first had the term “exploration” for referring to the “Knowledge discovery” concept, however when used this term we found it also referred to the concepts “attribute exploration” etc. Therefore we only used the specific variant such as “data exploration”, which always refers to the “Knowledge discovery” concept. We aimed at composing term clusters that are complete, i.e. we search for all terms typically referring to for example the “Information retrieval” concept. Both specificity and completeness of search terms and term clusters was analyzed and validated with FCA lattices on our dataset. We only used abstract, title and keyword because the full text of the paper may mention a number of concepts that are irrelevant to the paper. For example, if the author who wrote an article on information retrieval gives an overview of related work mentioning papers on fuzzy FCA, rough FCA, etc., these concepts may be irrelevant however they are detected in the paper. If they are relevant tot the entire paper we found they were typically also mentioned in the title, abstract or keywords.

The papers that were downloaded from the World Wide Web (WWW) were all formatted in pdf. These pdf-files were converted to ordinary text and the abstract, title and keywords were extracted. The open source tool Lucene was used to index the extracted parts of the papers using the thesaurus. The result was a cross table describing the relationships between the papers and the term clusters or research topics from the thesaurus. This cross table was used as a basis to generate the lattices.

2.2.3. Web portal

We plan to host these 700 papers and the lattices to browse them on the internet. The concept lattices are expanded with hyperlinks to allow easy access to the papers. The user will be able to dynamically compose the lattices with his topics of interest.

2.3 Dataset

This Systematic Literature Review (SLR) has been carried out by considering a total of 702 papers related to FCA published between 2003 and 2009 in the literature and extracted from the most relevant scientific sources. The sources that were used in the search for primary studies contain the work published in those journals, conferences and workshops which are of recognized quality within the research community. These sources are:

- IEEE Computer Society
- ACM Digital Library
- Sciencedirect
- Springerlink
- EBSCOhost
- Google Scholar
- Conference repositories: ICFCA, ICCS and CLS conference

Other important sources such as DBLP or CiteSeerX were not explicitly included since they were indexed by some of the mentioned sources (e.g. Google Scholar). In the selected sources we used various search strings including "Formal Concept Analysis ", "FCA", "concept lattices", "Temporal Concept Analysis". To identify the major categories for the literature survey we also took into account the number of citations of the FCA papers at CiteseerX.

Perhaps the major validity issue facing this systematic literature review is whether we have failed to find all the relevant primary studies, although the scope of conferences and journals covered by the review is sufficiently wide for us to have achieved completeness in the field studied. Nevertheless, we are conscious that it is impossible to achieve total completeness in the field studied. Some relevant studies may exist which have not been included, although the width of the review and our knowledge of this subject have led us to the conclusion that, if they do exist, there are probably not many. We also ensured that papers that appeared in multiple sources were only taken into account once, i.e. duplicate papers were removed.

2.4 Studying the literature using FCA

The 702 papers are grouped together according to a number of features within the scope of FCA research. We visualized the papers using FCA lattices, which facilitate our exploration and analysis of the literature. The lattice in Figure 2.2 contains 7 categories under which 55% of the 702 FCA papers can be categorized. Knowledge discovery is the most popular research theme covering 20% of the papers and will be analyzed in detail in section 2.4.1. Recently, improving the scalability of FCA to larger and complex datasets emerged as a new research topic covering 5% of the 702 FCA papers. In particular, we note that almost half of the papers dedicated to this topic work on issues in the KDD domain. Scalability will be discussed in detail in section 2.4.3. Another important research topic in the FCA community is information retrieval covering 15% of the papers. 25 of the papers on information retrieval describe a combination with KDD approach and in 20 IR papers authors make use of ontology's. 15 IR papers deal with the retrieval of

software structures such as software components. The FCA paper on information retrieval will be discussed in detail in section 2.4.2. In 13% of the FCA papers, FCA is used in combination with ontology's or for ontology engineering. FCA research on ontology engineering will be discussed in section 2.4.4. Other important topics are using FCA in software engineering (15%) and for classification (7%).

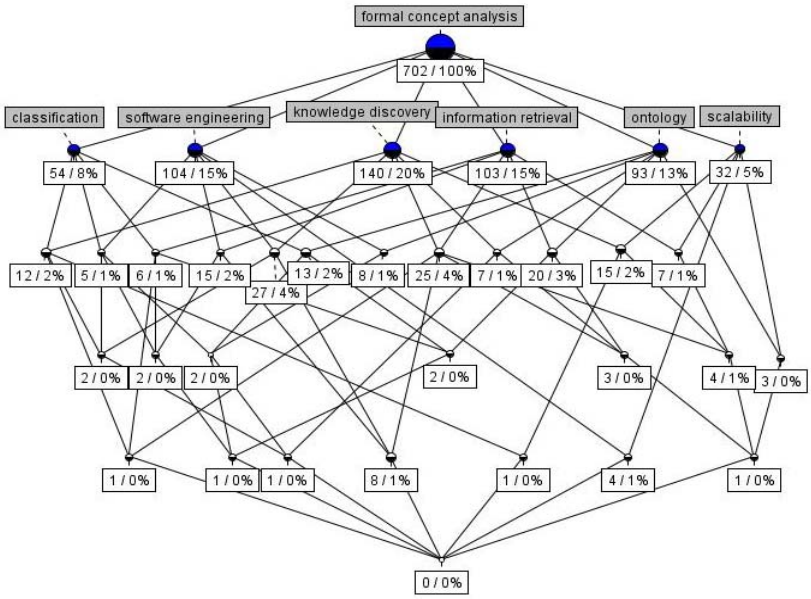


Fig. 2.2 Lattice containing 702 papers on FCA

2.4.1 Knowledge discovery and data mining

Knowledge discovery and data mining (KDD) is an interdisciplinary research area focusing upon methodologies for extracting useful knowledge from data. In the past, the focus was on developing fully automated tools and techniques that extract new knowledge from data. Unfortunately, these techniques allowed almost no interaction between the human actor and the tool and failed at incorporating valuable expert knowledge into the discovery process (Keim 2002), which is needed to go beyond uncovering the fool's gold. These techniques assume a clear definition of the concepts available in the underlying data which is often not the case. Visual data exploration (Eidenberger 2004) and visual analytics (Thomas et al. 2005) are especially useful when little is known about the data and exploration goals are vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary.

In Conceptual Knowledge Processing (CKP) the focus lies on developing methods for processing information and knowledge which stimulate conscious reflection, discursive argumentation and human communication (Wille 2006). The word "conceptual" underlines the constitutive role of the thinking, arguing and communicating human being and the term "processing" refers to the process in

which something is gained which may be knowledge. An important subfield of CKP is Conceptual Knowledge Discovery (Stumme 2003). FCA is particularly suited for exploratory data analysis because of its human-centeredness (Correira et al. 2003). The generation of knowledge is promoted by the FCA representation that makes the inherent logical structure of the information transparent. The philosophical and mathematical origins of using FCA for knowledge discovery have been briefly summarized in Priss (2006). The system TOSCANA has been used as a knowledge discovery tool in various research and commercial projects (Stumme et al. 1998).

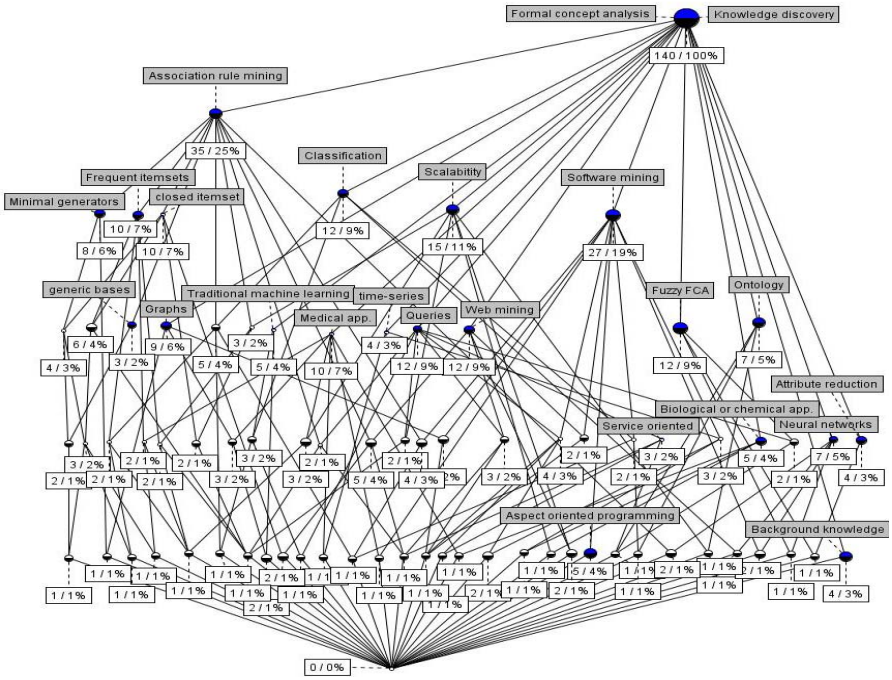


Fig. 2.3 Lattice containing 140 papers on using FCA in KDD

About 74% of the FCA papers on KDD are covered by the research topics in Figure 2.3. 35 papers (25%) are in the field of association rule mining. 19% of the KDD papers focus on using FCA in the discovery of structures in software. 9% of the papers describes applications of FCA in web mining. 11% of papers discuss some of the extensions of FCA theory for knowledge discovery. 10% of the KDD papers describe applications of FCA in biology, chemistry and medicine. The relation of FCA to some standard machine learning techniques is investigated in about 4% of papers. The applications on using Fuzzy FCA for KDD cover 9% of the papers.

2.4.2 Information retrieval

According to Manning et al. (2008), information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Information retrieval used to be an activity that only a few people engaged in: librarians, paralegals and similar professional searchers. The world has changed and hundreds of millions of people engage in information retrieval these days when they use a web search engine or search their email. Information retrieval systems can be distinguished by the scale at which they operate, and it is useful to distinguish three prominent scales. In web search, the system has to provide search over billions of documents stored on millions of computers. At the other extreme is personal information retrieval In the last few years, consumer operating systems have integrated information retrieval Email programs usually not only provide search but also text classification. In between is the space of enterprise institutional, and domain-specific search, where retrieval might be provided for collections such as a corporation's internal documents, a database of patents, etc.

The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. Given a set of topics, standing information needs, or other categories, classification is the task of deciding which classes, each of a set of documents belongs to.

The first attempts to use lattices for information retrieval are summarized in Priss (2000), but none of them resulted in practical implementations. Godin et al. (1989) developed a textual information retrieval system based on document-term lattices but without graphical representations of the lattices. The authors also compared the system's performance to that of Boolean queries and found that it was similar to and even better than hierarchical classification (Godin et al. 1993). They also worked on software component retrieval (Mili et al. 1997). In Carpineto et al. (2004), their work on information retrieval was summarized. They argue that FCA can serve three purposes. First, FCA can support query refinement. Because a document-term lattice subdivides a search space into clusters of related documents, lattices can be used to make suggestions for query enlargement in cases where too few documents are retrieved and for query refinement in cases where too many documents are retrieved. Second, lattices can support an integration of querying and navigation (or browsing). An initial query identifies a start node in a document-term lattice. Users can then navigate to related nodes. Further, queries are then used to "prune" a document-term lattice to help users focus their search (Carpineto et al. 1996). Third, a thesaurus hierarchy can be integrated with a concept lattice, an idea which was independently discussed by different researchers (e.g. Carpineto et al. 1996, Skorsky 1997, Priss 1997). For many purposes, some extra facilities are needed: process large document collections quickly, allow more flexible matching operations and allow ranked retrieval.

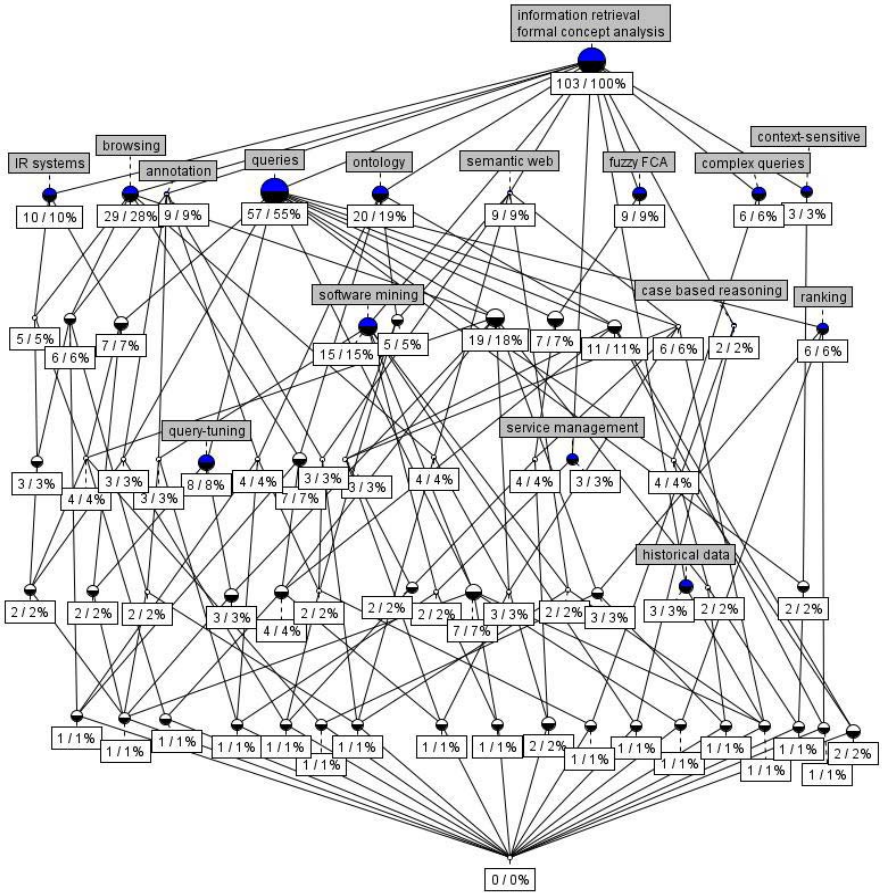


Fig. 2.4 Lattice containing 103 papers on using FCA in IR

86 % of the papers on FCA and information retrieval are covered by the research topics in Figure 2.4. 28 % of papers are about using FCA for representation of and navigation in document collections. The IR systems that were developed based on FCA cover 10 % of the papers. Query tuning and query result improvement covers 8% of the papers. Defining and processing complex queries covers 6% of the papers. The papers on contextual answers (6% of papers) and ranking of query results (6% of papers) cover 12% of the total amount. Finally 9% of papers are on fuzzy FCA in IR.

2.4.3 Scalability

At the international Conference on Formal Concept Analysis in Dresden (ICFCA 2006) an open problem of “handling large contexts" was pointed out. Since then, several studies have focused on the scalability of FCA for efficiently handling large and complex datasets. Many techniques have been devised including nested line diagrams for zooming in and out of the data, conceptual scaling for transforming many-valued contexts into a single-valued context, iceberg lattices and pruning strategies to reduce the size of the concept lattice, etc.

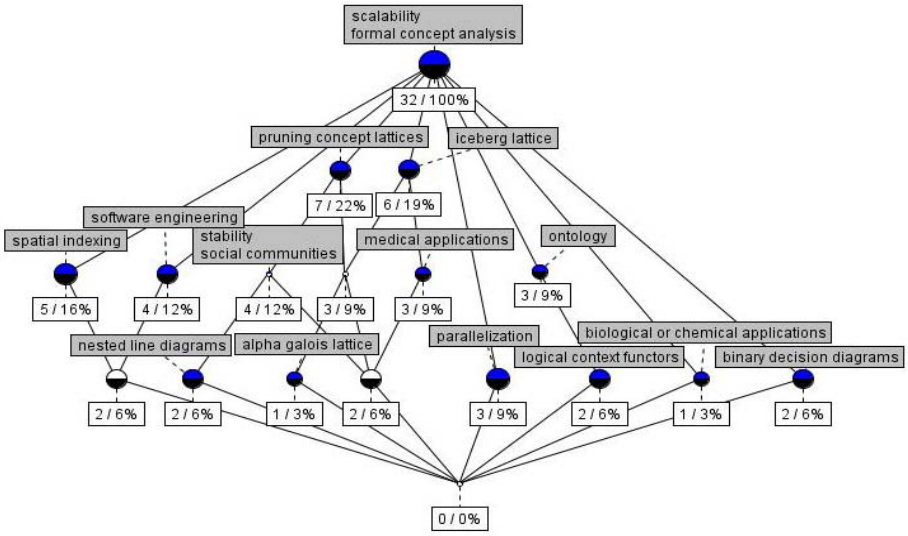


Fig. 2.5 Lattice containing 32 papers on FCA and scalability

81% of the papers on FCA and scalability are covered by the research topics in Figure 2.5. 19% of these papers use iceberg lattices. 22% papers are on reducing the size of concept lattices. 9% of papers discuss parallelization and 6% of papers the combination with binary decision diagrams and 16% of papers spatial indexing for improving the scalability of FCA-based algorithms.

2.4.4 Ontologies

Ontology’s were introduced as a means of formally representing knowledge. Their purpose is to model a shared understanding of the reality as perceived by some individuals in order to support knowledge intensive applications (Gruber 2009). An ontology typically consists of individuals or objects, classes, attributes, relations between individuals and classes or other individuals, function terms, rules, axioms, restrictions and events. The set of objects that can be represented is called the universe of discourse. The axioms are assertions in a logical form that together comprise the overall theory that the ontology describes in its domain of application. Ontologies are typically encoded using ontology languages, such as the Ontology Web Language (OWL). Whereas ontologies often use hierarchical representations

for modeling the world, FCA has the benefit of a non-hierarchical partial order representation which has a larger expressive power (Christopher 1965). A key objective of the semantic web is to provide machine interpretable descriptions of web services so that other software agents can use them without having any prior "built-in" knowledge about how to invoke them. Ontologies play a prominent role in the semantic web where they provide semantic information for assisting communication among heterogeneous information repositories

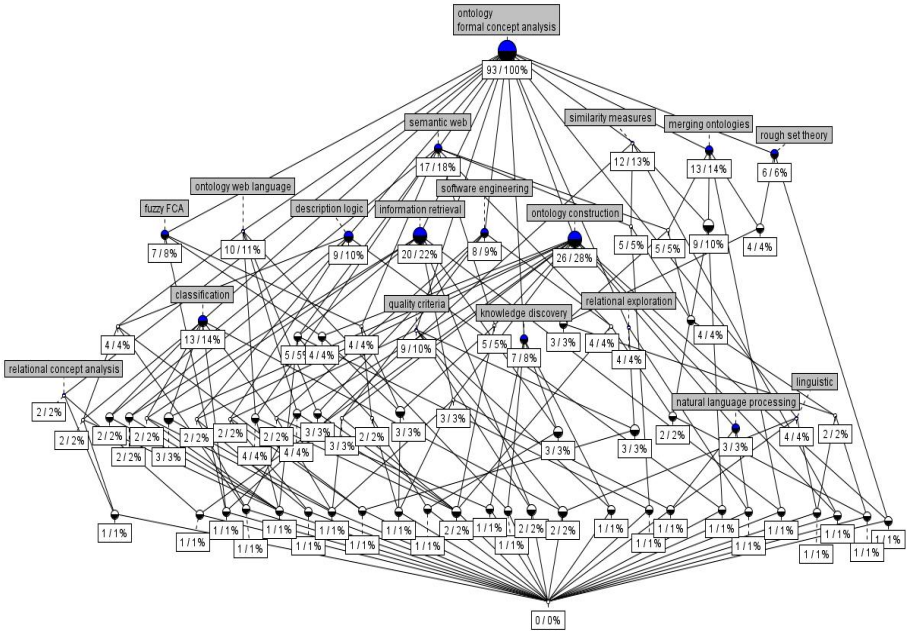


Fig. 2.6 Lattice containing 93 papers on FCA and ontologies

84 % of the FCA papers on ontologies are covered by the research topics in Figure 2.6. The construction of ontologies using FCA, covers 28% of the 93 papers. 10% of the papers are about improving the quality of ontology's. 6% of the papers describe linguistic applications of FCA and ontologies or the combination with natural language processing. 17% of papers are on developing FCA-based similarity measures and using FCA in ontology mapping and merging. 14% of the papers use rough set theory or fuzzy theory in combination with FCA for ontology construction or merging.

2.5 Conclusions

Since its invention in 1982 as a mathematical technique, FCA became a well-known instrument in computer science. Over 700 papers have been published over the past 7 years on FCA and many of them contained case studies showing the method's usefulness in real-life practice. This chapter showcased the possibilities of FCA as a

Meta technique for categorizing the literature on concept analysis. The intuitive visual interface of the concept lattices allowed for an in-depth exploration of the main topics in FCA research. In particular, its combination with text mining methods resulted in a powerful synergy of automated text analysis and human control over the discovery process.

One of the most notorious research topics covering 20% of the FCA papers is KDD. FCA has been used effectively in many domains for gaining actionable intelligence from large amounts of information. Information retrieval is another important domain covering 15% of the papers. FCA was found to be an interesting instrument for representation and navigation in large document collections. Multiple IR systems resulted from this research. FCA was also used frequently (13% of papers), amongst others in the context of semantic web, for ontology engineering and merging. Finally, 5% of the papers devoted attention to improve FCA's applicability to larger data repositories.

In 18% of the papers, traditional concept lattices were extended to deal with uncertain, three-dimensional and temporal data. In particular, combining FCA with fuzzy and rough set theory received considerable attention in the literature. Temporal and Triadic Concept Analysis received only minor attention.

CHAPTER 3

Curbing domestic violence: Instantiating C-K theory with Formal Concept Analysis and Self Organizing Maps

In this chapter we propose a human-centered process for knowledge discovery from unstructured text that makes use of Formal Concept Analysis and Emergent Self Organizing Maps[5]. The knowledge discovery process is conceptualized and interpreted as successive iterations through the C-K theory design square .To illustrate its effectiveness, we report on a real-life case study of using the process at the Amsterdam-Amstelland Police Department in the Netherlands aimed at distilling concepts to identify domestic violence from the unstructured text in actual police reports. The case study allows us to show how the process was not only able to uncover the nature of a phenomenon such as domestic violence, but also enabled analysts to identify many types of anomalies in the practice of policing. We will illustrate how the insights obtained from this exercise resulted in major improvements in the management of domestic violence cases and has replaced the knowledge rule "missing domestic violence label" of the in-triage system Trueblue.

3.1 Introduction

In this chapter we propose a human-centered process for knowledge discovery from unstructured text that makes use of Formal concept Analysis (FCA) (Wille 1982, Ganter 1999) and Emergent Self Organizing Maps (ESOM) (Ultsch et al. 2005a, Ultsch et al. 2005b). Human-centered KDD refers to the constitutive nature of human interpretation for the discovery of knowledge, and stresses the complex, interactive process of KDD as being led by human thought (Brachman et al. 1996). Data mining should be primarily concerned with making it easy, practical and convenient to explore very large databases for organizations and users with vast amounts of data but without years of training as data analysts (Fayyad 2002). A significant part of the art of data mining is the user's intuition with respect to the tools (Pednault 2000, Marchionini 2006).

Visual data exploration (Eidenberger 2004) and visual analytics (Thomas 2005) are especially useful when little is known about the data and exploration goals are vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary. In addition to the direct involvement of the user, the main advantages of visual data exploration over automatic data mining techniques from statistics or machine learning are: visual data exploration can easily deal with highly non-homogeneous and noisy data and visual

[5] Jonas Poelmans (2010) "Essays on using Formal Concept Analysis in information engineering", Katholieke Universiteit Leuven, PhD thesis. Chapter 3 was joint work between Paul Elzinga and Jonas Poelmans.

data exploration usually allows a faster data exploration and often provides better results, especially in cases where automatic algorithms fail. In addition, visual data exploration techniques provide a much higher degree of confidence in the findings of the exploration (Keim 2002).

This chapter extends but also synthesizes our previous work involving FCA and ESOM, two visually appealing data exploration aids, for knowledge discovery from unstructured text. In (Poelmans 2008), we first discussed the possibilities of using FCA for knowledge discovery in a police environment. A parallel research track consisted of investigating the potential of using ESOM for knowledge discovery. Our first findings using the ESOM are discussed in (Poelmans 2009a, Poelmans 2009c). We also compared ESOM's performance to that of other SOM's such as the spherical SOM and we found it to be superior (Poelmans 2009b). In (Poelmans 2009d), we briefly presented our idea to use FCA and ESOM together for domestic violence discovery. The ESOM functions as a catalyst for the FCA based discovery process. The proposed methodology recognizes the important role of the domain expert in mining real-world enterprise applications and makes efficient use of specific domain knowledge, including human intelligence and domain-specific constraints.

We chose for a semi-automated approach since the major drawback of all automated and supervised machine learning techniques, including decision trees, is that these algorithms assume that the underlying concepts of the data are clearly defined, which is often not the case. These techniques allow almost no interaction between the human actor and the tool and fail at incorporating valuable expert knowledge into the discovery process (Keim 2002), which is needed to go beyond uncovering the fool's gold (Smyth 2002). In the paper presented by Hollywood et al. (2009) these problems were clearly addressed in the context of terrorist threat assessment. The central question was whether it is possible to find terrorists with traditional automated data mining techniques and the answer was no.

The knowledge discovery process is conceptualized and interpreted as successive iterations through the C-K theory design square. C-K theory offers a formal framework that interprets existing design theories as special cases of a unified model of reasoning (Hatchuel 1996, Hatchuel 2002). It provides a clear and precise definition of design that is independent of any domain of professional tradition (Hatchuel 1999). C-K theory defines design reasoning dynamics as a joint expansion of the Concept (C) and Knowledge (K) spaces through a series of continuous transformations within and between the two spaces. The beauty of C-K theory is that it can provide insight into an iterative and expansive knowledge acquisition process (Hatchuel 2003, Hatchuel 2004)). One of the core characteristics of C-K theory is this focus on human intelligence as the driving force in expanding the space of knowledge. To our knowledge, this is the first systematic application of C-K theory to the information systems domain. C-K theory is used as a unifying framework to provide a clear structure to the discovery process based on FCA and ESOM. The combined use of FCA and ESOM in the C-K framework not only gives insight into the generic nature of the KDD activity but also makes for significantly improved results. Some of the aspects of this chapter have already been discussed in the

literature in a fragmented way (e.g. information retrieval, knowledge browsing, prior knowledge incorporation), but an integrated approach has never been pursued.

To illustrate its effectiveness, we report on a real-life case study on using the process at the Amsterdam-Amstelland Police Department in the Netherlands aimed at distilling concepts for domestic violence from the unstructured text in filed reports. The aim of our research was to conceptualize and improve the definition and understanding of domestic violence with the ultimate goal of improving the detection and handling of domestic violence cases. One important spin-off of this exercise that will be elaborated on in this paper was the development of a highly accurate and comprehensible classification procedure for automatically raising a domestic violence flag for incoming police reports. This procedure automatically classifies 91% of incoming cases correctly whereas in the past all cases had to be dealt with manually. We performed this classification exercise to measure the quality of our conceptualization of domestic violence. We have never seen a similar set up in the literature and to the best of our knowledge there is no packaged automated solution to do all the same at once.

Over 90% of the information available to police organizations is stored as plain text. To date, however, analyses have primarily focused on the structured portion of the available data. Only recently the first steps have been taken to apply text mining in criminal analysis (Chen 2004, Ananyan 2002). Domestic violence is one of the top priorities of the Amsterdam-Amstelland Police Department in the Netherlands (Politie Amsterdam-Amstelland 2009). In the past, intensive audits of the police databases of filed reports established that many of the reports tended to be wrongly labeled as domestic or as non-domestic violence cases. Previous attempts have mainly focused on developing a machine learning classifier that automatically classified incoming cases as domestic or as non-domestic violence. Unfortunately they were unsuccessful because the underlying concept of domestic violence was never challenged. These systems did not provide any insight into the problem, since they are black-boxes and their classification performance was around 80% only (Elzinga 2006). As a consequence, these systems never made it into operational policing practice. All of these previous attempts had in common that the concept of domestic violence was never challenged. The developers overlooked the complexity of the notion of domestic violence, were unaware that different people have different visions about the nature and scope of it and did not pay attention to niche cases. Moreover, the correctness of the labels assigned to cases by police officers was never verified. We found that different police officers regularly assigned different labels to the same situation. Finally, the developers did not dispose of a high-quality domain-specific thesaurus that contained sufficient discriminant terms for accurately classifying cases. In the past, several automated term extraction and thesaurus building techniques were used (Elzinga 2006). We interviewed several domain experts that were exposed to these efforts. All of them attested to their failure in constructing a useful thesaurus when we asked them for their appraisal of these prior initiatives.

The remainder of this chapter is composed as follows. In section 3.2 we discuss intelligence led policing, domestic violence and the motivation for this research. In section 3.3, we elaborate on the essentials of FCA, ESOM and C-K theory. In

section 3.4, we show how we used the synergistic combination of FCA and ESOM to institute the C-K framework. Section 3.5 then discusses the dataset, while section 3.6 showcases the knowledge discovery process and the four C-K operators described in section 3.3. In section 3.7, we summarize the actionable results of the iterative knowledge enrichment. Section 3.8 contains a comparative analysis of ESOM and multi-dimensional scaling. Finally, section 3.9 presents the main conclusions of this chapter.

3.2 Intelligence Led Policing

Policing is a knowledge intensive affair. Over the past fifteen years or so there have been calls for a shift from a more traditional reactive intuition led style of policing to a more proactive intelligence led approach (Collier 2006). Intelligence Led Policing (ILP) promotes this use of factual, evidence based information and analyses to provide management direction and to guide police actions at all levels of a police organization. The goal is specifically to complement intuition led police actions with information coming from analyses on aggregated operational data, such as crime figures and criminal characteristics (Collier 2004). While over 80% of all information available to police organizations resides in textual form, analysis has to date been primarily focused on the structured portion of the available data. Only recently the first steps for applying text mining in criminal analysis have been taken. Though text mining has been identified as a promising area in the formal framework for crime data mining by Chen et al. (2004), this work has hardly found its way into mainstream scientific literature. One of the notorious exceptions is the paper by Ananyan (2002) in which historical police reports were analyzed to identify hidden patterns.

In 1997, the Ministry of Justice of the Netherlands made its first inquiry into the nature and scope of domestic violence (Van Dijk 1997). 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. Acting firmly against this phenomenon became one of the pivotal projects of the Balkenende administration when it took office in 2003.

Domestic violence is nowadays one of the top priorities of the police organization of the region Amsterdam-Amstelland in the Netherlands (Politie Amsterdam-Amstelland 2009). Of course, in order to pursue an effective policy against offenders, being able to swiftly recognize cases of domestic violence and label reports accordingly is of the utmost importance. Still, this has proven to be problematic. In the past intensive audits of the police databases related to filed reports established that many reports tended to be wrongly classified as domestic or as non-domestic violence cases.

3.2.1 Domestic violence

According to the U.S. Office on Violence against Women, domestic violence is a "*pattern of abusive behavior in any relationship that is used by one partner to gain or maintain power and control over another intimate partner*" (Office on Violence

against Women 2007). Domestic violence can take the form of physical violence, which includes biting, pushing, maltreating, stabbing or even killing the victim. Physical violence is often accompanied by mental or emotional abuse, which includes insults and verbal threats of physical violence towards the victim, the self or others, including children. Domestic violence occurs all over the world, in various cultures (Watts 2002) and affects people throughout society, irrespective of economic status (Waits 1985).

The BVH database – the database of the Amsterdam-Amstelland Police Department – contains all documents with regard to criminal offences. Documents related to certain types of crime receive corresponding labels. It is of the utmost importance that a correct label is assigned to each of the filed police reports. First, there are some legal consequences. If the police judged an incident to be domestic violence, the public prosecutor can accuse the offender of committing a domestic violence crime. This is taken into account by the judge as an aggravating circumstance, often resulting in a more severe penalty. Second, police officers will be able to better assess new incidents between the perpetrator and the victim, resulting in a more effective way of tackling the problem. Finally, if a domestic violence label was incorrectly assigned to a case, this will result in a waste of the valuable time of the police officers assigned to the case.

Immediately after the reporting of a crime, police officers are given the possibility to judge whether or not it is a domestic violence case. If they believe it is, they can indicate this by assigning the label “domestic violence” to the report. However, not all domestic violence cases are recognized as such by police officers. This may have several reasons, for example, because of a lack of training, a lack of prior experience or new types of domestic violence occurring. As a consequence, many documents are lacking the appropriate label, which put on the agenda the need for a more efficient and effective case triage software program to automatically filter out suspicious cases for in-depth, manual inspection and classification. The in-place case triage system has been configured to filter out these reports for in-depth manual inspection and classification, with the aim of substantially reducing the number of domestic violence cases that are not recognized as such. It retrieves suspicious cases that lack the label of domestic violence and sends them back to the data quality management team. At present, each case retrieved by the in-place case triage system is subjected to an in-depth manual inspection by one of the co-workers of the quality control department. If analysis reveals that a case was wrongly classified as non-domestic violence, it is sent back to the police officer responsible for the case, who is obliged to re-examine and reclassify the police report. It is obvious that this is a very time-consuming and, by consequence, costly procedure. Given that it takes an individual at least five minutes to read and classify a case, it is clear that more accurate triage will result in major savings.

Currently the triage is based on either one or both of the following two criteria being met. The first criterion is whether the perpetrator and the victim live at the same address. The second criterion is whether any or a combination of the following expressions appear in the case documents: “ex-boyfriend”, “ex-girlfriend”, “ex-husband”, “ex-wife”, “domestic”, “stalk”, “lived together”, “live together”, “son and

scared", "child and scared", "child and threat", "son and threat", "daughter and threat" or "daughter and scared".

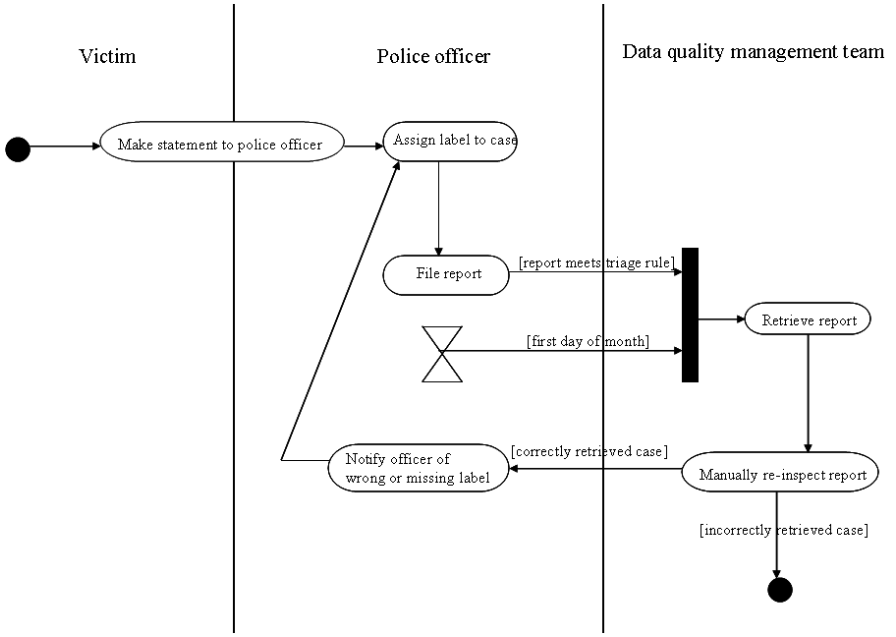


Fig. 3. 1 Current domestic violence reporting procedure

A summary of the current domestic violence reporting procedure is displayed in Figure 3.1. There are several problems associated with this process. First, recent audits have confirmed that many of the retrieved cases are wrongly selected for in-depth manual inspection. Going back to 2006, the system retrieved 1157 cases, 80% of which actually turned out to be non-domestic violence cases. For example, going back to 2007, the triage system retrieved 1091 of such cases in which the victim made a statement to the police. Second, because of a lack of manpower the data management quality team was not able to analyze each retrieved police report. Third, audits of the police databases revealed that not all domestic violence cases lacking the appropriate label were retrieved by the case triage system. Fourth, no actions have yet been undertaken to address the issue of the filed reports that were wrongly classified as domestic violence.

3.2.2 Motivation

According to R.S. Brachman et al. (1996), much attention and effort has been focused on the development of data mining techniques, but only a minor effort has been devoted to the development of tools that support the analyst in the overall discovery task. They argue for a more human-centered approach. Human-centered KDD refers to the constitutive character of human interpretation for the discovery of knowledge, and stresses the complex, interactive process of KDD as being led by human thought. In most real-world knowledge discovery applications, an

indispensable part of the discovery process is that the analyst explores and sifts through the raw data to become familiar with it and to get a feel for what the data may cover. Often an explicit specification of what one is looking for only arises during an interactive process of data exploration, analysis and segmentation. R.S. Brachman et al. (1993) introduce the notion of data archeology for KDD tasks in which a precise specification of the discovery strategy, the crucial questions and the basic goals of the task have to be elaborated during an unpredictable exploration of the data. Data archeology can be considered as a highly human-centered process of asking, exploring, analyzing, interpreting and learning by interacting with the underlying database. Comprehensible support should be provided to the analyst during the KDD process. According to Brachman et al. (1996) this should be embedded into a knowledge discovery support environment. How the process of human-centered KDD can be supported by Formal Concept Analysis (FCA) was for the first time investigated by Stumme et al. (1998).

Smyth et al. (2002) already stated that the algorithm designer and the scientist should be able to bring in prior knowledge so the data mining algorithm does not just rediscover what is already known. Moreover, the scientist should be able to "get inside" and "steer" the direction of the data mining algorithm. FCA fulfils these requirements. Starting from initial knowledge on the problem area, it provides the user with a visual display of the relevant concepts available in the dataset and their relationships. Additionally, the user can visually interact with the concept lattice and thereby steer the knowledge discovery process.

What makes FCA into an especially appealing technique for knowledge discovery in databases is that it meets the important requirement stated by, amongst others, Fayyad et al. (2002) that data mining should be primarily concerned with making it easy, convenient and practical to explore very large databases for organizations and users with vast amounts of data but without years of training as data analysts. FCA offers the user an intuitive visual display of different types of structures available in the dataset and guides the user in the exploration of the dataset. This end-user-friendly interface also makes the data mining more transparent to the user.

When compared to other, more traditional, techniques such as association rules, FCA has a larger explanatory power because of its underlying non-hierarchical structure (Christopher 1965). While traditional association rules are flat, FCA provides an order of significance, which makes its representation richer and more intuitive to use.

3.3 FCA, ESOM and C-K theory

3.3.1 Formal Concept Analysis

FCA arose twenty-five years ago as a mathematical theory (Ganter 1999, Stumme 2002b) and has over the years grown into a powerful tool for data analysis, data visualization and information retrieval (Priss 2005). The usage of FCA for browsing text collections has been suggested before by Cole et al. (2001). However, none of the papers have focused on how FCA can be used in an actionable environment for knowledge enrichment and for discovering different types of knowledge in

unstructured text. FCA has been applied in a wide range of domains, including medicine, psychology, social sciences, linguistics, information sciences, machine and civil engineering, etc (Stumme 2000). For instance, FCA has been applied for analyzing data of children with diabetes (Scheich 1993), for developing qualitative theories in music esthetics (Hereth 2000), for database marketing (Hereth 2000), and for an IT security management system (Becker 2000). In (Eklund 2004, Domingo 2005), FCA was used as a visualization technique that allows human actors to quickly gain insight by browsing through information. Full details on the use of FCA in KDD are given in chapter 2.

We previously applied FCA to a relatively small police dataset and were able to establish its practical usefulness (Poelmans 2008). FCA is particularly suited for exploratory data analysis because of its human-centeredness (Correira 2003, Valtchev 2004). It is a fundamental principle that the generation of knowledge from information is promoted by representations that make the inherent logical structure of the information transparent. FCA builds on the model that concepts are the fundamental units of human thought. Hence, the basic structures of logic and logical structure of information are based on concepts and concept systems (Stumme 1998, Stumme 2002a). Consequently, FCA uses the mathematical abstraction of the concept lattice to describe systems of concepts to support human actors in their information discovery and knowledge creation practice (Wille 2002).

Again, the starting point of the analysis is a database table consisting of rows M (i.e. objects), columns F (i.e. attributes) and crosses $T \subseteq M \times F$ (i.e. relationships between objects and attributes). The mathematical structure used to reference such a cross table is called a formal context (M, F, T). An example of a cross table is displayed in Table 3.1. Here, reports of domestic violence (i.e. the objects) are related (i.e. the crosses) to a number of terms (i.e. the attributes): a report is related to a term if the report contains this term. The dataset in Table 3.1 is an excerpt from the one we used in our research. Given a formal context, FCA then derives all concepts from this context and orders them according to a subconcept-superconcept relation, which results in a line diagram (a.k.a. lattice).

Table. 3.1. Example of a formal context

	kicking	dad hits me	stabbing	cursing	scratching	maltreating
report 1	X	X				X
report 2			X	X	X	
report 3	X	X	X	X	X	
report 4						X
report 5				X	X	

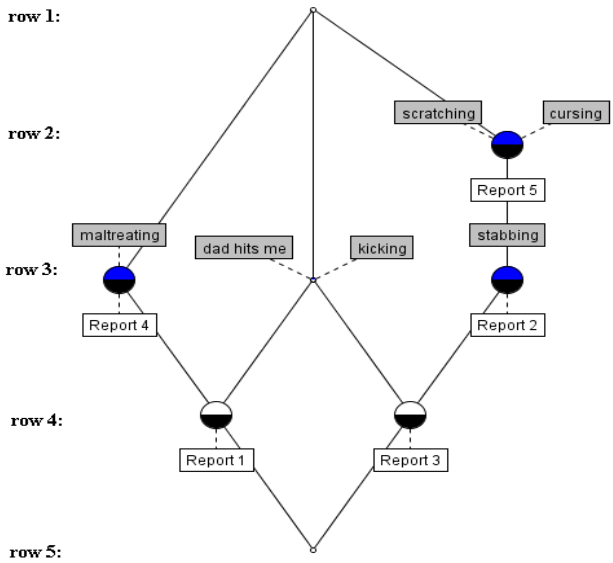


Fig. 3. 2 Line diagram corresponding to the context from Table 3.1

Retrieving the extension of a formal concept from a line diagram such as the one in Figure 3.2 implies collecting all objects on all paths leading down from the corresponding node. In this example, the objects associated with the third concept in row 3 are reports 2 and 3. To retrieve the intension of a formal concept, one traces

all paths leading up from the corresponding node in order to collect all attributes. In this example, the third concept in row 3 is defined by the attributes "stabbing," "cursing" and "scratching". The top and bottom concepts in the lattice are special: the top concept contains all objects in its extension, whereas the bottom concept contains all attributes in its intension. A concept is a subconcept of all concepts that can be reached by travelling upward and it will inherit all attributes associated with these superconcepts. Note that the extension of the concept with attributes "kicking" and "dad hits me" is empty. This does not mean that there is no report that contains these attributes. However, it does mean that there is no report containing only these two attributes.

In contrast to most data mining algorithms, the discovery process using FCA is human-centered. It is definitely not a black-box that runs and optimizes without intervention beyond specifying initial model choices and parameters.

3.3.2 Emergent Self Organizing Map

Emergent Self Organizing Maps (ESOM) (Ultsch 2005a) is a special and very recent type of topographic maps (Ritter 1999, Kohonen 1982, Hulle 2000). According to (Ultsch 2003), "*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 (Ultsch 2005b). In the traditional SOM, the number of nodes is too small to show emergence. ESOM is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of their structure (Ultsch 1990). From a practitioner's point of view, topographic maps are a particularly appealing technique for knowledge discovery in databases (Ultsch 1990, Ultsch 1999) because they perform a non-linear mapping of the high-dimensional data space to a low-dimensional space, usually a two-dimensional one, which facilitates the visualization and exploration of the data (Ultsch 2004). In the past, we applied the ESOM to a police dataset and found its performance to be superior to that of a spherical SOM tool (Poelmans 2009b). We made some interesting discoveries using the ESOM, although the obtained results were limited and not convincing enough to make it into operational policing practice (Poelmans 2009a).

It is claimed by Ultsch and co-workers that the topology preservation of the traditional SOM projection is of little use when the maps are small: the performance of a small SOM is argued to be almost identical to that of a *k*-means clustering, with *k* equal to the number of nodes in the map (Ultsch 2005a). Using large numbers of neurons, as in the ESOM, permits one to observe data at a higher level capturing the overall structures, disregarding the elementary ones and allowing the consideration of structures that otherwise would be invisible.

3.3.2.1 Emergent SOM

An ESOM map is composed of a set of neurons *I*, arranged in a hexagonal topology map or lattice. A neuron $n_j \in I$ is a tuple (w_j, p_j) in the map, consisting of a

weight vector $w_j = (w_{j1},\ldots,w_{jm})$ with $w_j \in R^m$ and a discrete position $p_j \in P$, where P is the map space. The data space D is a metric subspace of R^m. The training set $E = \{x_1,\ldots,x_k\}$ with $x_1,\ldots,x_k \in R^m$ consists of input samples presented during the ESOM training. The training algorithm used is the online training algorithm in which the best match for an input vector is searched for, and the corresponding weight vectors, and also those of its neighboring neurons of the map, are updated immediately.

When an input vector x_i is supplied to the training algorithm, the weight w_j of a neuron n_j is modified as follows:

$$\Delta w_j = \eta h(bm_i, n_j, r)(x_i - w_j)$$

with $\eta \in [0,1]$, r the neighborhood radius and h a non-vanishing neighborhood function. The best-matching neuron of an input vector $x_i \in D$

$$D \to I : bm_i = bm(x_i)$$

is the neuron $n_b \in I$ having the smallest Euclidean distance to x_i:

$$n_b = bm(x_i) \Leftrightarrow d(x_i, w_b) \le d(x_i, w_b) \forall w_b \in W .$$

Where $d(x_i, w_j)$ stands for the Euclidean distance of input vector x_i to weight vector w_j. The neighborhood of a neuron

$$N_f = N(n_f) = \{n_j \in M \mid h_{fj}(r) \neq 0\}$$

is the set of neurons surrounding neuron n_f and determined by the neighborhood set h. The neighborhood defines a subset in the map space of the neurons K, while r is called the neighborhood range.

The map produced maintains the neighborhood relationships that are present in the input space and can be used to visually detect clusters. It also provides the analyst with an idea of the complexity of the dataset, the distribution of the dataset (e.g. spherical) and the amount of overlap between the different classes of objects. The lower-dimensional data representation is also an advantage when constructing classifiers. ESOM maps can be created and used for data analysis by means of the publicly available Databionics ESOM Tool[6]. With this tool the user can construct ESOMs with either flat or unbounded (i.e. toroidal) topologies.

3.3.2.2 ESOM parameter settings

To simulate the ESOM, we used the Databionics software and its standard parameter settings (Hulle 2000). We did not attempt to optimize them. A SOM with a lattice containing 50 rows and 82 columns of neurons was used (50x82=4100 neurons in total). The weights were initialized randomly by sampling a Gaussian with the same mean and standard deviation as the corresponding features. A Gaussian bell-shaped kernel with initial radius of 24 was used as a neighborhood function. Further, an initial learning rate of 0.5 and a linear cooling strategy for the learning rate were used. The number of training epochs was set to 20. In the map displayed in Figure

[6] Databionics ESOM tool: http://sourceforge.net/projects/databionic-esom/

3.8, the best matching (nearest-neighbor) nodes are labeled in the two classes for the given test data set (red for domestic violence, green for non-domestic violence). The red squares in all figures represent neurons that mainly contain domestic violence reports, whereas the green squares represent neurons that mainly contain non-domestic violence reports. The U-Matrix (Ultsch et al. 2005) is used as background visualization in the ESOM. The local distance structure is displayed at each neuron as a height value creating a 3D landscape of the high-dimensional data space. The height is calculated as the sum of the distances to all immediate neighbors normalized by the largest occurring height. This value will be large in areas where no or few data points reside (white color) and small in areas of high densities (blue and green color).

3.3.3 C-K theory

The Concept-Knowledge theory (C-K theory) was initially proposed by Hatchuel et al. (1999), Hatchuel et al. (2002) and further developed by Hatchuel et al. (2004). C-K theory is a unified design theory that defines design reasoning dynamics as a joint expansion of the Concept (C) and Knowledge (K) spaces through a series of continuous transformations within and between the two spaces (Hatchuel 2003). C-K theory makes a formal distinction between Concepts and Knowledge: the knowledge space consists of propositions with logical status (i.e. either true or false) for a designer, whereas the concept space consists of propositions without logical status in the knowledge space. According to Hatchuel et al. (2003), concepts have the potential to be transformed into propositions of K but are not themselves elements of K. The transformations within and between the concept and knowledge spaces are realized by the application of four operators:

concept $\rightarrow$ knowledge, the conceptualization
knowledge $\rightarrow$ concept, the concept expansion
concept $\rightarrow$ concept, the concept activation and
knowledge $\rightarrow$ knowledge, the knowledge expansion.

These transformations form what Hatchuel calls the design square, which represents the fundamental structure of the design process. The last two operators remain within the concept and knowledge spaces. The first two operators cross the boundary between the Concept and Knowledge domains and reflect a change in the logical status of the propositions under consideration by the designer (from no logical status to true or false, and vice versa).

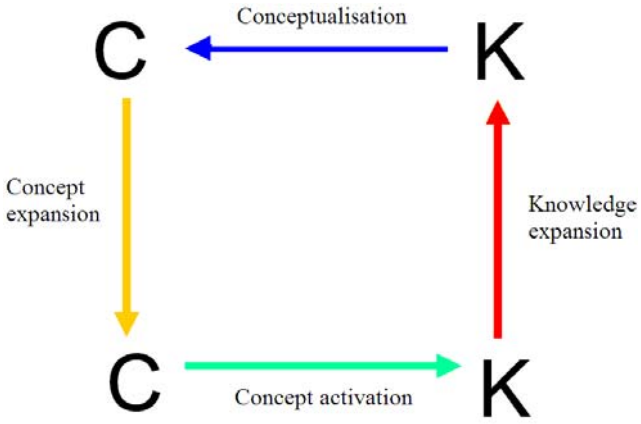


Fig. 3. 3 Design square (adapted from (Hatchuel 2003))

Design reasoning is modeled as the co-evolution of C and K. Proceeding from K to C, new concepts are formed with existing knowledge. A concept can be expanded by adding, removing or varying some attributes (a "partition" of the concept). Conversely, moving from C to K, designers create new knowledge either to validate a concept or to test a hypothesis, for instance through experimentation or by combining expertise. The iterative interaction between the two spaces is illustrated in Figure 3.3. The beauty of C-K theory is that it offers a better understanding of an expansive process. The combination of existing knowledge creates new concepts (i.e. conceptualization), but the activation and validation of these concepts may also generate new knowledge from which once again new concepts can arise.

However, one of the reasons why it is hard to apply traditional C-K theory in practice is that it lacks an actionable definition of the notions concept, partition and inclusion. In this paper, we show that these issues can be resolved by implementing the C-K framework with a synergistic combination of FCA and ESOM for modeling and expanding the space of concepts. One of the limitations of traditional C-K theory is that hierarchical representations are used to model and expand the concept space. These hierarchical representations are limited in their semantic expressiveness, which is one of the reasons why we chose for the non-hierarchical concept representation of FCA. Complementary to FCA, the ESOM functions as a catalyst to make the knowledge discovery process with FCA more efficient. One of the issues we encountered while using FCA was the scalability of the techniques for larger datasets. We choose to solve this problem by using the ESOM maps, which provide a clear picture of the overall distribution of the entire dataset and the available clusters. The combination of the maps and lattices allows for an efficient exploration of the data, leading, amongst other things, to a better selection of police reports for in-depth manual inspection.

3.4 Instantiating C-K theory with FCA and ESOM

In this section, we elaborate on the applied process for knowledge discovery based on the visually appealing discovery techniques presented in section 3.3. FCA as a

standalone technique suffers from scalability issues when the number of attributes is increased. Exploring high-dimensional data and discovering new concepts with FCA while little is known about the contents is a difficult task. Although the ESOM can provide some insights into the overall distribution of the data and may help in discovering new concepts and knowledge in the data, its capacities for knowledge discovery are limited. The ESOM as a standalone technique does not allow gaining thorough insights into the conceptual structure of the data and the underlying knowledge of police officers. This is important since we want to improve our understanding of the gaps in the current domestic violence definition, the knowledge of police officers concerning the problem, etc. In this paper, we go beyond the use of either one of these techniques and use them in combination as part of a unifying framework based on C-K theory. The unifying framework gives insight into the generic nature of the KDD activity and is a necessary precondition for successfully embedding the knowledge discovery process based on the synergistic combination of FCA and ESOM in daily policing practice. In this setup, FCA is used as a concept engine, distilling formal concepts from unstructured text. We complement knowledge discovery with the capabilities of ESOM, which functions as a catalyst for the FCA based knowledge extraction. Our approach to knowledge discovery is framed in C-K theory. The K space could be viewed as being composed of actionable information. It contains the existing knowledge used to operate and steer the action environment. The C space, on the other hand, can be considered as the design space. Whereas K is used as the basis for action and decision making, C puts this actionability under scrutiny for potential improvement and learning. At the basis of the knowledge discovery process is much fast iteration through the C-K loop.

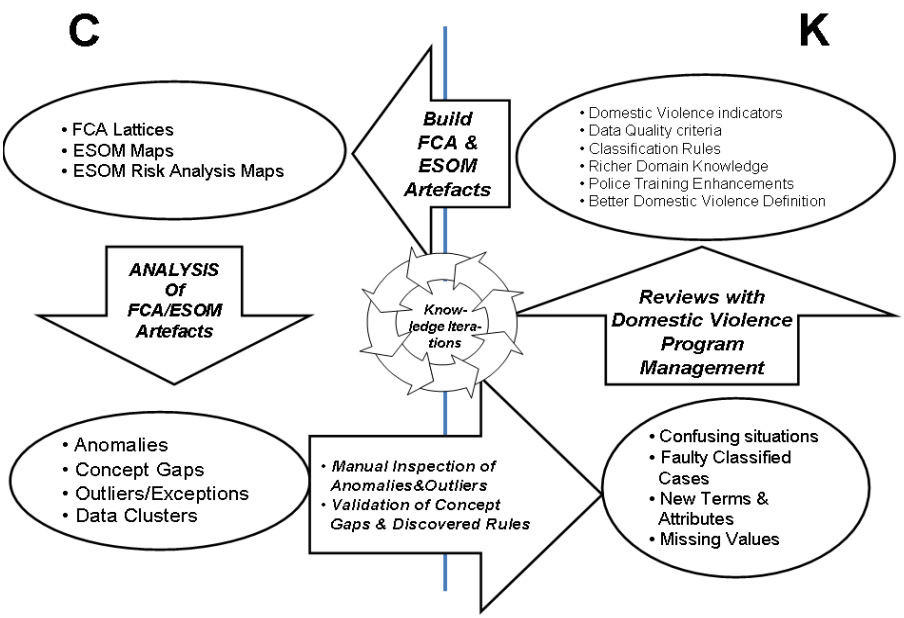


Fig. 3. 4 Knowledge discovery process

During the mining process, two persons, an exploratory data analyst and a domain expert are the driving force behind the exploration and collaborate intensively. There is a continuous process of iterating back and forth between the FCA lattices, the ESOM maps and the police reports. The knowledge discovery process using the combination of FCA and ESOM is summarized in Figure 3.4. It basically consists of iteratively applying the four operators from the design square in Figure 3.3.

Initially, an FCA lattice and an ESOM map are constructed by the exploratory data analyst based on the domain expert's prior knowledge of the problem area, the police reports contained in the dataset and the terms contained in the thesaurus (i.e. K $\rightarrow$ C). The lattice and the ESOM map provide a reduced search space to the domain expert, who then visually inspects and analyzes the lattice and ESOM map (i.e. C $\rightarrow$ C). The synergistic combination of FCA and ESOM can be considered as a knowledge browser. Our contention is that it allows for an effective interaction between the human actors and the underlying information. Using FCA, police reports are selected for in-depth manual inspection based on observed anomalies and counter-intuitive facts (i.e. C $\rightarrow$ K). Using the ESOM map, police reports are selected based on the analysis of outliers, clusters and areas of the map containing a mixture of domestic and non-domestic violence cases (i.e. C $\rightarrow$ K). These police reports are then used to discover new referential terms to improve the thesaurus, to enrich and validate prior domain knowledge, to discover new classification rules or for operational validation (i.e. K $\rightarrow$ K).

Additionally, based on the classification rules discovered using FCA, we label/relabel cases and use these cases to construct an ESOM risk analysis map. We then project the unlabeled cases onto this map (i.e. K $\rightarrow$ C). Subsequently, this map is analyzed by the exploratory data analyst and the domain expert, who search the map for outliers, clusters of cases in different areas of the map and areas containing a mixture of domestic and non-domestic violence cases (i.e. C $\rightarrow$ C). Based on the observations made, representative police reports are again selected for in-depth manual inspection (i.e. C $\rightarrow$ K). The obtained results, together with the relevant prior knowledge of the domain expert, are then incorporated into the existing visual representation, resulting in a new lattice and ESOM map (i.e. K $\rightarrow$ C).

3.5 Dataset

Our dataset consists of a selection of 4814 police reports describing a whole range of violent incidents from the year 2007. All domestic violence cases from that period are a subset of this dataset. The selection came about amongst others by filtering out those police reports that did not contain the reporting of a crime by a victim, which is necessary for establishing domestic violence. This happens, for example, when police officers are sent to investigate an incident and afterwards write a report in which they mention their findings, but the victim ends up never making an official statement to the police. The follow-up reports referring to previous cases were also removed. From the 4814 police reports contained in the dataset the following information was extracted: the person who reported the crime, the suspect, the persons involved in the crime, the witnesses, the project code and the statement made by the victim to the police. Of those 4814 reports, 1657 were classified by police officers as domestic violence. These data were used to generate the 4814 html-documents that were used during our research. An example of such a report is displayed in Figure 3.5.

The validation set for our experiment consists of a selection of 4738 cases describing a whole range of violent incidents from the year 2006 where the victim made a statement to the police. Again, the follow-up reports were removed. Of these 4738 cases 1734 were classified as domestic violence by police officers.

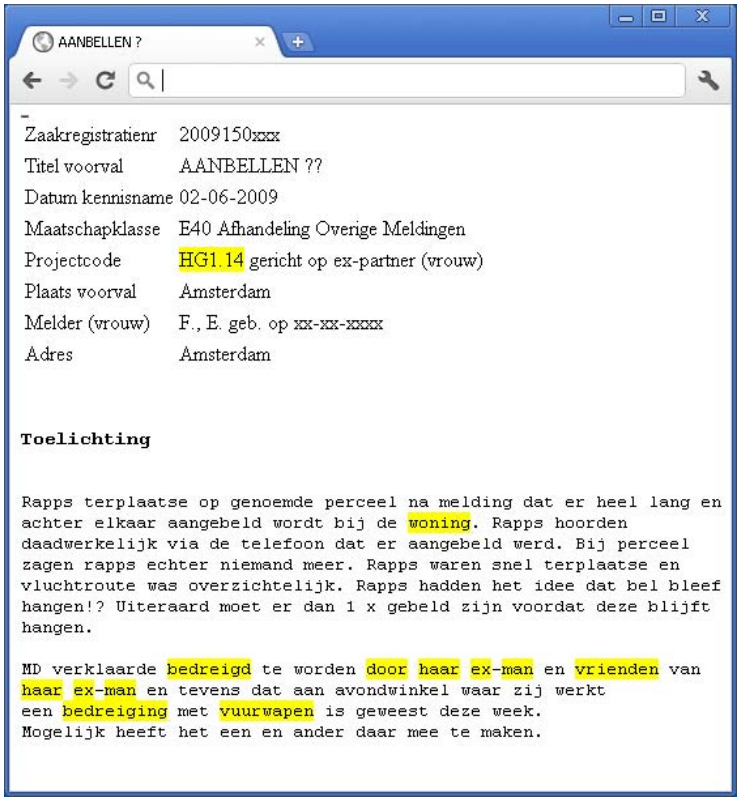


Fig. 3. 5 Example police report

The initial phase of the knowledge acquisition process consists of translating the area under investigation into objects, terms and attributes. We considered the police reports from the dataset as objects and the relevant terms contained in these reports as attributes. The terms and term clusters (see section 6) are stored in a thesaurus.

We composed an initial thesaurus of which the content was based on expert prior knowledge such as the domestic violence definition. We enriched the thesaurus with terms referring to the different components of the definition such as “hit”, “stab”, “my mother”, “my ex-boyfriend”. Since domestic violence is a phenomenon that according to the literature typically occurs inside the house, we also added terms such as “bathroom”, “living room”. We made an explicit distinction from public locations such as “under the bridge”, “on the street”. The initial thesaurus contained 123 elements.

The reports were indexed using this thesaurus. For each report the thesaurus elements that were encountered were stored in a collection. This collection would be used as input for both the FCA and the ESOM procedure. The thesaurus was refined after each iteration of re-indexing the reports and visualizing and analyzing the data with the FCA lattice and ESOM maps. This process is demonstrated in detail in section 3.6.

3.5.1 Data pre-processing and feature selection

Our initial steps consisted of data pre-processing and applying traditional classification techniques. We have applied feature selection to reduce the input space dimensionality, prior to applying the ESOM tool. We chose to select the 65 most relevant features. Feature selection comprises the identification of the most characterizing features of the observed data. Given the input data D consisting of N samples and M features $X = \{x_i, i = 1 \ldots M\}$, and the target classification variable c, the feature selection problem is to find from the M-dimensional observation space, R^M, a subspace of m features, R^m, that optimally characterizes c. A heuristic feature selection procedure, known as minimal-redundancy-maximal-relevance (mRMR), as described in (Peng 2005), was considered. In terms of mutual information, the purpose of feature selection is to find a subset S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c. This is called the Max-Dependency scheme:

$$\text{Max } D(S,c),\ D = I(x_1,\ldots,x_m;c) \quad (1)$$

As the Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features satisfying (2), which approximates $D(S,c)$ in (1) with the mean value of all mutual information values between individual feature x_i and class c:

$$\max D(S,c), D = \frac{1}{|S|}\sum_{xi \in S} I(x_i;c) \quad (2)$$

Features selected according to Max-Relevance could have redundancy, i.e., the dependency among these features could be large. When two features highly depend on each other, the respective class-discriminative power would not change much if one of them was removed. Therefore, the following minimal redundancy (Min-Redundancy) condition can be added to select mutually exclusive features (Ding 2003).

$$\min R(S), R = \frac{1}{|S|^2}\sum_{xi,xj \in S} I(x_i, x_j) \quad (3)$$

The criterion combining the above two constraints is called "minimal-redundancy-maximal-relevance (mRMR). The operator $\Phi(D,R)$ is defined to combine D and R and the following is the simplest form to optimize D and R simultaneously:

$$\max \Phi(D,R), \Phi = D - R \quad (4)$$

The outcome of this filter approach is a ranked list of features. To decide on where to cut off this list we use the classifiers discussed in the next section.

3.5.2 Initial classification performance

To obtain the optimal feature set, an SVM, a Neural Network, a kNN (k-nearest-neigbor with k=3) and a Naïve Bayes classifier were used to measure the classification performance for an increasing number of features.

Naïve Bayes is based on the Bayes rule and assumes that feature variables are independent of each other given the target class.

Given a sample $s=\{x_1,\ldots, x_m\}$ for m features, the posterior probability that s belongs to class c_i is

$$p(c_i \mid s) \propto \prod_{j=1}^{m} p(x_j \mid c_i)$$

where $p(x_j \mid c_i)$ is the conditional probability table learned from examples in the training process. Despite the conditional independence assumption, Naïve Bayes has been shown to have good classification performance for many real data sets (Cover 1991). We have used the WEKA package (Weka 2009). We used 10-fold cross-validation.

The Support Vector Machine (SVM) (Vapnik 1995) is a more modern classifier that uses kernels to construct linear classification boundaries in higher dimensional spaces. We make use of the LibSVM package (Hsu 2002). A Radial Basis Function (RBF) was chosen as kernel, the kernel parameter was set to 0.05 and 10-fold cross-validation was used.

Nearest neighbor methods estimate the probability $p(t|x)$ that an input vector $x \in R^n$ belongs to class $t \in \{0,1\}$ by the proportion of training data instances in the neighborhood of x that belong to that class. The metric used for evaluating the distance between $a,b \in R^n$ is the Euclidean distance:

$$dist(a,b) = \| a-b \|_2 = \sqrt{(a-b)^T (a-b)}$$

The version of *k*-nearest neighbor that was implemented for this study was chosen because it is especially appropriate for handling discrete data (Webb 1999). The problem with discrete data is that several training data instances may be at the same distance from a test data instance x as the kth nearest neighbor, giving rise to a non-unique set of k-nearest neighbors. The k-nearest neighbor classification rule then works as follows. Let the number of training data instances at the distance of the kth nearest neighbor be n_k, with n_{k1} data instances of class $t = 1$ and n_{k0} data instances of class $t = 0$. Let the total number of training data instances within, but excluding this distance be N_k, with N_{k1} data instances of class $t = 1$ and N_{k0} data instances of class $t=0$ if

$$N_{k1} + \frac{k - N_k}{n_k} \times n_{k1} \geq N_{k0} + \frac{k - N_k}{n_k} \times n_{k0}$$

where $N_k < k \leq N_k + n_k$. Now all training data instances at the distance of the *k*th nearest neighbor are used for classification, although on a proportional basis. The parameter *k* was set to 2 and 10-fold cross-validation was used.

We also used a feed-forward multiplayer perceptron (MLP) with one hidden layer consisting of 10 neurons and an output layer consisting of one neuron (Matlab Arsenal 2008). The weight decay parameter was set to 0.2 and the number of training cycles to 10. Again we used 10-fold cross-validation.

The classification performance is plotted as a function of the number of features in Figure 3.6. The result of the mrmr algorithm is a ranked list of the best features. The x-axis indicates how many of these best features were used to train the classifiers. The y-axis shows the classification performance for these different feature subsets. We opted to retain the best 44 features which is a compromise for the 4 classifiers. 44 features was one of the points in the curve where the sum of classification performances for the different classifiers was highest. We also tested other maxima such as 15 and 30 but this resulted in a less qualitative graphical image. A toroidal ESOM map was trained on this dataset with a reduced number of features and was compared to that of Figure 3.8. It shows that the density problem (one class label for each density peak) was not solved by lowering the number of features (result not shown).

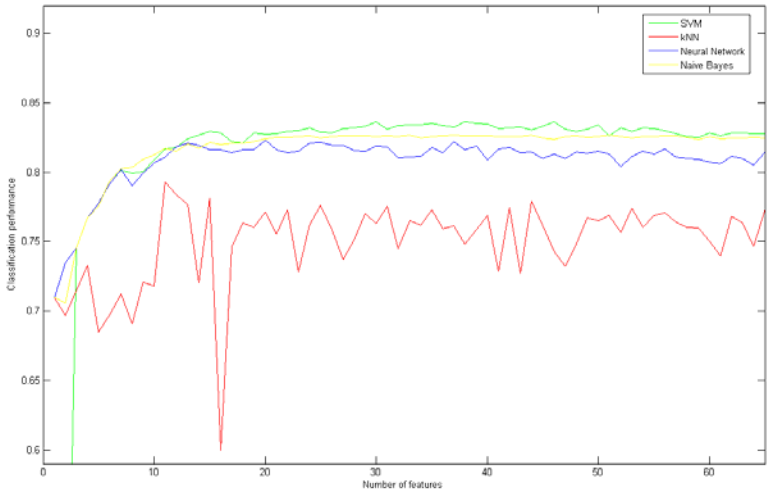


Fig. 3. 6 Classification performance for different subsets of the ranked list of features produced by the mrmr algorithm

3.6 Iterative knowledge discovery with FCA and ESOM

In this section, we illustrate the abstract description of the knowledge discovery process provided in section 3 with a real life case study with the Amsterdam-Amstelland Police Department on domestic violence. We have chosen not to present the sequential build-up of the lattices and ESOM maps, but to make a selection from these lattices and maps, just to help the reader become familiar with the explorative possibilities of the method presented here.

The process displayed in Figure 3.4 contains an iterative learning loop. During the successive iterations through the C-K loop, multiple interesting results emerged from the research. These different types of results will now briefly be described. The analysis process is showcased in detail in the next subsections. The FCA lattices and ESOM maps are mainly used as an instrument to efficiently select representative reports for in-depth manual inspection, to discover new classification rules, to enrich, test and refine expert prior knowledge, to browse and annotate the collection of police reports, etc.

An important aspect of the process consists in searching these reports for new attributes that can be used to discriminate between the domestic and non-domestic violence reports or that may lead to an enrichment of existing domain knowledge. New referential terms were not selected using a term extractor, but they were obtained by carefully reading some representative reports and then selecting relevant terms as attributes. We built in the necessary validation mechanisms to ensure the completeness of the thesaurus:

1. Word stemming. Each word is reduced to word-stem form.
2. Stop wording. A stop list is used to delete from the texts the words that are insufficiently specific to represent content. The stop list contains many common function words, such as “the”, “or”, etc.
3. Synonym lists. Synonym lists are used to add semantically similar words.
4. Spelling checking. Spelling checking is used to validate the correctness of the term added to the thesaurus and the correctness of the words in the police reports.

During the research the thesaurus was under constant evolution: when new terms and concepts were discovered, the terms were added to the thesaurus. This approach ensured that the thesaurus remained at all times a reflection of the knowledge already gained. Because of the large number of police reports in the dataset, it was not possible to visually analyze concept lattices containing more than 14 attributes. Therefore, terms with a similar semantic meaning or referring to the same domain concept were clustered by the domain experts. When these term clusters were used to create an FCA lattice, they were considered as attributes.

During the exploration, we also verified the correctness of the labels assigned by police officers to the selected cases and we searched the reports for new interesting concepts, inconsistencies, etc. This led amongst others to the discovery of faulty case labelling and situations that were often not recognized by police officers as domestic or as non-domestic violence. This information was used by the data quality management team to significantly improve the quality of the data in the police databases and to improve the way police officers handle domestic violence cases. The information was also useful for the domestic violence program manager to improve the training of police officers. We also found some regularly occurring confusing situations that could not be uniquely classified as domestic or non-domestic violence based on the domestic violence definition. These situations were presented to the program manager and were used to enrich, improve and refine the concept and definition of domestic violence.

During the discovery and conceptualization of the nature of domestic violence from the data at hand, we were able to define a set of accurate and comprehensible classification rules to automatically classify incoming cases as domestic or as non-domestic violence. In the past developing an accurate classifier using decision trees, SVM's, Neural Networks, etc. turned out to be impossible. We found that this was largely due to the incorrect labels assigned by police officers to cases, to the vagueness of the domestic violence definition and to the lack of a high-quality thesaurus. We managed to resolve many of these problems during the exploration with FCA and ESOM, resulting in a set of highly accurate and comprehensible classification rules. All these different aspects of the process, which have only been briefly introduced so far, are discussed more extensively in the next sections.

3.6.1 Transforming existing knowledge into concepts

The process of design reasoning starts by making the transition from the knowledge space to the concept space. The process of transforming propositions of K into concepts of C is called disjunction. The corresponding operator in the design square from Figure 3.3 is the knowledge $\rightarrow$ concept operator. This operator expands the space of C with elements from K. We used two techniques to perform this knowledge to concept transformation. First, we constructed an FCA lattice based on expert prior knowledge, the police reports in the dataset and the term clusters in the thesaurus. Second, we designed an ESOM map based on the terms in the thesaurus and the police reports in the dataset. Both methods are further discussed in this section.

The definition of domestic violence employed by the police organization of the Netherlands is as follows:

> *"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 (Keus 2000, Van Dijk 1997)".*

The lattice in Figure 3.7 was fundamentally influenced by this domestic violence definition. Prior to the analysis with FCA, certain terms were clustered in term clusters based on this definition and added to the thesaurus. We clustered the terms contained in the thesaurus into term clusters associated with one of the two components of the definition (i.e. prior knowledge incorporation).

We first attempted to verify whether a report could be classified as domestic violence by checking it for the occurrence of one or more terms related to each of the two components of the domestic violence definition. In other words, a case would be labeled as domestic violence if the following two conditions were fulfilled. First, a criminal offence had occurred. To verify whether a criminal offence had occurred, the report was searched for terms such as "hit", "stab" and "kick". These terms were grouped into the term cluster "acts of violence". Second, a person in the domestic circle of the victim was involved in the crime. Therefore, the report was

searched for terms such as “my dad”, “my ex-boyfriend” and “my uncle”. These terms were grouped into the term cluster “persons of domestic sphere”. It should be noted that a report is always written from the point of view of the victim and not from the point of view of the officer. A victim always adds “my”, “your”, “her” and “his” when referring to the persons involved in the crime. Therefore, the report is searched for terms such as “my dad”, “my mom” and “my son”. These terms are grouped into the term cluster “family members”. The report is also searched for terms such as “my ex-boyfriend”, “my ex-husband”, and “my ex-wife”. These terms are grouped into the term cluster “ex-partners”. Furthermore, the report is searched for terms such as “my nephew”, “her uncle”, “my aunt”, “my step-father” and “his step-daughter”. These terms are grouped under the term cluster “relatives.” Then the report is searched for terms such as “family friend” and “co-occupant”. These terms are grouped into the term cluster “family friends”. Reports that were assigned the label “domestic violence” have been classified as such by police officers. The remaining reports were categorized as non-domestic violence. This results in the lattice displayed in Figure 3.7.

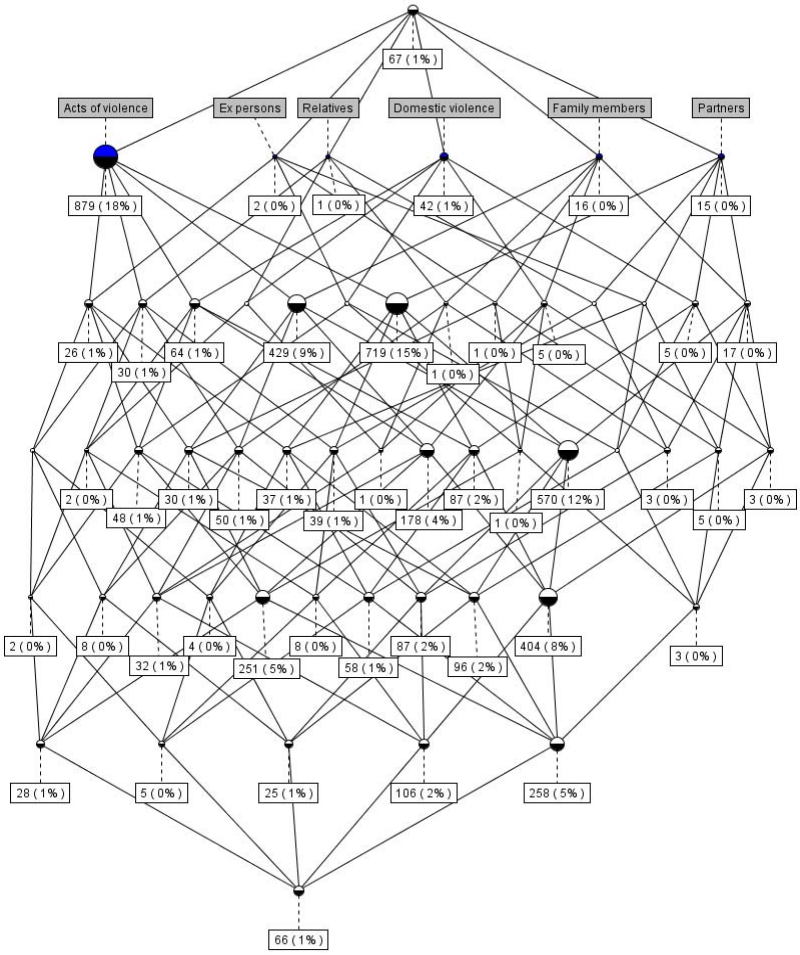


Fig. 3. 7 Initial lattice based on the police reports from 2007

Indexing the 4814 reports from 2007 with the initial thesaurus from section 5 resulted in a cross table with all reports as objects and all terms as attributes. This cross table is used for training a toroidal ESOM. The ESOM is represented in Figure 3.8: the green squares refer to neurons that dominantly contain non-domestic violence cases, while the red squares refer to neurons that dominantly contain domestic violence cases.

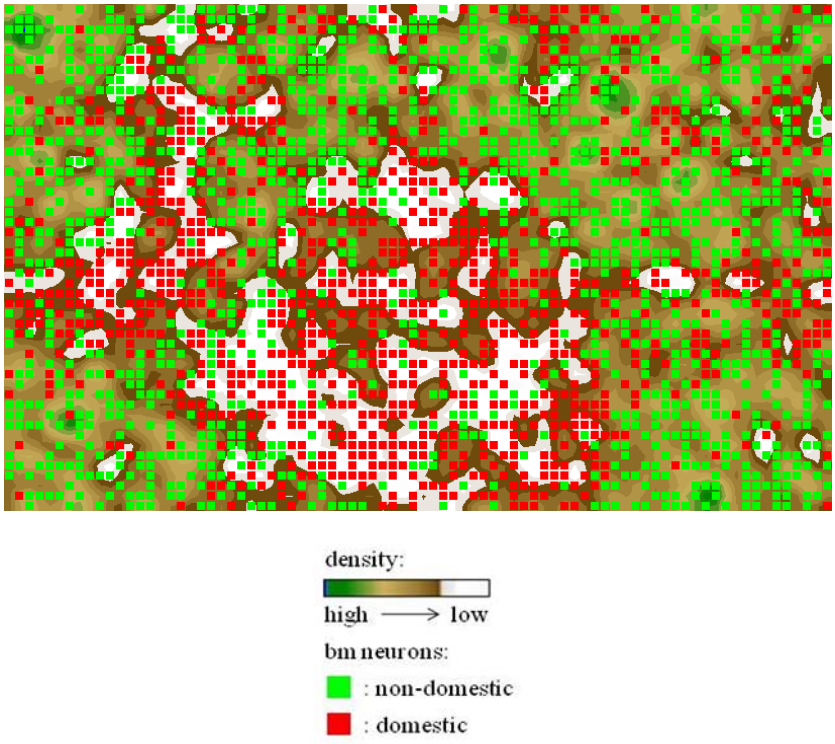


Fig. 3. 8 Toroidal ESOM map

Using the reference definition of domestic violence employed by the police was but one way to identify term clusters to structure the lattices. Term clusters also emerged from in-depth scanning of certain reports highlighted during a knowledge iteration cycle. This is how, for example, the term cluster “relational problems” was created. We discovered terms such as “relational problems”, “I had a relationship with”, which refer to a broken relationship. A distinction was made between a broken relationship and an ongoing relationship. Terms such as “I have a relationship with” and “live together” were brought together in the cluster “in a relationship”.

According to the literature, domestic violence is a phenomenon that mainly occurs inside the house (Vincent 2000, Black 1999, Beke 2003). Therefore, an attribute called “private locations” was introduced. This term cluster contained terms such as “bathroom”, “living room” and “bedroom”. An attribute called “public locations” was also introduced. The redefined lattice structure, taking into account the analyses of the previous iterations, is displayed in Figure 3.9. In order to keep the lattice comprehensible, the terms belonging to the clusters “family members”, “relatives”, “partners”, “ex-partners” and “family friends” have been lumped into a cluster “persons”.

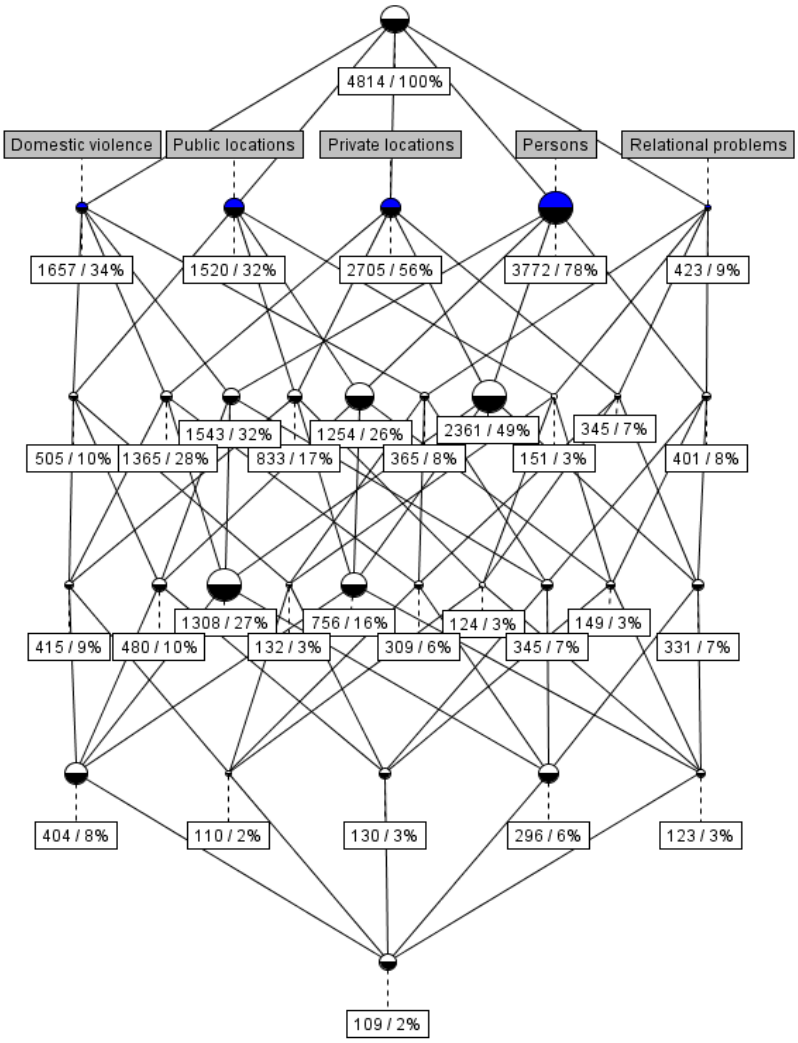


Fig. 3. 9 First refined lattice based on the police reports from 2007

In the analysis of some of the reports selected using ESOM during an earlier iteration, we also found that many cases did not have a formally labeled suspect. This attribute is also incorporated in the lattice in Figure 3.10. We also found a lot of cases with a description of the suspect. Whether or not perpetrator and victim lived at the same address at the time of the incident was also included as attribute.

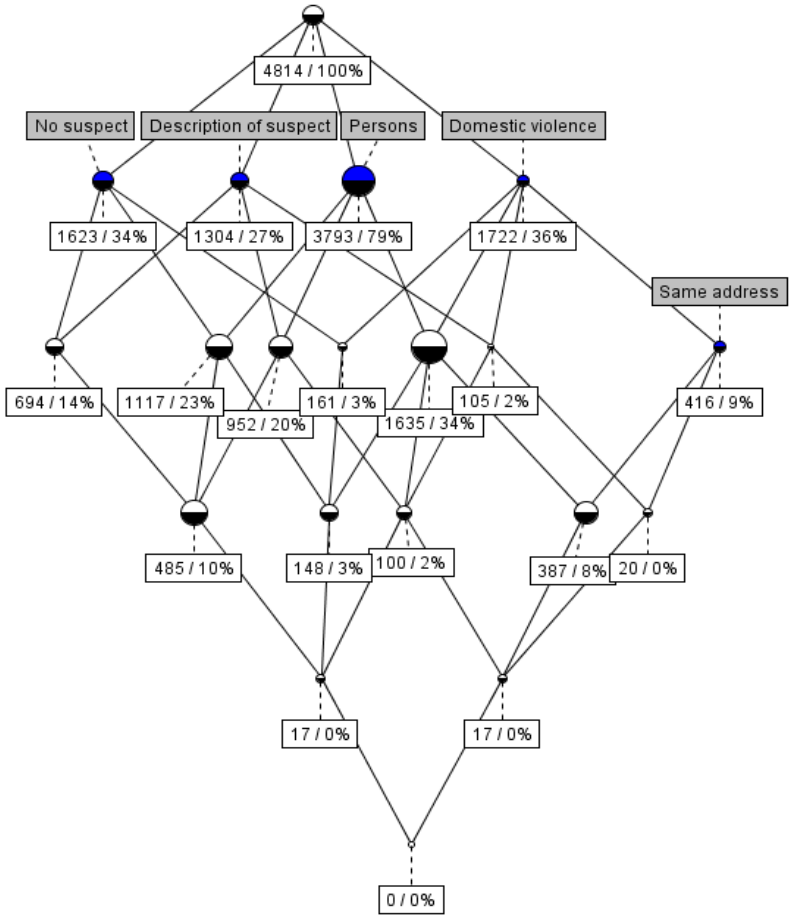


Fig. 3. 10 Second refined lattice based on the police reports from 2007

While further exploring the domestic violence reports during successive knowledge creation iterations, it became apparent that in many cases the victim made statements such as “I want to institute legal proceedings against my husband” and “I want to institute legal proceedings against my brother”. These sentences were brought together into the cluster “legal proceedings against domestic sphere”. Another type of phrasing that was regularly used by victims of domestic violence was, for example, “the crime was committed by my dad” or “the crime was committed by my ex-boyfriend”. These sentences were brought together into the cluster “committed by domestic sphere”. Yet another type of wording that was also frequently used by a victim was phrases such as “I was maltreated by my husband” and “I was threatened by my ex-partner”. These sentences in turn were brought together into the cluster “threatened by domestic sphere”. Finally, neighborhood quarrels (non-domestic violence) often made reference to phrases such as “I want to

institute legal proceedings against my neighbor" and "committed by the man next door", so these sentences were combined into the cluster "neighbors". These attributes were included in the lattice of Fig. 3.11.

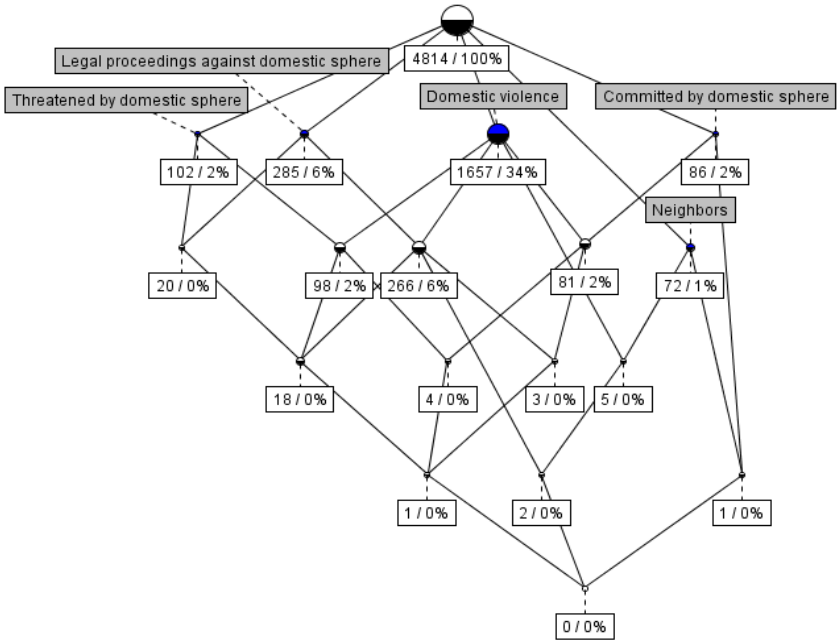


Fig. 3. 11 Third refined lattice based on the police reports from 2007

We also use FCA for the validation of some aspects of operational policing practice. For some specific situations it was verified whether police officers disposed of sufficient knowledge about the problem area to recognize these cases as domestic violence. Some very important special domestic violence situations were considered, including incest and honor-related violence. For the first type of situation, reports were searched for terms such as "incest" and "sexual abuse by my father". For the second type of situation, reports were searched for terms such as "marriage of convenience" and "marry off". The resulting lattice after incorporating these special cases is displayed in Figure 3.12.

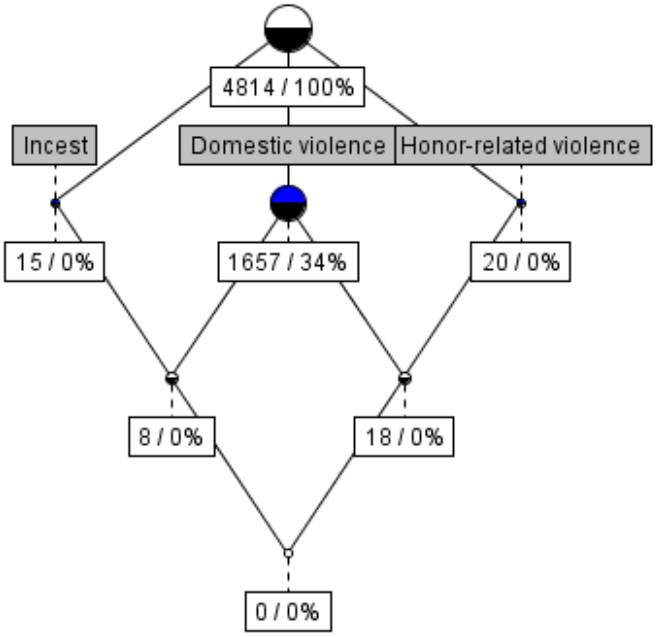


Fig. 3. 12 Fourth refined lattice based on the police reports from 2007

3.6.2 Expanding the space of concepts

The notion of expansion plays a key role in C-K theory. An analyst's ability to recognize an expansion can depend on his sensitivity to these opportunities, his training or the knowledge at his disposal. In (Hatchuel 2004) it is stated that expansion is a K-relative notion, which means that its significance depends on the knowledge of a designer or any other observer or user. In this paper, we argue that FCA and ESOM help analysts recognize and exploit these opportunities. Basically, C space expansion is driven by the analyst's detection and investigation of anomalies, outliers, clusters and concept gaps with these visual exploration tools. Based on these observations, police reports are selected for in-depth manual inspection. This section describes in more detail these two ways of expanding the space of concepts.

We first explain how we used FCA to expand the space of concepts. FCA was used to efficiently explore the data based on the prior knowledge of the domain expert. Some interesting findings emerged from the interactive exploration of the lattice in Figures 3.7 – 3.12 and warranted further investigation.

Table. 3.2. Interesting observations from the lattices in Figures 3.7 – 3.12

	Non-domestic violence	**Domestic violence**
No “acts of violence”	128	60
No “acts of violence” and “persons of domestic sphere”	63	18
“Acts of violence” and no “persons of domestic sphere”	863	72
“Relational problems”	58	365
“private locations”	1340	1365
“public locations”	1015	505
Acts of violence and same address	37	379
Acts of violence and no suspect and description of suspect	695	16
Acts of violence and no suspect	1442	181
“legal proceedings against domestic sphere”	19	266
“committed by domestic sphere”	5	81
“threatened by domestic sphere”	4	98
“neighbors"	67	5
“incest”	7	8
“honor-related violence”	2	18

As can be seen from Table 3.2, a total of 60 domestic violence cases did not contain a term from the “acts of violence” term cluster. Of these 60 cases 18 contained a term from the clusters containing terms referring to a person in the domestic sphere of the victim. Interestingly, some 28% (i.e. 863) of the non-domestic violence reports only contain terms from the “acts of violence” cluster, while there are only 72 domestic violence reports in the dataset that share that characteristic. Apparently, some cases that were labeled as domestic violence did not fit the definition of domestic violence that was used to start this discovery exercise in the first place. The reports in question were therefore selected for in-depth investigation.

It should be clear from the lattice in Figure 3.9 that the terms contained in the cluster “relational problems” tend to be associated with domestic violence cases. Apparently, only 58 non-domestic violence reports contained one or more terms from the “relational problems” cluster. We concluded that the presence of at least one of the terms from of this cluster in a police report seemed to be a strong indication for domestic violence. This was enough evidence to warrant manual inspection of these 58 police reports.

We also used FCA to verify the correctness and the practical usefulness of this prior knowledge. Most of the domestic violence cases under scrutiny (1365 cases or 82%) contained one or more terms from the “private locations” term cluster. However, 1340 (42%) of the non-domestic violence cases also contained one or

more terms from this same term cluster. In addition, a hypothesis that was formulated prior to the data exploration was that almost no domestic violence case was expected to have taken place on the street. Surprisingly, this hypothesis was proven incorrect by the data. In about one-fourth of the domestic violence cases there had been an incident at a public location. While scrutinizing these police reports, we discovered that this was often the case when ex-partners were involved. It became apparent that it was not possible to distinguish domestic from non-domestic violence reports by means of the type of locations mentioned in the reports. Combining the clusters "private locations" and "public locations" with clusters such as "family members" or "ex-persons", for example, did not yield the expected results in terms of discriminatory power. We noticed that in a large number of the domestic violence cases (416 cases or 28%) the perpetrator and the victim happened to live at the same address at the time the victim made their statement to the police. Most of these cases (379 cases or 91%) were classified as domestic violence.

Visual inspection of the patterns produced by the ESOM map in Figure 3.8 also allowed us to make some interesting observations. For example, color coding made it easy to detect outlying observations: some red squares are located in the middle of a large group of green squares and vice versa. For further examination we made use of the ESOM tool's functionality to select neurons and display the cases that had this neuron as their best match. We thought that these neurons were associated with cases that might have been wrongly classified by police officers. Therefore, these cases were also selected for in-depth manual inspection.

3.6.3 Transforming concepts into knowledge

The concept $\rightarrow$ knowledge operator from Figure 3.4 transforms concepts in C into logical questions in K. In our case an answer to such a question is found by manually inspecting the selected police reports. We refer to this manual analysis as the validation of concept gaps, giving rise to multiple types of discoveries: confusing situations, new referential terms, faulty case labelling, niche cases and data quality problems.

For example, and with reference to Table 3.2, the 18 cases labeled by police officers as domestic violence that contained a term from the "persons of domestic sphere" but no violence term were selected for manual inspection. Is it possible that there are domestic violence reports in which the victim does mention a person of the domestic sphere, but does not mention an act of violence? In-depth analysis showed that these 18 reports contained violence related terms that were originally lacking from the initial thesaurus, such as "abduction", "strangle" and "deprivation of liberty". Another example is the discovery of 42 cases that did not contain a violence term or a term referring to a person of the domestic sphere. These cases turned out to be wrongly classified as domestic violence. We also analyzed the reports that contained a violence term but no term referring to a person of the domestic sphere. This inspection revealed that more than two thirds of these reports were wrongly classified as domestic violence. In the next section, we will focus on the causes of these labelling errors and the extraction of actionable intelligence from

these individual cases that can be used to improve the domestic violence definition and the training of police officers.

Table 3.2 also indicates that there were 58 police reports that were classified as non-domestic violence while containing a term from the "relational problems" cluster. This investigation revealed that a startling 95% of these cases had been wrongly labeled as non-domestic violence. Moreover, about 70% of these cases had as a common feature that a third person made a statement to the police for someone else. Analysis of the remaining 30% of these misclassified cases led to the discovery of an important new concept that was lacking from the domain expert's initial definition of domestic violence. Many of the reports included expressions such as "I was attacked by the ex-boyfriend of my girlfriend" and "I was harassed by the ex-girlfriend of my boyfriend". These terms were grouped into the cluster "attack by ex-person against new friend". This situation is analyzed in detail in the next section together with the resulting actionable intelligence. The term cluster is also used to distil new classification rules in one of the subsequent iterations.

Another interesting finding emerged from our search for novel and potentially interesting classification attributes. The lattice in Figure 3.10 shows that some 34% of the reports (1623 cases) did not mention a suspect. According to the domestic violence definition (which specifies that the perpetrator must belong to the domestic circle of the victim), the offender has to be known in domestic violence cases. Naturally, we had assumed that these reports described non-domestic violence cases. Nevertheless, when looking into these cases, we found that 181 of them turned out to describe domestic violence cases after all. In the next section, we uncover the causes of this phenomenon. Additionally, we found out that some 44% of the reports (711 cases) that lacked a labeled suspect did contain a description of the actual suspect. Of these 711 cases, only 16 reports were classified as domestic violence. After studying these 16 reports, we discovered that the majority of them were wrongly classified as domestic violence.

When studying the remaining 37 non-domestic violence cases more carefully, we found, much to our surprise, that the perpetrator and the victim often lived together in the same institution (e.g. a youth institution, a prison or a retirement home). It turned out that of the 41 cases where the perpetrator and the victim lived in the same institution only 30 actually had been classified as cases of domestic violence. The non-domestic violence cases where the perpetrator and the victim lived at the same address and were not inhabitants of an institution turned out to be wrongly classified as non-domestic violence. Therefore, a new attribute called "institution" was introduced. After browsing the 19 non-domestic violence cases in which the victim used one or more terms from the "legal proceedings against domestic sphere" cluster, it turned out that these reports should have been classified as domestic violence. The same observation was made when the 5 non-domestic violence reports containing a term from the "committed by domestic sphere" cluster and the 4 non-domestic violence cases containing a term from the "threatened by domestic sphere" cluster were analyzed. In-depth investigation of the 5 domestic violence cases in which a term from the "neighbors" cluster occurred, showed that these reports should have been classified as non-domestic violence.

After an in-depth manual inspection of the police reports corresponding to the ESOM outliers, interesting discoveries were made. For example, we observed that many of these outlier reports contained several important new features that were lacking in the domain expert's understanding of the problem area. Every time new and important features were discovered in this way, they were used to enrich the thesaurus. A selection of these features is displayed in Table 3.3 and 3.4.

Table. 3.3. Newly discovered features by studying the domestic violence outliers in the ESOM map.

Pepper spray
Homosexual relationship, lesbian relationship
Sexual abuse, incest
Alternative spelling of some words (e.g. ex-boyfriend, exboyfriend, ex boyfriend)
Weapons lacking in the thesaurus: belt, kitchen knife, baseball bat, etc
Terms referring to persons: partner, fiancée, mistress, concubine, man next door, etc.
Terms referring to relationships: romance, love affair, marriage problems, divorce proceedings, etc
Reception centers: woman's refuge center, home for battered woman, etc. Gender of the perpetrator: mostly male
Gender of the victim: mostly female
Age of the perpetrator: mostly older than 18 years and younger than 45 years
Age of the victim: mostly older than 18 years and younger than 45 years
Terms referring to an extra marital affair: I have an another man, lover, I am unfaithful, etc

Table. 3.4. Newly discovered features by studying the non-domestic violence outliers in the ESOM map.

Places of entertainment: club, disco, bar, etc.
Crime locations: on the street, on a bridge, under a viaduct, on a crossing, etc.
Public locations: metro station, bus stop, tram stop, etc.
Reception centers: refugee center, shelter for the homeless, relief center, etc.
Drugs: drug abuse, drug joint, etc.
Addresses of youth institutions, prisons, etc.
Hotel: hotel room, hotel, etc.
Description of suspect's origin: Turkish descent, white man, North-African descent, etc.
Description of suspect's body: 1.75 meters tall, 119 centimeters tall, muscular appearance, etc.
Description of suspect's hair: curly haired, blond hair, redhead, etc.
Description of suspect's clothes: black jacket, leather shoes, blue pants, jeans, etc.
Description of suspect's face: beard, moustache, facial hair, etc.
Description of suspect's accent
Unknown person is involved in the crime
Attack by unknown person
Corporate body
Neighborhood quarrel

The reports also contained multiple confusing situations. When more detailed information was disclosed to us, these cases were also used to refine the domestic violence definition.

3.6.4 Expanding the space of knowledge

The expansion of the space K constitutes validation or testing of the proposed expansion with the ultimate goal of producing actionable intelligence. K-validation of a concept boils down to a confrontation of the output from the C-K transformation with knowledge sources available to the K space (e.g. cross-checking with other databases, setting up field experiments, soliciting expert advice). These new propositions have logical status. In this section, we show how we obtain actionable intelligence from the observations made during the Concept $\rightarrow$ Knowledge phase.

Analysis of the misclassified police reports described in the previous section revealed that for some unknown reason police officers regularly seem to misclassify burglary, car theft, bicycle theft and street robbery cases as domestic violence. Therefore, terms such as "street robbery", burglary" and "car theft" were combined into a new term cluster called "burglary cases". This term cluster was then used in one of the subsequent iterations through the C-K loop.

In the previous section, we also described how the analysis of the police reports revealed that a situation in which a third person makes a statement for somebody

else can be confusing for police officers. For example, one case described a father who made a statement to the police about the sexual abuse of his daughter by her stepfather. This is a clear case of domestic violence, but since it was not the victim who made the statement to the police, the police officer did not recognize it as such. This type of situation is now specifically addressed in police training.

In the previous section, we also described how the analysis of police reports revealed interesting cases in which the ex-boyfriend attacked the new boyfriend. We presented these ambiguous cases to the board members responsible for the domestic violence policy. Police officers and policy makers confirmed that this type of situation was to be seen as domestic violence, mainly because the perpetrator often intends to emotionally hurt the ex-partner. Consequently, the expectation was for the terms contained in this cluster to frequently occur in domestic violence reports. However, this turned out to be incorrect. It became clear from the investigation that in general this type of situation was very confusing to police officers. A quick scan revealed that more than 50% of police officers actually had trouble with such cases. The ensuing investigation and discussions with police officers and policy makers revealed that this situation needed to be addressed during the training of police officers. Several interesting cases like the previous one were identified during the data exploration. All of them resulted in a clearer insight into the nature of domestic violence.

In the previous section, we found that some domestic violence cases did not mention a formally labeled suspect. Analysis revealed that this was a result of police officers' rather haphazard ways of registering suspects for these cases. Apparently, while some officers immediately registered a suspect at the moment the victim mentioned this person as a suspect, others preferred to first interrogate these suspects before officially labeling them as such. In the latter case, the person would just be added to the list of persons who were said to be involved in or to have witnessed the crime. Because such lists included friends, family members or bystanders, they could potentially be very extensive and diverse. Which is why suspects easily got lost in these lists. When we inquired about the proper policy regarding the labelling of suspects, we were told there simply was none. Our analysis made a strong case for the need for such a policy. In the end, the quick-win proposal that could be implemented to solve this issue involved a relatively simple change to the registration software: an additional data entry field would need to be introduced for police officers to register the persons that were mentioned by the victim as offenders.

The same address finding brought about a lively discussion amongst the police officers of the Amsterdam-Amstelland Police Department. More importantly, it exposed the discord amongst police officers on how to classify such cases. We took note of all their reflections and presented them to the board members responsible for the domestic violence policy. After intensive debate the classification guidelines, displayed in Table 3.5, were obtained. Careful inquiry into the incest and honor-related violence cases taught us that police officers regularly misclassified incest cases as non-domestic violence. On the other hand, even for insiders it was quite surprising to observe how almost all honor-related violent incidents ended up being correctly classified as domestic violence. The latter was probably attributable to the

intensive sensitization campaigns organized to inform police officers of this important societal problem.

The newly obtained knowledge led to a new iteration of the FCA analysis, supported by another run of the ESOM tool. In each iteration, it is possible that one or more new classification rules are discovered. The attribute “corporate body”, for example, was found by first analyzing a cluster of green squares that was located within a group of red squares in an ESOM map. With FCA we found that the presence of a corporate body in a police report almost always excludes domestic violence. Therefore, we introduced a new domestic violence classification rule named “corporate body”. An other example of a classification rule is when a case has no formally labeled suspect and contains a description of a suspect, it can be labeled as non-domestic violence.

3.7 Actionable results

Several iterations through the design square resulted in truly valuable upgrades of the K space from the perspective of improving action in the field. This section provides an overview of some of the most important achievements of our work.

First, we were able to refine the definition of domestic violence that would act as a principle guideline for labeling cases. During the exploration, several types of niche cases were identified as valid exceptions to the general definition. No clear labeling guidelines were available, so we formulated advice, grounded in evidence, to redesign the general policy. Eventually, we obtained the classification guidelines displayed in Table 3.5.

Table. 3.5. Classification guidelines for incidents involving inhabitants of the same institution

Perpetrator	**Victim**	**Classification**
Caretaker	Inhabitant	Domestic violence
Inhabitant	Caretaker	Non-domestic violence
Inhabitant younger than 18y	Inhabitant younger than 18y	Domestic violence
Inhabitant older than 18y	Inhabitant older than 18y	Non-domestic violence
Inhabitant of prison older than 18y	Inhabitant of prison older than 18y	Individual evaluation
Inhabitant older than 18y	Inhabitant younger than 18y	Domestic violence
Inhabitant younger than 18y	Inhabitant older than 18y	Individual evaluation

In the end the presence or absence of a dependency relationship between the perpetrator and the victim was the decisive factor for classifying a case as either domestic or as non-domestic violence. Nevertheless, we also discovered some regularly occurring situations in which there is a clear dependency relationship

between the perpetrator and the victim, but that were typically classified as non-domestic violence by police officers. A selection of these circumstances is listed in Table 3.6. These confusing situations helped to expose the mismatch between the management's conception of domestic violence and that of police officers. We found that the management employed a much broader definition of domestic violence than most police officers.

Table. 3.6. Circumstances in which the offender abuses the dependency relationship with the victim, but that are not recognized by police officers as domestic violence.

Circumstance	**Dependency relationship**
Lover boys	The victim is in love with the lover boy, who abuses this dependency relationship to force her into prostitution.
Extramarital relationship	If the mistress of an adulterer blackmails him, for example by threatening to reveal their affair to his wife, the mistress abuses the dependency relationship that exists between her and the man.
Violence between a caretaker and an inhabitant of an institution	If the caretaker threatens or harasses the inhabitant (for example, a nurse who maltreats an elderly woman in a retirement home), the latter is often helpless because she depends on the caretaker.
Violence between colleagues	If two colleagues had a relationship and one keeps stalking the other, this is domestic violence between ex-persons.
An ex-boyfriend attacks the new boyfriend	This is considered to be domestic violence because the ex-boyfriend often intends to emotionally hurt his ex-girlfriend.
Third person makes statement to the police for somebody else	Police officers regularly fail to recognize cases in which a third person makes a statement to the police for somebody else (e.g. a father who makes a statement about the sexual abuse of his daughter by her stepfather) as domestic violence.

Second, a set of 22 domestic violence and 15 non-domestic violence classification rules were extracted. Using these rules, 75% of cases from the year 2007 could be labeled automatically as either domestic or non-domestic violence. We also applied these rules to two validation sets containing unstructured police reports from the year 2006 and from the year 2008, which yielded similar results, i.e. 72% and 73% respectively. These rules are now fully operational and used to automatically and correctly classify the majority of incoming cases, while in the past all cases had to be dealt with manually. Ten of these domestic violence and five of these non-domestic violence classification rules are displayed in Table 3.7

Table. 3.7. Excerpt of discovered classification rules

	Domestic violence classification rules
1	Legal proceedings against domestic sphere
2	Committed by domestic sphere
3	Relational problems and living together
4	Relational problems and institutions
5	Honor related violence
6	Incest
7	(Court) injunction
8	Fear of domestic sphere
9	Attack by ex-person against new friend
10	Problems with domestic sphere
	Non-domestic violence classification rules
1	Unknown perpetrator
2	Corporate body
3	Burglary cases
4	Road rage
5	Violence at school

Third, the set of newly identified classification rules did not just allow the police to classify incoming cases. The rules could also be employed to reclassify cases from the past to result in more correct performance management and reporting over time. Domestic violence cases that were not recognized as such in the past might also be re-opened for investigation. In total, we found 420 filed reports that were wrongly labeled as domestic violence and 912 filed reports that were wrongly labeled as non-domestic violence. Table 3.8 presents an overview of these results.

Table. 3.8. Number of filed reports that were incorrectly classified, but corrected by means of the 37 rules

	Non-domestic corrected to Domestic	Domestic corrected to Non-domestic	Total
Year 2006	307	136	443
Year 2007	290	115	405
Year 2008	315	169	484
Total	912	420	1332

Finally, based on the cases that could be labeled using the classification rules that were discovered, we constructed an ESOM risk analysis map. For each neuron, the number of domestic and non-domestic violence cases contained in the neuron and the 32 surrounding neurons was counted and used to calculate the probability that a police report that has this neuron as its best match described a domestic violence incident. For the visualization, a color scheme consisting of 5 different colors was used. Red indicates a 90-100% probability rate of domestic violence,

orange a 70-90% probability rate, yellow a 30-70% probability rate, green a 10-30% probability rate and dark green a 0-10% probability rate. The labels of the cases that could not be categorized using the new classification rules were not used to construct this risk analysis map. However, we projected these remaining cases onto this map afterwards. The map for the dataset of the year 2007 is shown in Figure 3.13. Cases that were labeled by police officers as domestic violence are represented as black dots, while the cases that were labeled as non-domestic violence, are represented as light blue dots.

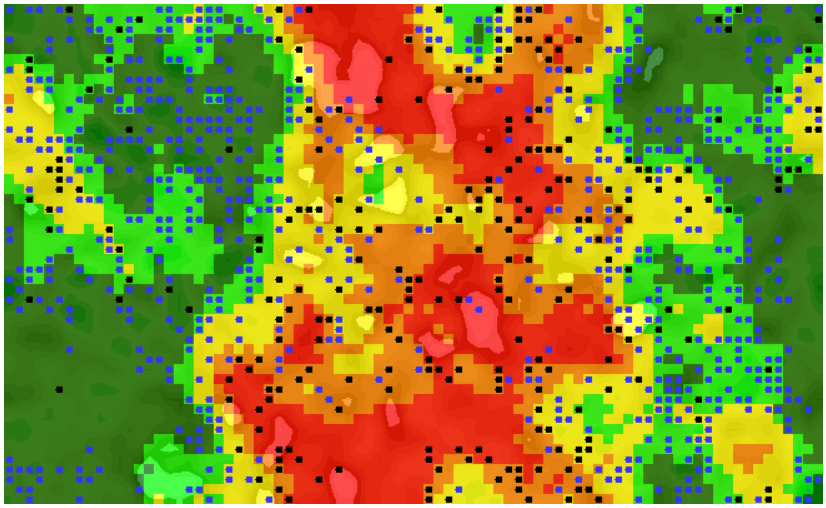


Fig. 3. 13 ESOM risk analysis map for the year 2007 and remaining cases made visible

It was remarkable to observe that some of the remaining cases were located in the red area of the map, but were not classified by police officers as domestic violence. About 6.4% of the remaining cases were located in the red area of the map displayed in Figure 3.13. About 22.1% of the cases located in the red area of the map were classified as non-domestic violence by police officers. In-depth analysis of these police reports revealed that the majority of these cases should have been classified as domestic violence. On the other hand, only a small percentage of the cases located in the dark green and green areas of the map were classified as domestic violence by police officers (4.8% and 12.4% respectively). Further scrutiny revealed that all of these cases actually described non-domestic violence incidents.

Table. 3.9. Distribution of remaining cases of 2007 over different map areas

Domestic violence probability	Map area color	% of remaining cases located in map area	% classified as domestic violence	% classified as non-domestic violence
0-10%	dark green	28.1%	4.8%	95.2%
10-30%	green	30.0%	12.4%	88.6%
30-70%	yellow	21.5%	37.7%	62.3%
70-90%	orange	14.0%	64.3%	35.7%
90-100%	red	6.4%	77.9%	22.1%

Based on the map displayed in Figure 3.13, a correct label can be automatically assigned to 64.5% of the remaining cases of the year 2007 (i.e. the cases located in the dark green, green and red areas of the map). The other cases (i.e. the cases located in the yellow area of the map) have to be classified manually. A similar result was obtained for the cases of the year 2006 and 2008. Based on the comprehensible classification rules discovered during the knowledge discovery exercise, we developed a Tomcat-based system to assist analysts in their labelling of cases. The system is currently used as a stand-alone application by the data quality management team (i.e. the back office). The long term goal is to make it available to all police officers in the organization (i.e. the front office) to assist them in their labelling of cases.

The labelling process, as performed by the data quality management team, consists of a number of steps that are, to a large extent, automated by the newly introduced system. First, the user can select a set of police reports for labelling (e.g. all police reports from the month October 2008). Subsequently, the classification rules that were discovered during the exploration of the data are applied to the cases. When a case comes in for labelling, the first step consists in verifying whether one of the domestic violence rules is satisfied. If this is the case, the case is classified as domestic violence. If not, it is verified whether one of the non-domestic violence rules is applicable. If this is the case, the case is classified as non-domestic violence. Otherwise, the case is left unclassified. The remaining cases are projected onto the ESOM risk analysis map based on the cases labeled with the FCA rules. Using the combination of the classification rules and the ESOM risk analysis map, 91.0 % of cases can be classified automatically and correctly. This is a major improvement compared to the past situation where each incoming case had to be dealt with manually.

In 2010 a new version of the in-triage system, Trueblue, has been released. Trueblue is used to detect faulty cases in the BVH system and is based on knowledge rules. The majority of the knowledge rules are related to the structured information from the BVH system. Only a few rules, like “label domestic violence is missing”, “weapons involved” and “discrimination against police officers” uses a combination of structured and unstructured information.

One of the new functionalities of the new version of Trueblue is showing the detected suspicious cases when clicking on it in a web browser, which reduces the

inspection time of the individual cases from 5 minutes to 2 minutes. In the previous version they have to activate two sessions in two separate windows, a Trueblue session and a BVH session, and copy and paste the detected case number from the Trueblue session to the BVH session.

Unfortunately, this new version turns to be unstable during the year 2010. A number of knowledge rules which uses unstructured text were deactivated because of memory problems. This means that the knowledge rule of domestic violence is not applied over 2010 and can still not be applied at the moment of writing this thesis. In appendix D we have simulated the knowledge rule of domestic violence with the Cordiet toolbox (see chapter 5) and detected 5,367 suspicious domestic violence cases over 2010. The quality team has no time to inspect all 5,367 cases manually within a short period and has asked us to apply the found domestic violence rules from section 3.7 over the dataset of 2010.

The Trueblue knowledge rule does not include two important classes of general reports and restricts the dataset to 37,294 cases. We decided to add two classes of general reports to our dataset because the BVH system has 9302 labeled domestic violence cases in 2010 where the dataset of Trueblue has only 6092 labeled domestic violence cases (see appendix D). The difference (35%) is mainly attributable to the two classes of general reports.

The cases are divided into three groups based on the class of the report. First the violence and vandalism reports, where serious offences were committed. Second the quarrel and stalking reports, where no serious offences are committed, but may contain signals of domestic violence and third the general reports.

We developed a rule based application which used the rules detected from FCA, from which table 3.7 shows an excerpt. Appendix E shows the description of the rule based application we applied for domestic violence. Each rule associates with a classification probability and the corresponding color. Green corresponds with 95%, orange corresponds with 85% and yellow with 75%. The results are stored in a new created table within the Trueblue database.

Our dataset of 2010 consists of 90.385 reports: 15,433 violence and vandalism, 13,688 quarrel and stalking and 61,254 general reports. If a case meets at least one of the rules, a record is added to a table of Trueblue with the information about the case itself, the detected rule and all detected concepts from the thesaurus. The application is extended with two functionalities. The first is generating a HTML file for the quality team to show and inspect the detected cases and their probability. The case can be selected and inspected in detail showing the report with highlighted terms. Appendix E shows examples of the HTML file and an example of detected domestic violence with highlighted terms. The second functionality produces an input file for generating a FCA lattice to investigate the results of the classification. The result of this lattice is show in Figure 3.14.

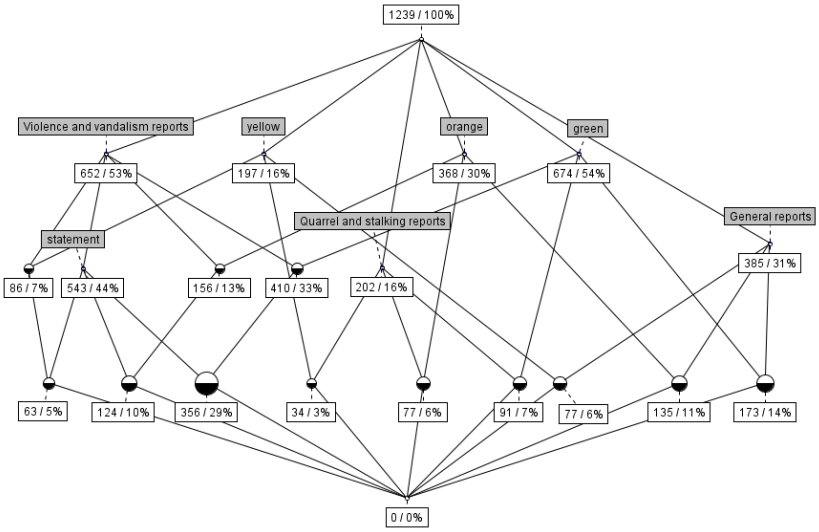


Fig. 3. 14 detected domestic violence cases of 2010

Out of a total of 90,385 cases we detected 1,239 suspicious cases of domestic violence. The results can be retrieved from the lattice. Starting from the node "green" we find 674 (54%) cases. Going down to the three connecting nodes of the reports and go up to the corresponding node of the report group, we find 410 violence and vandalism reports, 91 quarrel and stalking reports and 173 general reports. In the same way we can retrieve the values for orange and yellow. The overall result is presented in table 3.10

Table. 3.10. Results of the classification of the BVH cases.

	Violence reports	**Quarrel reports**	**General reports**	**Total**
Green	410 (63%)	91 (45%)	173 (45%)	674 (54%)
Orange	156 (24%)	77 (38%)	135 (35%)	368 (30%)
Yellow	86 (13%)	34 (17%)	77 (20%)	197 (16%)
Detected	652(100%)	202(100%)	385(100%)	1,239(100%)

The table shows that 674 cases out of 1,239 (54%) can be classified as domestic violence with a probability of 95%, 368 cases with 85% probability and 197 with 75% probability. The table also shows the importance of reading the general reports where 385 cases are found (31% of all cases), which would never have been detected by Trueblue.

Our approach does have two significant better results applying Trueblue given the scenario from Figure 3.1 with inspecting all detected cases. First the reduction in inspecting the number of detected cases from 5,367 to 1,239. Depending on the experience of the member of the quality team, inspecting a domestic violence case takes 2 to 5 minutes. Not only must the case itself be investigated, but also the

history of the involved persons. An experienced member will save (5,367 – 1,239) * 2)/60 = 138 hours on a total of 179 hours. Second, more suspicious cases will be detected, because the two classes of general reports are responsible for 385 suspicious cases which Trueblue would never have been found.

3.8 Comparative study of ESOM and multi-dimensional scaling

In this section we compare the usability of ESOM and MDS as text exploration instruments in police investigations. Multi-dimensional scaling (MDS) is a method that uses the similarity or dissimilarity among pairs of objects in the original space to represent the objects in a lower dimensional space for visualization purposes. In our case, we have used the classical metric MDS algorithm (Gower 1966) to visualize, in a two-dimensional space, the distribution of police reports in the sense that two reports will be close to each other when their correlation is high (Borg et al. 2005). Both the MDS and ESOM can be used for detecting closely related data points, but each one has its own focus. Contrary to the ESOM, which starts directly from the document vectors, we first have to construct a dissimilarity matrix prior to the MDS calculation. In our case, it is a (symmetric) 4814 x 4814 matrix containing the Euclidean distances between each pair of normalized document vectors. The MDS algorithm (Kruskal et al. 1978), starts from this calculated distance matrix and uses a function minimization algorithm to find the best configuration in a lower dimension, i.e. a mapping of the original space on a two-dimensional space, thereby minimizing the overall error. The error is defined as the sum of the squared differences between the distances in the original space (as present in the Euclidean distance matrix) and the corresponding ones in the lower dimensional space. We used the cmdscale algorithm from the R package for calculating the MDS map (Gower 1966).

The output of an ESOM calculation is different from that of a metric MDS. The metric MDS algorithm concentrates on the largest dissimilarities whereas ESOM concentrates on the largest similarities. ESOM tries to reproduce the topology of the data in a 2D grid, instead of reproducing distances. Similar documents are represented by neighboring neurons in an ESOM, while a distance in an MDS map can be interpreted as an estimate of the true distance between both (Wehrens et al. 2007). The MDS map trained on the same initial dataset is displayed in Figure 3.15. The red dots indicate police reports labeled as domestic violence, whereas green dots indicate police reports labeled as non-domestic violence.

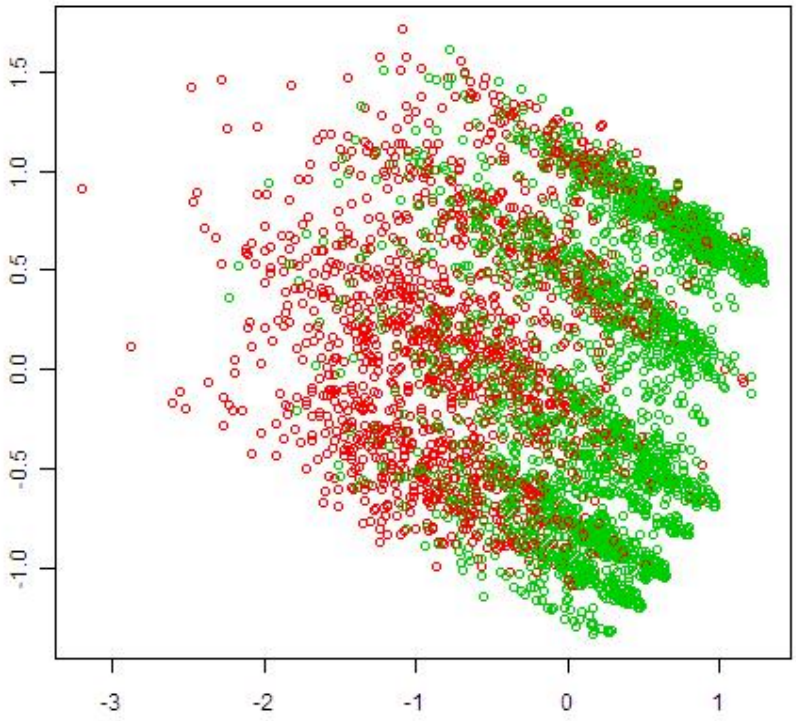


Fig. 3. 15 MDS map trained on the categorical dataset

After multiple successive iterations of refining the thesaurus, training a new map, and analyzing the resulting ESOM, our thesaurus contained more than 800 domain-specific terms, term combinations and term clusters. We found that the classification accuracy of the SVM, Neural network, Naïve Bayes and kNN classifiers improved significantly after adding the newly discovered features to the thesaurus. For example, for the SVM, the best classification accuracy on the initial dataset was around 83%, while the best classification accuracy on the dataset with the refined thesaurus was around 89%. These classification accuracies are again plotted in function of the best selected features in Figure 3.16.

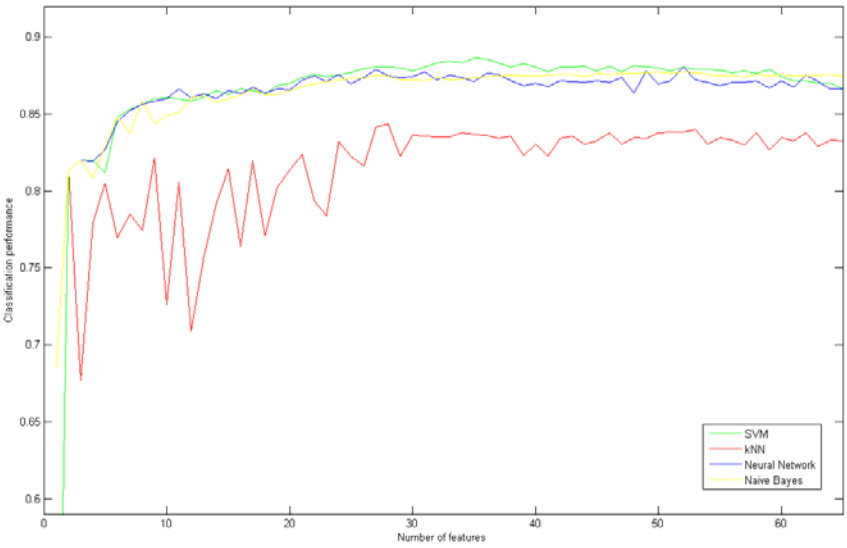


Fig. 3. 16 Classification performance

During one of the final iterations and before correcting the wrongly labeled cases, we trained a new toroidal ESOM map and MDS map on the dataset based on the refined thesaurus. The resulting map is displayed in Figure 3.17. The resulting MDS map is displayed in Figure 3.18.

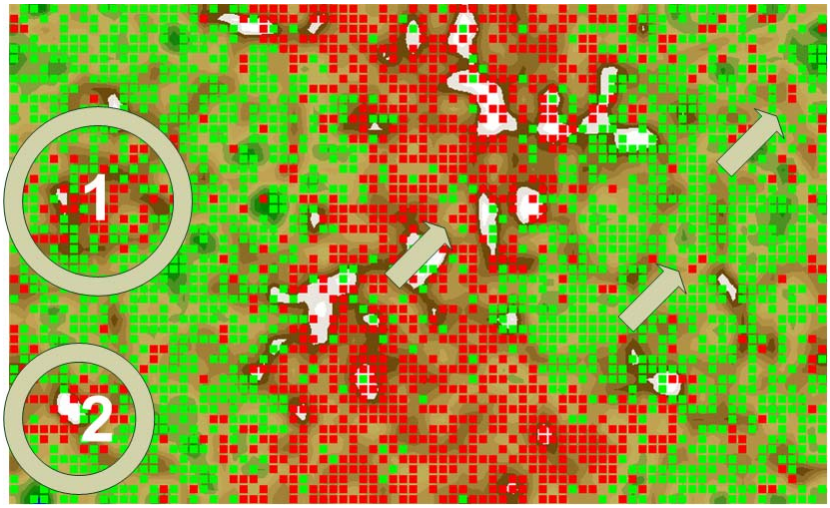


Fig. 3. 17 Toroid ESOM map trained on the categorical dataset. See text for an explanation of the arrows and circles.

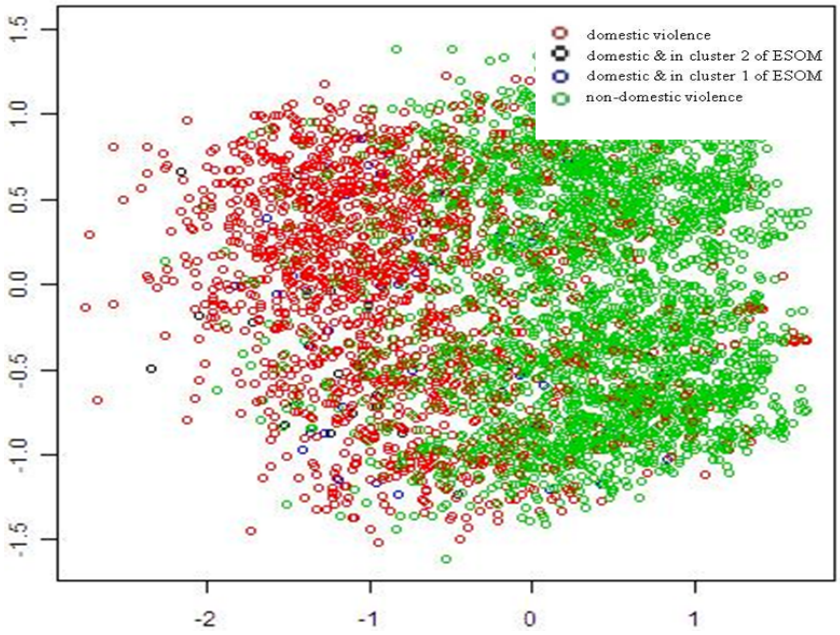


Fig. 3. 18 MDS map trained on the categorical dataset.

Comparing the ESOM map of Figure 3.8 to that of Figure 3.17 reveals that the amount of overlap between the two classes has decreased significantly after the refined thesaurus was introduced. The map in Figure 3.17 shows 3 different clusters that mainly contain cases labeled as domestic violence. When we inspected the cases contained in the top left cluster (circle marked as 1), we found that this cluster mainly contained the burglary cases that for some unknown reason were wrongly labeled by police officers. During analysis of the cluster located at the left and bottom of the map (circle marked as 2) and some of the outliers (arrows), we discovered a large number of situations that were found to be confusing for police officers. Their opinions differed on how these cases should be labeled. No such clusters were found in the MDS map. Finally, we found that the outliers mostly contained cases that were wrongly labeled as either domestic or non-domestic violence. We conclude that ESOM is better suited in our case for knowledge discovery purposes.

3.9 Conclusions

In this chapter, we proposed an approach to knowledge discovery from unstructured text using FCA and ESOM. The approach was framed within C-K theory (i.e. the design square) to provide a deeper understanding of the nature of the exploration process, a process that is essentially human-centered. In this chapter we argued for the discovery capabilities of FCA and ESOM, acting as information browsers in the hands of human analysts. The tools were shown to help analysts proceed with knowledge expansion by progressively looping through the design square in an effective way. We demonstrated the method using a real-life case study with data from the Amsterdam-Amstelland Police Department. The case focused on the problem of distilling concepts indicating domestic violence from the unstructured text in police reports. The data exploration for this case study resulted in several improvements to the way domestic violence cases are dealt with and reported on in practice. This included the implementation of an effective early case filter to identify cases that truly warrant in-depth manual inspection.

Intensive audits of the police databases revealed that many police reports tended to be wrongly classified as domestic or as non-domestic violence. Our approach was used to discover new features that better distinguish domestic from non-domestic violence cases resulting in higher classification accuracy and an improvement of the domestic violence definition. Additionally, we found some regularly occurring situations that were often wrongly labeled as non-domestic violence by police officers (e.g. lover boys). Eventually, we managed to build an accurate and comprehensible classifier that automatically assigns a correct label to more than 90% of incoming cases.

We applied the detected rules on a dataset of BVH with 90,385 cases from the year 2010 and found 1,289 suspicious domestic violence cases. The quality team will save at least 138 out of 179 hours time compared to inspecting the 5,369 cases detected by the in-triage system and also detect 385 more suspicious cases because the general reports were included in the dataset.

We have also performed a comparative study of ESOM and MDS for analyzing large amounts of unstructured text. The ESOM was able to recognize two extra data clusters that were of significant importance but not found by MDS.

Potentially, in future work one could investigate how iceberg lattices and alpha lattices could be used to prune the FCA lattices. One could also investigate the potential of conceptual scaling for improving scalability of the lattices which is however very labor-intensive.

CHAPTER 4

Formal concept analysis of temporal data.

In this chapter we investigate the power of the combination of FCA and TCA from real life cases. FCA is used to detect potential suspects and TCA is used to profile the potential suspects. The first case study uses a newly developed behavioral model of classifying (potential) jihadists in four sequential phases of radicalism[7]. FCA and TCA are for the first time used to actively find new subjects. The second case study uses FCA and TCA to identify and profile Human Trafficking and Loverboy suspects[8]. Both cases have in common that they rely heavily on reported observations of suspicious situations made by police officers on the street.

4.1 Terrorist threat assessment with Temporal Concept Analysis

The National Police Service Agency of the Netherlands developed a model to classify (potential) jihadists in four sequential phases of radicalism. The goal of the model is to signal the potential jihadist as early as possible to prevent him or her to enter the next phase. This model has up till now, never been used to actively find new subjects. In this section, we use Formal Concept Analysis to extract and visualize potential jihadists in the different phases of radicalism from a large set of reports describing police observations. We employ Temporal Concept Analysis to visualize how a possible jihadist radicalizes over time. The combination of these instruments allows for easy decision-making on where and when to act.

4.1.1 Introduction

In the modern day globalized world, the ease of terrorist network information exchange is characterized by contact moments through the internet and an absence of time and location restrictions. The amount of information available to police forces is continuously increasing and many police forces are not ready for handling data amounts of this size. As a consequence, pro-actively observing potential threats to our national security becomes increasingly difficult. The National Police Service Agency (KLPD) of the Netherlands started a new Intelligence Led Policing (ILP) project with the aim of collecting terrorist-related information in visually appealing

[7] Part of this section has been published in 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.

[8] Part of this section is submitted and accepted for the 19th International Conference on Conceptual Structures, Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011) A concept discovery approach for fighting human trafficking and forced prostitution.

models (Knowledge in Models, KiM, project) to ease the extraction and sharing of actionable knowledge. The KiM project is part of the Program Improvement by Information Security Awareness (VIA). This program is a partnership between the National Coordinator of Counterterrorism (NCTb), the National Forensic Institute (NFI), the General Intelligence and Security Service (AIVD) and the KLPD. Shortly described, the program includes research on and implementation of methods and techniques for supporting police services in their fight against terrorism.

One of the results of this project was the development of a model describing the radicalization process a potential jihadist may pass through before committing attacks. This model consists of 4 phases and its feasibility and practical usefulness have been validated by members of the intelligence services on known suspects. After the validation of this model on known jihadists, the next and probably most important, step consists of finding unknown potential suspects by applying the model to the large amounts of structured and unstructured text available in the police databases.

In this paper, we make use of the techniques known as Formal Concept Analysis (FCA) (Ganter et al. 1999, Priss 2005) and Temporal Concept Analysis (TCA) (Wolff 2005). Contextual attribute logic (Ganter et al. 1999a) is used to group and transform the terrorism indicators available into new attributes for generating the concept lattices. After extracting the potential suspects for each phase of the model, a detailed profile based on TCA is constructed giving the history of the suspect and his current level of threat to national security.

The remainder of this section is composed as follows. In section 4.2.2, we give some background on Jihadism in the Netherlands and the four phase model of radicalism used by the KLPD. In section 4.2.3, we elaborate on the dataset used. In section 4.2.4, the essentials of FCA and TCA theories are introduced. In section 4.2.5, the research methodology is explained and the results of the analysis are presented. Finally, section 4.2.6 concludes the section.

4.1.2 Backgrounder

4.1.2.1 Home-grown terrorism

In November 2004 the Dutch society was confronted for the first time with an act of terrorism, namely the brute murder of the Dutch film maker Theo van Gogh. The people suddenly realized that the ideology of violent jihad against the West had also established a foothold in the Netherlands and that the Netherlands as well had become a scene of terrorist violence. It ensued that the murderer, and most other members of the extremist network to which he belonged, were young Muslims born and bred in the Netherlands (AIVD 2006, AIVD 2007).

The latter fact has been seen as a confirmation of the new phase in Islamist terrorism, the phase in which the threat emanates principally from extremist European Muslims who are prepared to commit attacks in their own country, also known as the *European jihad*. The AIVD formulated four general trends in the development of jihadism (AIVD 2006). The first and most important is the evolvement from exogenous foreign terrorist threat to indigenous *home-grown* terrorism. This threat has led to the project VIA, Information Security Awareness

coordinated by the National Coordinator of Counterterrorism (NCTb). One of the results of this project is the development of a four phase model of Muslim radicalization by the National Police Service Agency. This model will be discussed in detail in the next section.

4.1.2.2 The four phase model of radicalism

The four phase model of radicalism, displayed in Figure 4.1, developed by the National Police Service Agency, is based on the idea a jihadist might pass through several phases before he or she might commit serious acts of terrorism. Several indicators (i.e. words and/or sentences) are associated with each phase which are used to decide based on automated text analysis, to which phase a subject belongs. Due to National Security reasons the indicators can not be published. Interested and authorized intelligence services can contact the National Police Service Agency[9]. An exception is made for the indicator "change of behavior" from type 2-A. Some of the keywords belonging to this indicator are the phrases "not shaking hands with women", "wearing traditional clothes", "suddenly let grow a beard" and "Islamistic marriage".

The model should be interpreted in a bottom up fashion. If 4 or more indicators of type 1 become true or 2 or more of type 2-A, than the subject enters the preliminary phase. But if the number of type 2-A comes below 2 or the number of type 1 comes below 4, then the subject will leave the preliminary phase. If 5 or more indicators of type 1 becomes true and 6 of type 2-B then the subject will enter the Social alienation phase, etc.

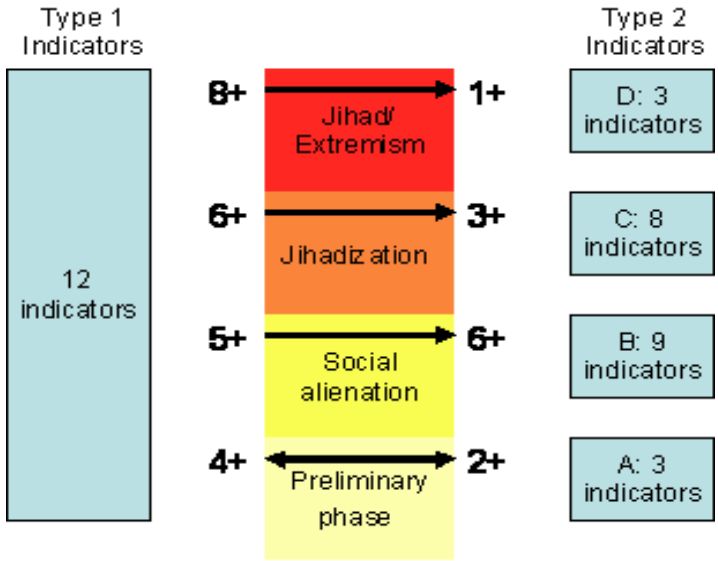


Fig. 4. 1 The four phase model of radicalism

[9] http://www.politie.nl/klpd

In the preliminary phase the subject experiences a crisis of confidence; confidence in the authorities is undermined. At this point it is not so much a matter of an ideological rift, but certainly of distrust. Many young Muslims, especially those who have grown up in the West, turn to Islam in search of their identity and place in Western society. Often their parents still live in accordance with the strict traditional norms and values of Islam, which the young people can no longer relate to. They seek a new place for themselves within Dutch society, where their ethnic, religious and national identity can be a balanced part of the whole.

In the social alienation phase a small minority of these young Muslims cannot handle this situation. On the one hand, the first generation of migrants looks down on them for becoming too 'Dutchified'; on the other hand, they do not really fit in with their Dutch peers because they are viewed in terms of their origins. This is generally where a shift occurs from the desire for a national identity to the desire for a religious identity. Strict Islam, as a guiding principle in their lives, provides certainty and stability because it tells them precisely what to do and what not to do. This increases their susceptibility for the ideology of strict, extremist religion, and makes them feel alienated from the rest of society. This alienation finds expression in increasing rejection of Dutch society.

In the Jihadization phase the subjects are characterized by strong radical Islamic convictions and the fact that they condone violence. Strong involvement in a radical group may ultimately lead to a willingness to support terrorists in the Netherlands or elsewhere in the world. This may include all sorts of support (e.g. funding). This phase entails further alienation from society and an even greater readiness to make an active contribution to the Jihad. The subjects' firm belief in the rightness of their radical ideology and of radical Islam may lead to recruitment activities to convince others of radical Islamic beliefs and possibly also of the necessity of the Jihad. Isolation from the rest of the world is part of a gradual process.

The last phase, Jihad/Extremism, is a phase of total isolation. The subjects' entire lives are governed by their radical Islamic beliefs. This is the last step before carrying out Islamist terrorist acts. In this phase, subjects are prepared to use violence themselves to achieve their objectives. In most cases, the definitive preparation for perpetrating an attack takes the form of physical training, often at a training camp abroad. The final step is actually carrying out violent activities.

4.1.2.3 Current situation

All police forces in the Netherlands (25 police regions and the National Police Service Agency) have a monitoring task of collecting information about potential jihadists. Due to the nature of their activities, potential jihadists will avoid contact with the police and other legal authorities as much as possible. The consequence is that finding new potential jihadists is like finding a needle in the haystack. Attempts were made to search the national police database BlueView containing over 50 million documents. Unfortunately this turned out to be a laborious task.

The four phase model is not used yet as an instrument for finding new potential jihadists from large datasets, but as a checklist. To apply the model on large datasets, the KLPD has started a partnership with the Amsterdam-Amstelland Police Department who is investigating intelligent text mining applications, like the

classification system for domestic violence (Elzinga et al. 2009). This application has been used to evaluate the first version of the four phase model on the police dataset of the region of Amsterdam-Amstelland. The results of this investigation have led to fine tuning the conditions imposed by the model to maximize the recall and to find as many potential jihadists as soon as possible.

4.1.3 Dataset

Our first dataset consists of 166,577 general police reports from the years 2006 (41990), 2007 (54799) and 2008 (69788) from the region Amsterdam-Amstelland, which holds the communities Amsterdam, Amstelveen, Uithoorn, OuderAmstel and Diemen. These general reports contain observations made by police officers during motor vehicle inspections, during a police patrol, when a known subject was seen at a certain place, etc. This dataset is extended by activity and incident reports which are labeled with the projectcode “TERR” (terrorism) or “EXPL” (weapons and explosives). Next to general reports there are incident reports like car accidents, burglary, violence cases, etc. There are two reasons why we have chosen for analyzing the general reports. Since the implementation of an Intelligence Led Policing program at the Amsterdam-Amstelland Police Department, the number of general reports has been growing rapidly over the years. The unstructured text describing the observations made by police officers has a lot of underexploited potential for finding potential extremists or radicalizing subjects. The challenge is to find new potential jihadists within the huge amount of general reports.

An example of a report is displayed in Figure 4.2 where two police officers asked two repeat offenders information about a third subject, called C. The reason of the inquiries is that the officers might have an indication that C. might be a recruiter.

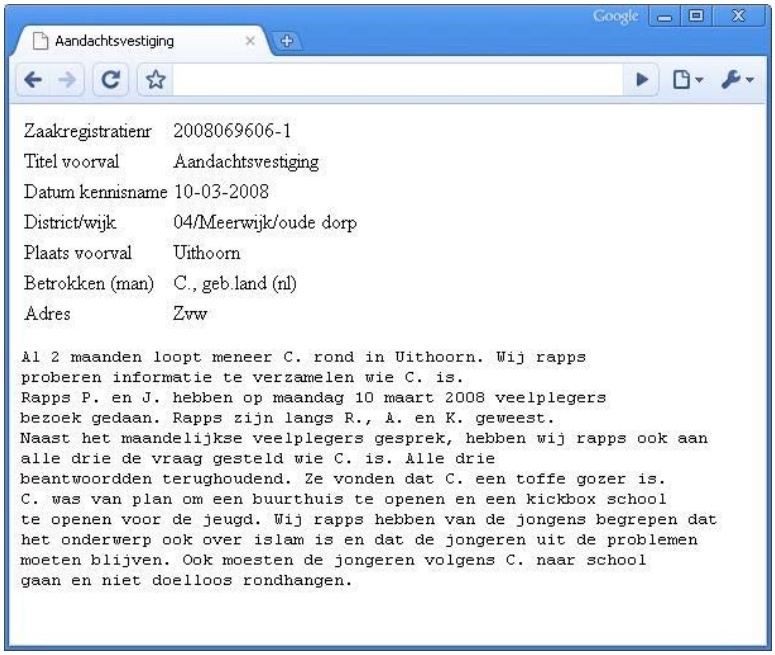


Fig. 4. 2 Example police report

4.1.4 Temporal Concept Analysis

Temporal Concept Analysis (TCA) is a mathematical theory that was introduced in scientific literature about nine years ago. TCA is based on Formal Concept Analysis (FCA) and addresses the problem of conceptually representing time. TCA is particularly suited for the visual representation of discrete temporal phenomena. In the following sections, we first introduce the essentials of FCA theory. Then, we discuss the extension of FCA with a time dimension, i.e. TCA

4.1.4.1 FCA essentials

FCA concept lattices are used to describe the conceptual structures inherent in data tables without loss of information by means of line diagrams yielding valuable visualizations of real data (Stumme et al. 1998). In a previous paper, we analyzed the concept of domestic violence using FCA (Poelmans 2009). The main difference with domestic violence is that there is a time dimension involved in human trafficking. Suspects are often spotted several times by the police and it is important to incorporate this time dimension in the visualization of the data. FCA can be used as an unsupervised clustering technique (Wille 2002, Stumme 2002) and police reports containing terms from the same term clusters are grouped in concepts.

The starting point of the analysis is a database table consisting of rows M (i.e. objects), columns F (i.e. attributes) and crosses $T \subseteq M \times F$ (i.e. relationships between objects and attributes). The mathematical structure used to represent such a cross table is called a formal context (T, M, F). An example of a cross table is

displayed in Table 4.1. In this table, collected reports of police observations of a person (i.e. the objects) are related (i.e. the crosses) to a number of terms (i.e. the attributes); here a report is related to a term if the report contains this term. The dataset in Table 4.1 is an excerpt of the one we used in our research. Given a formal context, FCA then derives all concepts from this context and orders them according to a subconcept-superconcept relation, which results in a line diagram (a.k.a. lattice). Full details on FCA can be found in chapter 2.

Table 4.1 Example of a formal context

Person	Anti-western	Orthodox-religion	Change behavior
A	X	X	X
B		X	X
C	X		X
D		X	X
E	X		

The set of all concepts of a formal context combined with the subconcept-superconcept relation defined for these concepts gives rise to the mathematical structure of a complete lattice, called the concept lattice of the context, which is made accessible to human reasoning by using the representation of a (labeled) line diagram. The circles or nodes in this line diagram represent the formal concepts. The shaded boxes (upward) linked to a node represent the attributes used to name the concept. The non-shaded boxes (downward) linked to a node represent the objects used to name the concept. The information contained in the formal context can be distilled from the line diagram in by applying the following reading rule: an object "*g*" is described by an attribute "*m*" if and only if there is an ascending path from the node named by "*g*" to the node named by "*m*".

Retrieving the extension of a formal concept from a line diagram such as the one in Figure 4.2 implies collecting all objects on all paths leading down from the corresponding node. To retrieve the intension of a formal concept, one traces all paths leading up from the corresponding node in order to collect all attributes. The top and bottom concepts in the lattice are special: the top concept contains all objects in its extension, whereas the bottom concept contains all attributes in its intension. A concept is a subconcept of all concepts that can be reached by travelling upward. This concept will inherit all attributes associated with these superconcepts.

4.1.4.2 TCA essentials

The pivotal notion of TCA theory (Wolff 2002, Wolff et al. 2003) is that of a conceptual time system (Wolff 2005). An example of a data table of a conceptual time system is displayed in Table 4.2.

Table 4.2 Data table of a conceptual time system

Time-part		Event part		
Time granule	Date	Anti-western	Orthodox religion	Change behavior
0	2008-01-26			X
1	2008-02-24	X	X	
2	2008-03-28		X	X
3	2008-04-06	X		X
4	2008-05-01		X	
5	2008-06-14		X	
6	2008-07-25			X
7	2008-08-14	X		

Table 4.2 contains the observations of one real person at several points of time. To make a single observation, police officers needed some time, varying from a few minutes to a few hours. We abstract from the duration of an observation and use the notion of a point of time, also called time granule. We thus start from a set of which the elements are time granules. In table 4.2 for example, we have 8 time granules. For describing the observations, we use a single valued context with G as its set of formal objects. This context consists of an event part and a time part. The indicators observed at each of these time granules are described in the event part of the data table. In contrast to (Wolff 2005), where a multi-valued context was used, we only need a single-valued context here. Formally, the conceptual time system we use can be described as follows.

Let $T := (G, M, I_T)$ and $C := (G, E, I_C)$ be two single valued contexts respectively on the same object set G. Then the pair (T, C) is called a conceptual time system on the set G of time granules. T is called the time part and C the event part or space part of (T, C). The combination of T and C is denoted by $K_{TC} := T|C$. It is the context of the conceptual time system (T, C). The object concepts of K_{TC} are called situations, the object concepts of C are called states and the object concepts of T are called time states. The sets of situations, states and time states are called the situation space, the state space and the time state space of (T, C) respectively. In the visualization of the data, we want to express the "natural temporal ordering" of the observations. In the TCA lattice, a time relation R is introduced on the set G of time granules of a conceptual time system. We speak of a conceptual time system with a time relation (CTST).

Let (T, C) be a conceptual time system on G and $R \subseteq G \times G$. Then the triple (T, C, R) is called a conceptual time system (on G) with a time relation.
On the set G :={0,1,2,3,4,5} of time granules we introduce the relation

R :={(0,1), (1,2), (2,3), (3,4), (4,5)} shortly described as
$0 \to 1 \to 2 \to 3 \to 4 \to 5$.

We also need the notions of transitions and life tracks. The basic idea of transition is a “step from one point to another”. The transitions in Figure 4.5 in section 4.1.5.3 form an example of a suspect who radicalizes over a given period.

4.1.5 Research method

The method we propose is summarized in Figure 4.3. First, we extract all subjects who have at least one attribute from the large set of observations with FCA. Second, we construct lattices for each phase of jihadism. Third, we use TCA to profile the selected subjects and their evolution over time.

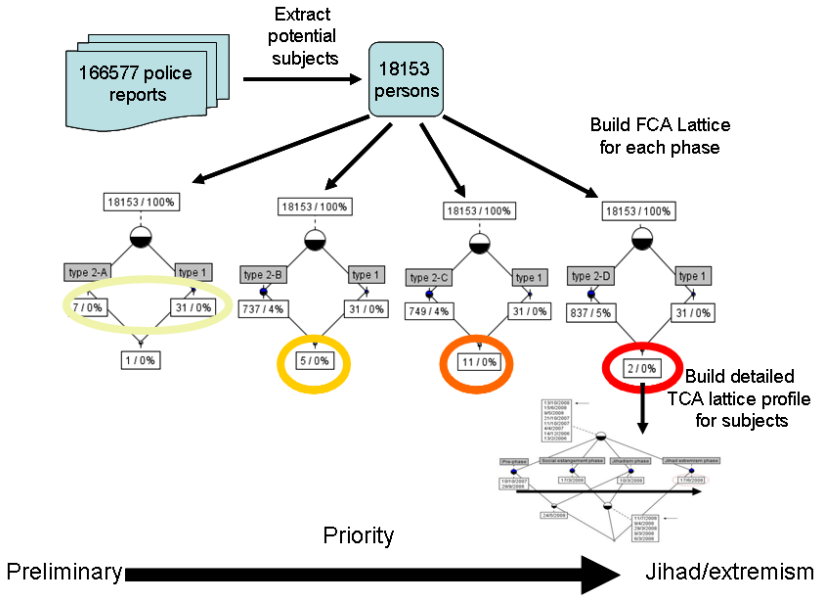


Fig. 4. 3 The process model of extracting and profiling potential jihadists

We used a toolset which was developed for text mining in large sets of documents, extracting entities from these sets and generating cross tables in various formats. It has been used to develop and to apply knowledge models for amongst others detecting domestic violence (Elzinga et al. 2009). The toolset uses a thesaurus where indicators can be defined with specific properties. For the purpose of this investigation, the thesaurus and toolset are extended with the property of range of numbers of different occurrences of an indicator which must true.

4.1.5.1 Extracting potential jihadists with FCA

For detecting potential jihadists from the large amount of observations, we make use of an FCA lattice. The subjects mentioned in the reports are the objects of the lattice. The indicators observed during the observations for these subjects are combined in one feature vector. This results in an FCA lattice as displayed in Figure 4.4.

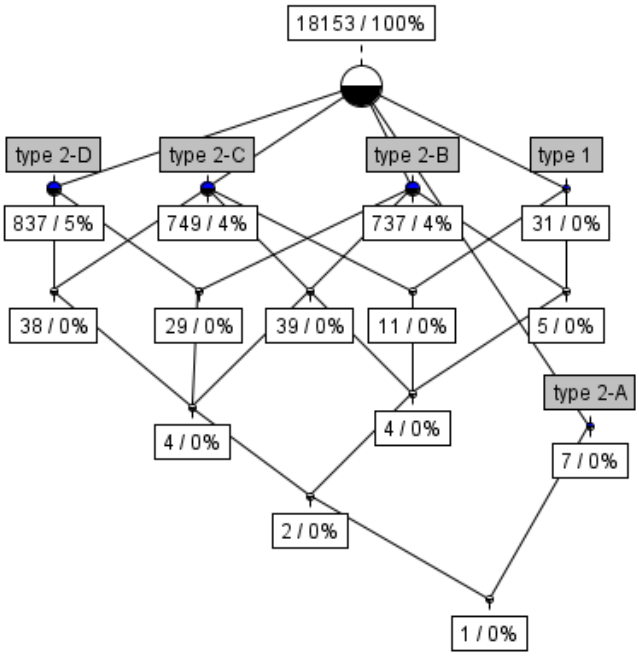


Fig. 4. 4 FCA lattice used for extracting potential jihadists

Out of 166578 documents 18153 subjects are selected into the cross table of FCA. Each of the subjects meets at least one of the 35 indicators of the original model. These indicators are grouped together based on the four phase model of the KLPD.

Table 4.3 Results of extraction of subjects

Attribute	# of subjects applied to
1	31
2-A	7
2-B	737
2-C	749
2-D	837

Table 4.3 shows the number of subjects who meet the requirements of the attributes of the four phase model. In the next section we will showcase how the combinations of type 1 and type 2 indicators will reveal the subjects in the different phases. One of the advantages of FCA theory is the ability to zoom in and out on the data and to create smaller lattices by amongst others deselecting the attributes from the main lattice.

4.1.5.2 Constructing Jihadism phases with FCA

The next step consists of constructing a lattice for each phase of jihadism and showing subjects. The FCA lattice serves as an intuitive knowledge browser making the interaction between the police officer and data more efficient. Based on this lattice, police officers can easily extract subjects for in-depth investigation. Figure 5.8 shows the process model of finding potential jihadists. The four lattices will be discussed from left to right.

The first lattice shows the preliminary phase where 38 subjects are detected. An in-depth search after these 38 subjects revealed that 19 were highlighted correctly. The other 19 subjects were mostly persons of the domestic sphere of the subject and therefore frequently reported in the same documents with the potential jihadists. Out of the 19 correctly highlighted subjects, 3 were previously unknown by the Amsterdam-Amstelland Police Department, but known by the National Police Agency Service. The second lattice shows 5 subjects for the social alienation phase which were all highlighted correctly. The third lattice shows 11 subjects for the Jihadization phase where 8 subjects are highlighted correctly. The fourth and last lattice shows 2 potential jihadists, who both are highlighted correctly.

4.1.5.3 Build detailed TCA lattice profiles for subjects

To show how the selected subject radicalizes over time a TCA lattice is constructed. C. from the example report is a subject who satisfies the conditions of all phases. Figure 4.5 shows the TCA lattice of C. We clearly see his radicalization process over time in action (black arrow). There were 8 observations of C. that did not trigger sufficient conditions for entering one of the four terrorism threat phases. In 29/9/2006, C. for the first time appeared under the preliminary phase and 13 months later again he was observed and again fulfilled the requirements of the preliminary phase. 5 months later, C. for the first time had all the properties of the social alienation phase and climbed from the fourth to the third phase of alert. Afterwards he was categorized 6 times under the second phase of alert: jihadism. In 17/6/2008 he reached the highest point of alert: Jihad extremism (red oval). Afterwards he was spotted 2 times by the police, once in the Jihadism phase and once outside any phases (2 arrows).

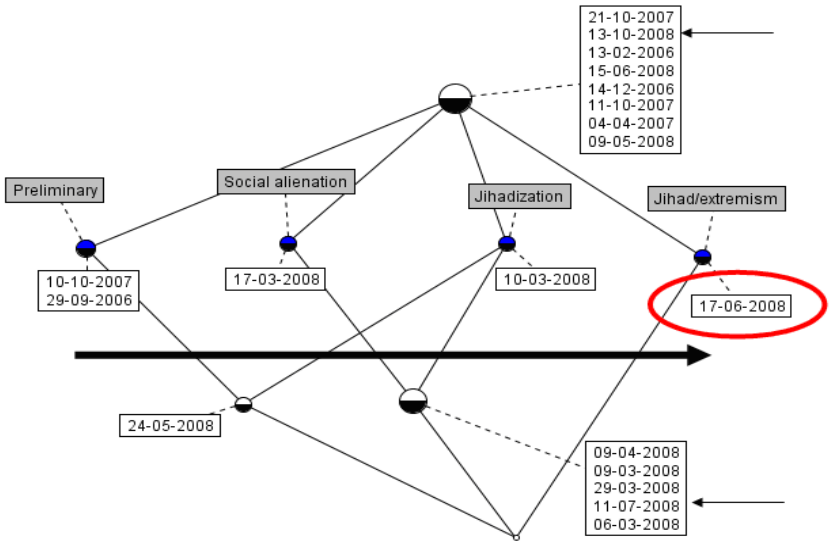


Fig. 4. 5 TCA lattice for subject C

4.1.6 Conclusions

In this section, we showed that the combination of techniques known as Formal Concept Analysis and Temporal Concept Analysis provides the user with a powerful method for identifying and profiling potential jihadists. We built a set of attributes based on the original knowledge model of radicalism which is used when searching the police reports. Out of 166,577 police reports we distilled and visualized 38 potential jihadism suspects using FCA. TCA is used to analyze the radicalization over time of the potential jihadists. Avenues for future research include the embedding of this sandbox discovery model into operational policing practice and applying.

4.2 Identifying and profiling human trafficking and loverboy suspects

4.2.1 Introduction

Irina, aged 18, responded to an advertisement in a Kiev, Ukraine newspaper for a training course in Berlin in 1996. With a fake passport, she travelled to Berlin, Germany where she was told that the school had closed. She was sent on to Brussels, Belgium for a job. When she arrived she was told she needed to repay a debt of US$10,000 and would have to earn the money in prostitution. Her passport was confiscated, and she was threatened, beaten and raped. When she didn't earn enough money, she was sold to a Belgian pimp who operated in Rue D'Aarschot in the Brussels red light district. When she managed to escape through the assistance of police, she was arrested because she had no legal documentation. A medical exam verified the abuse she had suffered, such as cigarette burns all over her body (Hughes et al. 2003).

In 2009, Amsterdam was shocked by the brute murder on an Eastern European woman who was forced to work in prostitution but resisted to her pimps[10]. Girls of Dutch nationality who were forced to work in prostitution in Amsterdam typically fell prey to a loverboy. The loverboy is a relatively new phenomenon (Bovenkerk et al. 2004) in the Netherlands. A loverboy is a man, mostly with Moroccan, Antillean or Turkish roots who makes a girl fall in love with him and then uses her emotional dependency to force her to work as a prostitute. Forcing girls and women in prostitution through a loverboy approach is seen as a special kind of human trafficking in the Netherlands (article 273f of the code of criminal law).

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 (United Nations 2004). 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” (Equality division 2006). Rough estimates suggest that 700,000 to 2 million women and girls are trafficked across international borders every year (O’Neill 1999, U.S. Department 2008). Since the fall of the Iron Curtain, the impoverished former Eastern bloc countries such as Albania, Moldova, Romania, Hungary, Bulgaria, Russia, Belarus and Ukraine have been identified as major trafficking source countries for women and children (Levchenko 1999, Dettmeijer-Vermeulen et al. 2008). The majority of transnational victims are trafficked into commercial sexual exploitation.

Because of the overload of mostly textual information in police databases and a lack of adequate supporting instruments to make this data more accessible, it becomes increasingly difficult to identify potential suspects and gather all available information about them. In this section we aimed at describing the new investigation procedures we developed with the Amsterdam-Amstelland Police Department for identifying and profiling potential suspects from this large amount of textual reports.

[10] http://www.politie-amsterdam-amstelland.nl/get.cfm?id=7963

Since the introduction of Intelligence Led Policing (Collier 2006, Viaene et al. 2009) in 2005, a management paradigm for police organizations which aims at gathering and using information to allow for pro-active identification of suspects, police officers are required to write down everything suspicious they noticed during motor vehicle inspections, police patrols, etc. These observational reports, 34,817 in 2005, 40,703 in 2006, 53,583 in 2007, 69,470 in 2008 and 67,584 in 2009 may contain indications that can help reveal individuals who are involved in human trafficking, forced prostitution, terrorist activities, etc. However, till date almost no analyses were performed on these documents.

We chose for a semi-automated approach with Formal Concept Analysis (FCA) lattices at its core (Ganter et al. 1999). These lattices are used to display the persons found in the available police reports and the early warning indicator observed for each of them. Police officers can then extract persons in whom they are interested and create a detailed profile for them. This profile is also an FCA lattice which displays all available information about this suspect, including social structure and temporal information, in one appealing visual picture. Our approach promotes efficient decision- making and significantly outperformed the currently employed manual investigation methods. The concept lattices revealed some cases where there were sufficient indications for starting an in-depth investigation. We applied FCA and its temporal variant to zoom in on some real life cases and suspects, resulting in actual arrestment's being made and/or illegal prostitution locations closed down.

4.2.2 Human trafficking and forced prostitution

The most popular destinations for trafficked women are countries where prostitution is legal such as the Netherlands (Hughes 2001). According to Shelley et al. (1999) most of these women are in conditions of slavery. Human trafficking and illegal forced prostitution are typically organized by international crime networks that make large amounts of money through the exploitation of young women and children. The money made by the criminal networks does not stay in poor communities but is laundered through bank accounts of criminal bosses in financial centers such as the US, Western Europe and off-shore accounts (Savona 1998). In Amsterdam, in particular Bulgarian and Hungarian criminals are active. Women who have been forced into prostitution can keep little or nothing of the money they earned. If they manage to escape they will return home in poverty and physically and emotionally damaged for life (Farley 1998). One of her only ways to escape the unwanted sex with multiple men each day is becoming a perpetrator herself. Women who fell prey to traffickers sometimes return home to recruit new victims. According to (Hughes 2003), 70 % of pimps in Ukraine are women. A recruiter gets US $2000 to $5000 for each woman recruited. Pimps can make 5 to 20 times as much from a woman as they paid for her in a short time.

4.2.2.1 Human trafficking model

Victims of human trafficking rarely make an official statement to the police (Tyldum 2005). The human trafficking team of the Amsterdam-Amstelland Police Department is installed to proactively search police databases for any signals of human trafficking. Unfortunately, this turns out to be a laborious task. The

investigators have to manually read and analyze the police reports, general reports and incident reports with human trafficking, one by one. The general reports are very poor labeled, only 10 to 15% of the total collection of general reports has a so-called project label like “prostitution”, “sex-related”. “domestic violence”, “explosives” and so on. As soon as the investigators find sufficient indications against a person, a document based on section 273f of the code of criminal law is composed for the person under scrutiny. Based on this report, a request is sent to the Public Prosecutor to start an in-depth investigation against the potential suspects. After permission is received from the Public Prosecutor, the use of special investigation techniques such as phone taps and observation teams is allowed.

The Attorney Generals of the Netherlands developed a set of guidelines based on which police forces can gather evidence of human trafficking and forced prostitution against potential suspects. These guidelines mention indications of human trafficking and forced prostitution and define in which cases pro-active intervention by police may be necessary. This information had not yet been used to actively search police databases for suspicious activity reports. Table 4.4 contains the five main types of indicators contained in these guidelines and two illustrative examples for each of them. The full list of indicators can be found in the first section of Appendix H.

Table 4.4 Human trafficking indicators

Dependency of the exploiter
The woman has a fake or counterfeit passport The woman does not know properly what her working address is.
Deprivation of liberty
The victim does not receive necessary medical treatment The victim does not carry her own identity papers
Being forced to work under bad circumstances
The victim receives an unusually low wage compared to the market. The victim has to work under all circumstances and unreasonably long
Violation of bodily integrity of the victim
Threatened or confronted with violence Certain things that may indicate the dependence of the exploiter such as tattoos or voodoo material.
Non-incidental pattern of abuse by suspect(s)
Working at different places from time to time Tips of reliable third parties

4.2.2.2 Loverboy model

Another model we discuss was developed by Bullens et al. (2000) for the identification of loverboys who typically force girls of Dutch nationality into prostitution. Loverboys use their love affair with a woman to force her to work in prostitution. Forcing girls and women in prostitution through a loverboy approach is seen as a special kind of human trafficking in the Netherlands (article 273f of the code of criminal law) as soon as the victim is 18 years or older. This model is a

resource used the Amsterdam-Amstelland Police Department during the trainings of police officers about this topic. A typical loverboy approach consists of four main phases. Table 4.5 contains the four main types of indicators and two illustrative examples for each of them. The full list of indicators can be found in the second section of Appendix H.

Table 4.5 Loverboy indicators

Preparatory activities to recruit girls
Actual recruitment and arranging residence and shelter locations for the girls During the first meeting, they estimate how vulnerable a girl is to attention and flattery. Their sensitivity to attention, presents, etc. made her fall in love with the pimp.
Forcing her into prostitution
Deflowering and forcible rape: In particular Islamic girls, deflowering and the threat of being brought back home increase their anxiety to say no to the pimp's demands, because it can result in her abandonment by her family. Blackmailing: If the girls don't want to work in prostitution, the pimps threaten to bring her back to her parents.
Keeping the girl in prostitution
Emotional dependence: Feelings of love, nobody else to support her, the pimp is the father of her child, etc. Social isolation: She becomes isolated from the outside world and only meets people from the prostitution circuit.
The pimp will also try to protect his organization
Internal protection measurements: He will make sure that the girls are constantly under surveillance and with the threat of physical violence he completely dominates her life. External protection: The pimp will threaten, bribe, interrogate, etc. the girls who have been in contact with the police.

4.2.3 Dataset

Our second dataset has been extended with general reports of 2005 and 2009 compared to the dataset used in the previous section, consists of 266,157 suspicious activity police reports, 34,817 in 2005, 40,703 in 2006, 53,583 in 2007, 69,470 in 2008 and 67,584 in 2009 and consists of general reports only, the labeled activity and or incident reports from the previous section are excluded.

4.2.4 Method

Our investigation procedure consists of multiple iterations through the square of Figure 4.6. For background information on FCA and its applications in KDD we refer to chapter 2. The guidelines of the human trafficking model contain a non-limitative list of indications and the indications can be subdivided into 5 main categories. If at least one of the thesaurus elements corresponding to these indications is present for a person or a group of persons, we might be dealing with a case of human trafficking or forced prostitution. From the 266,157 reports in our

dataset, the relevant reports which contain at least one indicator are selected. Then, the persons mentioned in these reports are extracted and FCA lattices are created, showing all the indications observed for each person. From these lattices containing persons, potential suspects or victims can be distilled and they can be further analyzed in detail with FCA and temporal concept lattices. If sufficient indications are available, a document based on article 273f of the code of criminal law can be created and sent to the Public Prosecutor with the request for using advanced intelligence gathering instruments such as observation teams, phone taps, etc. If the suspects are indeed involved in human trafficking and forced prostitution they can be taken into custody.

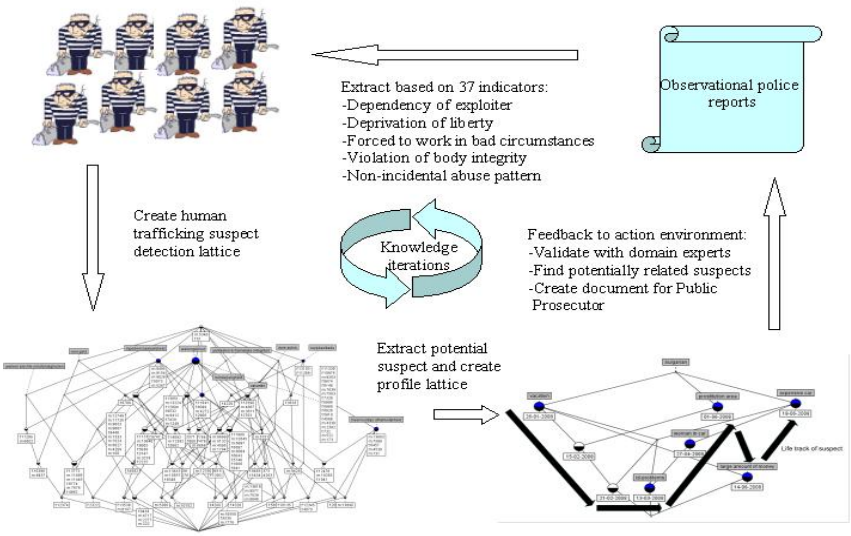


Fig. 4. 6 Criminal intelligence process

4.2.4.1 FCA analysis

Our method based on FCA consists of 4 main types of analysis that are performed:

- Concept exploration of the domestic violence problem of Amsterdam: In (Poelmans et al. 2010a, Poelmans et al. 2010b) our FCA-based approach for automatically detecting domestic violence in unstructured text police reports is described in detail. We not only improved the domestic violence definition but also found multiple niche cases, confusing situations, faulty case labelling, etc. that were used to amongst others improve police training. Part of the research reported on in this paper such as the construction of the thesaurus, consisted of repeating the procedures described in our domestic violence case study papers.
- Identifying potential suspects: Concept lattices allow for the detection of potentially interesting links between independent observations made by different police officers. When grouping suspicious activity reports

on a per person basis, the available information about the individuals is displayed in one intuitive and understandable picture that facilitates efficient decision making on where to look. In particular persons lower in the lattice can be of interest since they combine multiple early warning indicators.

- Visual suspect profiling: Some FCA-based methods such as Temporal Concept Analysis (Wolff 2005) were developed to visually represent and analyze data with a temporal dimension. Temporal Concept lattices were used in (Elzinga et al. 2010) to create visual profiles of potentially interesting terrorism subjects. Scharfe et al. (2009) used a model of branching time in which there are alternative plans for the future corresponding to any possible choice of a person and used it as the basis of an ICT toolset for supporting autism diagnosed teenagers. For creating the temporal profile of individual suspects, we use traditional FCA lattices and the timestamps of the police reports on which these lattices are based are used as object names. The nodes of the concept lattice can then be ordered chronologically.
- Social structure exploration: Concept lattices may help expose interesting persons related to each other, criminal networks, the role of certain suspects in these networks, etc. With police officers we discussed and compared various FCA-based visualization methods of criminal networks. Individual police reports mentioning network activity were used by us as objects and the timestamps of these police reports together with each suspect name mentioned in these reports as object names.

4.2.4.2 Thesaurus

The thesaurus constructed for this research contains the terms and phrases used to detect the presence or absence of indicators in these police reports. This thesaurus consists of two levels: the individual search terms and the term cluster level which was used to create the lattices in this paper. We used a semi-automated approach as described in (Poelmans et al. 2010a). Search terms and term clusters were defined in collaboration with experts of the anti-human trafficking team and gradually improved by validating their effectiveness on subsets of the available police reports. Each of these search terms were thoroughly analyzed for being sufficiently specific. The quality of the term clusters was determined based on their completeness. The validation of the quality of the thesaurus and the improvements were done by us and in conjunction with members of the anti-human trafficking team. Concept structures were created on multiple randomly selected subsets of the data. It was manually verified if all relevant indicators were found in these reports and no indicators were falsely attributed to these reports. For example, the term cluster “prostitute” in the end contained more than 20 different terms such as “prostituee”, “dames van lichte zeden”, “prosti”, “geisha”, etc. used by officers to describe a prostitute in their textual reports. To create the formal contexts in this paper, the term clusters in the thesaurus were used as attributes and the police reports as objects. Appendix C

shows an excerpt of this thesaurus with the term clusters and corresponding search terms.

A prototype of the FCA-based toolset CORDIET, which is described in detail in chapter 5, was used during the analysis process (Poelmans et al. 2010c). A new version is currently being developed under collaboration between the Katholieke Universiteit Leuven and the Moscow Higher School of Economics.

4.2.5 Analysis and results

Traditional data mining techniques often focus on automating the knowledge discovery process as much as possible. Since the detection of actual suspects in large amounts of unstructured text police reports is still a process in which the human expert should play a central role, we did not want to replace him, but rather empower him in his knowledge discovery task. We were looking for a semi-automated approach and in this section we try to illustrate the main reasons why FCA was ideal for this type of police work. With FCA at the core, we were able to offer police officers an approach which they could use to interactively explore and gain insight into the data to find cases of interest to them on which they could zoom in or out. Section 4.2.5.1 shows a lattice which was of significant interest to investigators of the anti-human trafficking team. For the first time, the overload of observational reports was transformed into a visual artifact that showed them a set of 1255 persons potentially of interest to the police and the indicators observed for each of them. The lattice visually summarizes the data and makes it more easily accessible for officers who want to efficiently explore it and extract unknown suspects. We chose first to highlight the case of the Turkish human trafficking network in section 4.2.5.2. From the lattice in section 4.2.5.1, two potential suspects were distilled since they were regularly spotted performing illegal activities. We found the name of a bar was mentioned a couple of times and used this information to build the concept lattice of section 4.2.5.2. This lattice was of particular interest to police officers since FCA quickly gave them a concise overview of the persons that were observed to be involved around a suspicious location and the lattice structure helped them to identify the most important suspects in this network. In particular the visualization of persons in a lattice was helpful during their exploration. FCA's partial ordering gave them clues on where to look first. The lower a person appears in the lattice, the more indicators he has. Section 4.2.5.3 showcases how the FCA visualization was used to combine temporal and social structure information in one easy to interpret picture. Such profile lattices were of significant interest to police officers since they allow for quick decision making on whether or not a person might be involved in illegal activities. Moreover, the lattices may help infer the roles of the persons mentioned in the network. The fourth case in section 4.2.5.4 is of interest, since it shows how an FCA lattice can give insight into the evolution of a person over time, in this case to detect the special case of a woman who was first victim and then became a suspect. Finally section 4.2.5.5 shows how an FCA lattice can give insight into the evolution of a person over time, in this case of a loverboy. The remaining part of this section describes cases of human trafficking and forced prostitution and two of them were identified in the lattice in Figure 4.7 and further investigated with FCA. Note that real names were replaced by false names because

of privacy reasons.

4.2.5.1 Detection of suspects of human trafficking and forced prostitution

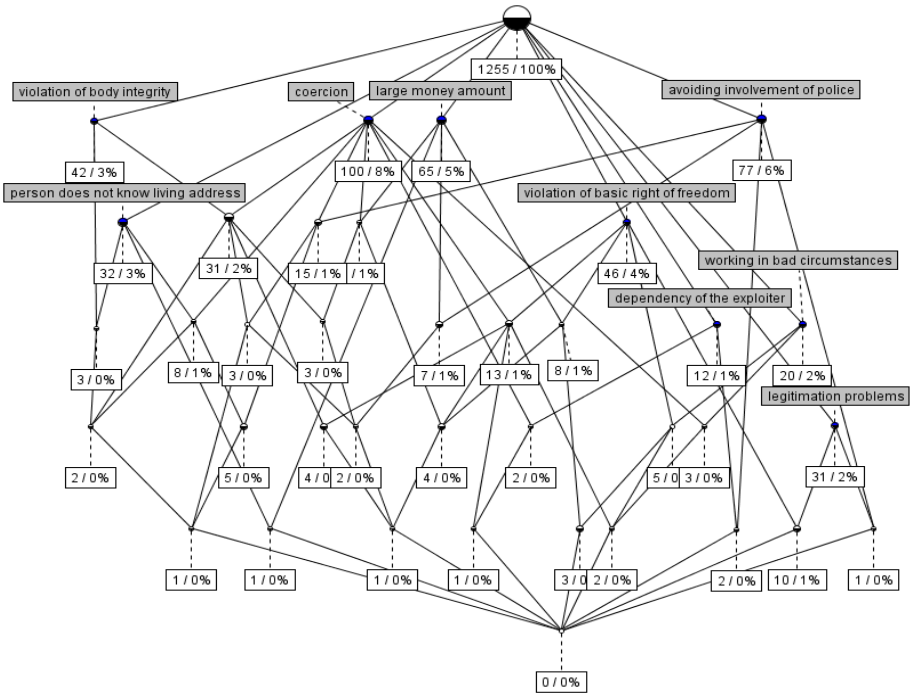


Fig. 4. 7 Human trafficking suspect detection lattice

Multiple concept lattices were created for detecting human trafficking suspects in the set of persons. Each of these concept lattices contained over 200 concepts and was based on different combinations of attributes. Since the format of this paper does not allow visualizing the entire lattices in a readable way, we chose to simplify one of these lattices and zoomed in on its most important aspects. Figure 4.7 contains the lattice with 1255 Bulgarian, Hungarian and Romanian persons. The concept containing some of the suspects of section 4.2.5.2 was found on the right and bottom part of the lattice and has 10 persons in its extent. The concept containing the main suspect of section 4.2.5.3 was found on the left and bottom part of the lattice and has 1 object in its extent. The next two sections will be used to describe and profile each of these suspects in detail.

4.2.5.2 Case 1: Turkish human trafficking network

By analyzing the concept lattice based on observational reports, we were able to expose a criminal network operating in Amsterdam, involved in illegal and forced prostitution. The concept lattice in Figure 4.8 contains the 61 persons and indicators found in the police reports mentioning activity around a bar in Amsterdam that played a central role in the network's activities and was closed down in 2009.

Multiple suspects operating in this network were found and some of the observations will be described in this section. The most important suspects are the persons with indication legitimating problems, since they were carrying the id papers of the girls. The police reports contained many indications of illegal and forced prostitution taking place, activities that were run by the owners or acquaintances of the owners of the bar. We found out the bar was used as a central hub, where mostly Turkish men met up with Bulgarian girls who had been forced into prostitution and took them to another location. We found at least two pimps who have multiple girls working for them.

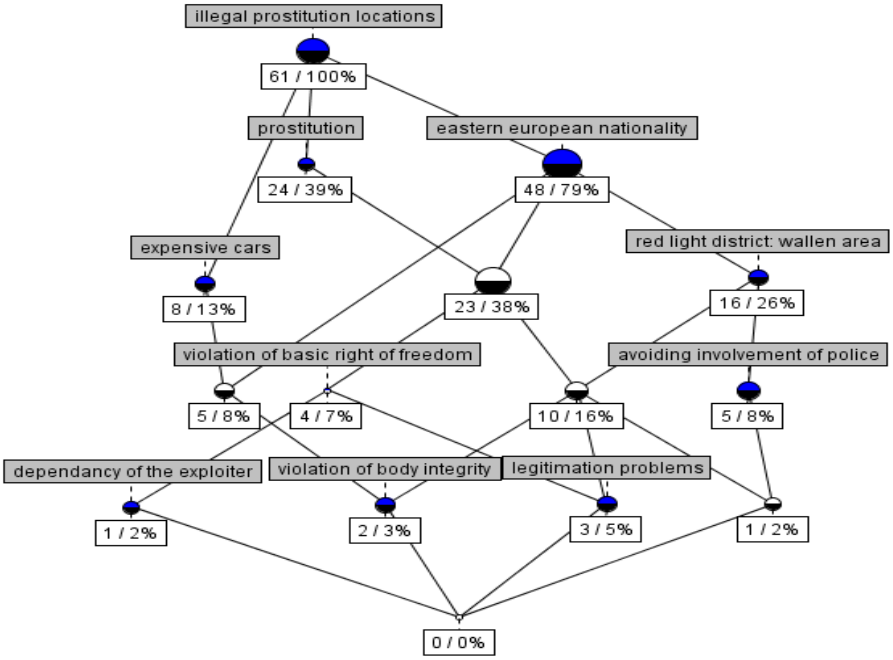


Fig. 4. 8 Concept lattice of human trafficking network

Starting in 2007, the first observations were made that hinted at illegal and forced prostitution being organized from within this bar. On 2 June 2008, victim H declared to the police that she was forced to work as a prostitute in the bar and did not get any money for that. She was never allowed to leave the house alone and the door of her apartment was locked from the outside such that she couldn't leave. On 12 December 2008, suspect A came out of the bar with a girl, their statements to the police did not match and moreover the girl was dressed in sexy clothing. Most likely the girl works as a prostitute and the driver is her pimp. On 25 January 2009, police officers stopped a car and behind the wheel was suspect B and next to him the victim E. We found woman E is often sitting at the bar and also the car is regularly parked in front of the bar. Suspect B gave the passport of victim E to the police and afterwards he placed it back in his pocket. Moreover, suspect B was carrying a large

amount of cash money, 1000 euro's in his pocket. On 26 January 2009, police did a check-up on the guests in the bar. One girl was new and told she only just arrived by train, she had no train tickets with her and she did not know her living address. Suspect B was also there and told the police he is a car trader so he travels a lot between Bulgaria and Netherlands. An excuse typically used by criminals responsible for the logistics of a trafficking network. Also victim E and two other girls, victims F and G were there. On 20 February 2009, police officers saw suspect A talking to the driver of a car with Bulgarian license plate. Afterwards he forced a girl to follow him and when the police asked about their relationship they told they had been friends for 3 months. The girl did not have her id-papers with her and the police went to her living address. In the house there were many mattresses and another girl. Both of them told they have no job. Most likely the house serves as an illegal prostitution location for the criminal gang.

Sufficient indications were found and on 17 June 2009, an observation team observed the bar during the evening. Eastern European women were sitting at the bar and mostly Turkish, Moroccan and Eastern European men at the tables. During the evening, the team saw multiple girls that were taken out of the bar by a customer to a hotel, house, etc. and brought back to the bar afterwards. On 15 July 2009 sufficient evidence was gathered that illegal prostitution was organized from within this bar and authorities closed down the bar.

4.2.5.3 Case 2: Bulgarian male suspect

In this section we describe a profile of a Bulgarian suspect who was also operating in Amsterdam. The lattice in Figure 4.9 shows that on 3 October 2007, suspect A was observed for the first time during a police patrol. An officer told the driver of a BMW car with Bulgarian license plate to turn right instead of left, the driver however ignored the instructions he received and quickly drove to the left with squeaking tires. The officer went after and in the end stopped the car. There were 3 men and one woman in the car. Suspect B was the driver and suspect A was sitting next to him. On the backseat of the car were woman F and man K. They told the officer they only arrived 3 days ago in the Netherlands and are a couple. Suspect A and suspect B were taken to the police office, the man and the woman walked away and was followed by a second officer. He saw that K was strongly holding the hand of F and forced her into a home at the corner of a street in central Amsterdam. In the police office, suspect B was not able to tell the address of the apartment he was going to rent. Suspect A was carrying a large amount of cash money in his pocket.

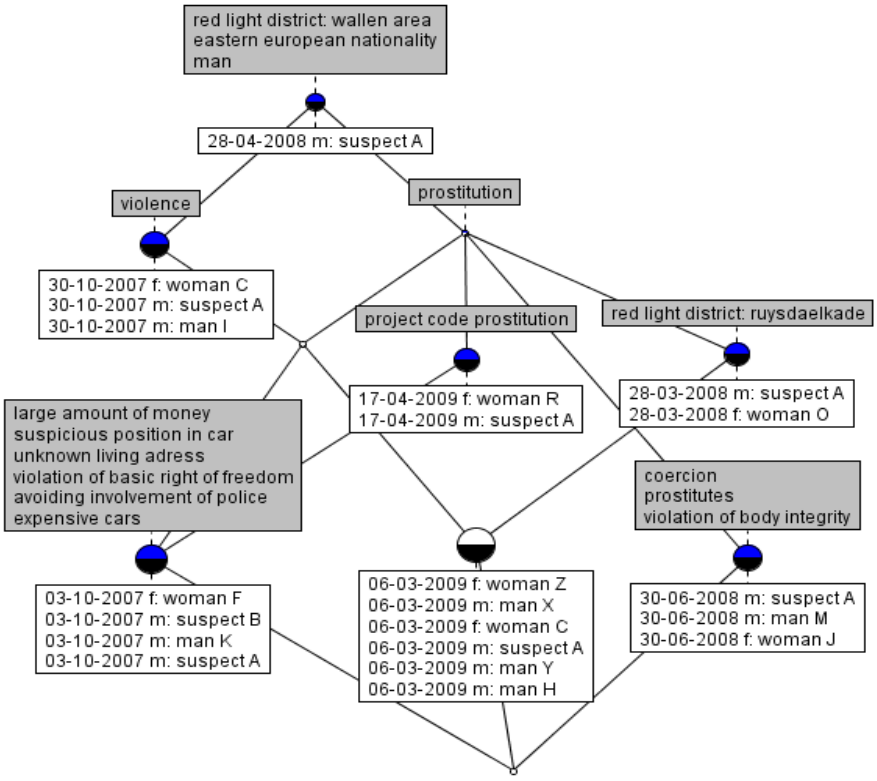


Fig. 4. 9 Profile lattice of individual suspect and his network

On 30 June 2009, woman J went to the police to ask if they could supervise the undersigning of a tenancy agreement of an apartment by man M who promised her accommodation. She told suspect A was intimidating and trying to scare away man M because suspect A wanted to rent the apartment for prostitution purposes. She was very afraid of suspect A and the officer noted that she might have been forced in prostitution by him. On 30 October 2007, the police did a routine inspection of 2 individuals who were waiting with two motorcycles in a street that had been plagued by street robberies. This was the second observation of suspect A by the police and his motorcycle was registered by the name of woman C who had been involved in human trafficking activities as a victim. On March 6th the police received a tip that a fugitive Colombian criminal might be living at a certain address owned by professional criminal H. When they entered the apartment they found 2 men and 2 women of Bulgarian nationality. Man X and woman C declared to be on holiday and would go back to Bulgaria although we found suspect A was driving around with a scooter registered at C's name in 2007. Man Y declared he exports expensive cars to Bulgaria and regularly drives back and forth between Netherlands, an excuse typically used by suspects taking care of logistics of a human trafficking gang. Woman Z declared to work in prostitution in Groningen. When the officers left the apartment they found a motorcycle registered on the name of suspect A. The last

observation dates back to 17 April 2009 when the police saw suspect A call somebody while standing in the entrance hall of prostitute R. He tells the police he has nothing to do with prostitution and owns a restaurant in Bulgaria. After his phone call he gives the cell phone to the prostitute.

To conclude, suspect A and B are most likely involved in human trafficking and there were sufficient signals found to request the use of special investigation techniques. Permission was granted, our suspicions were confirmed and both A and B were arrested by the police in 2010. Moreover these lattices showed some other people who are involved in the same gang and could be monitored.

4.2.5.4 Case 3: Hungarian woman both victim and suspect

In this section we describe a girl who was first a victim and then became a suspect of human trafficking. The concept lattice in Figure 4.10 contains indications that SV1 has been forced to work in prostitution but now also takes part in criminal activities such as "facilitating" new girls in the prostitution circuit.

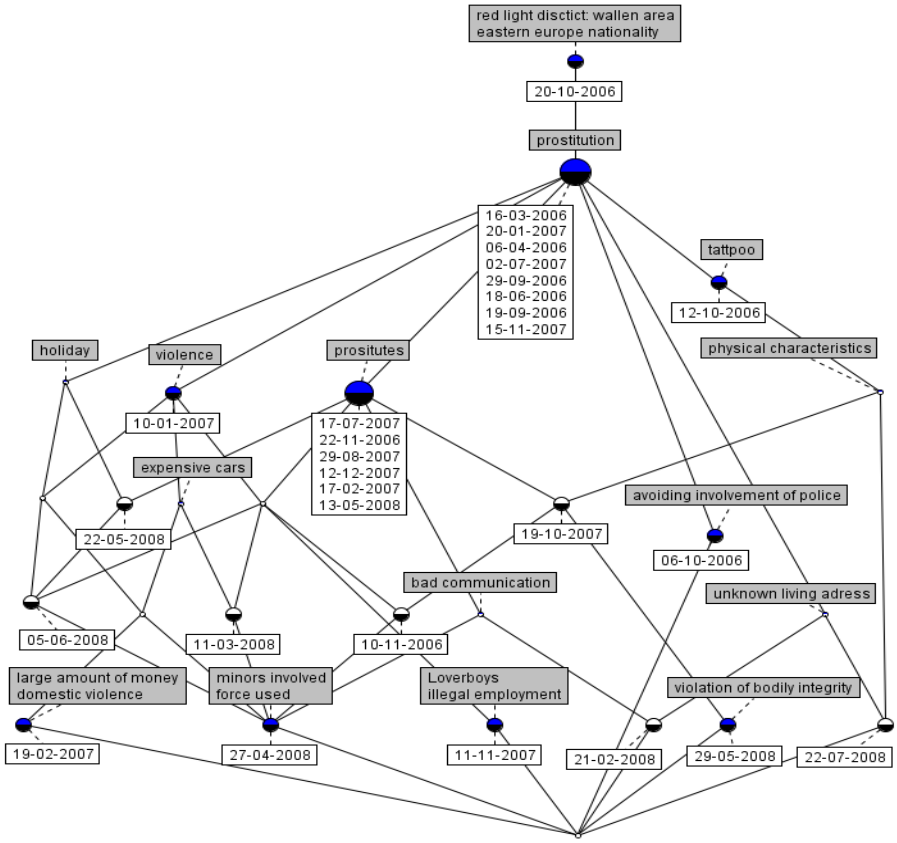


Fig. 4. 10 Profile lattice of women who was first victim and then became suspect

On 16-03-2006, woman SV1 was for the first time observed by the police in the red light district. She did not speak Dutch, English or any other language spoken by police officers in the Netherlands. She had all the indications of a woman who was lured into prostitution in her home country and trafficked to the Netherlands by a criminal gang. On 18-06-2006, the id-papers of SV1 and another girl I were checked and both pictures were very similar and had almost nothing in common with SV1 or I. Their id-cards were counterfeit, something regularly done by criminal gangs who took away their real identity papers. On 19-02-2007, prostitute Q declared to the police she had to give all her money to a Hungarian pimp who worked for a large criminal network. She told that also SV1 works for one of the pimps of this network and most likely undergoes the same treatment. On 19-10-2007, SV1 was observed with a new tattoo. Tattoos are regularly used by gangs to clearly show whose property the girl is. On 29-05-2008, officers saw SV1 underwent a breast enlargement.

From 2007 onwards, police officers started to see more and more indications that SV1 is becoming a perpetrator herself by facilitating girls in the prostitution circuit. On 02-07-2007, officers noticed that SV1 always pays the rent of the prostitution room for a new Hungarian girl L. On 17-07-2007, the police asked the id-card of the unknown woman who works as a prostitute and only resides in the Netherlands since 14 days. She did not know her living address; she lives with SV1 and was brought every day from and to her working place by SV1. The police asked if she likes her job but she had a very despairing look and could not answer their question. On 11-11-2007, police went to the lodging-house keeper of a room often rented by a Dutch girl D who worked in prostitution but mysteriously disappeared for multiple weeks. She told she was threatened by a group of Hungarian persons whom she met through SV1. They were trying to force her to work for them and give the money she earns away, amongst others through blackmailing, threatening and emotional manipulation. Amongst others on 15-11-2007, police saw SV1 having long conversations with Hungarian men for who she most likely works. She is granted more liberty than the other girls and seems to function as a kind of supervisor over the new girls who come into the business. On 13-05-2008, police did a routine inspection of 3 girls in the red light district but they only spoke Hungarian and SV1 was asked to translate their questions. When the police asked the girls about the place where they live, they became very nervous, tried to invent the name of a hotel, etc. In the end they asked to SV1 if they could tell their real address but SV1 answered no and if the police would try to force them they must first call the men of their network to ask for permission. On 22-07-2008, officers did a routine inspection in the red light district. Woman C was found to live together with SV1 and when the police asked her about their living address, C turned to SV1 who said in Hungarian “say whatever you want but don't tell the address”.

SV1 has many indications of a former victim of forced prostitution who had no better choice than becoming part of the criminal activities herself. She was part of a big network of Hungarian criminals that might be of interest to the police.

4.2.5.5 Case 4: Loverboy suspect

In this section we describe a loverboy case which we exposed by gathering evidence from multiple observational reports. This person was not found by analyzing the lattice in Figure 4.10 but by investigating a lattice based on Antillean, Moroccan and Turkish persons. Victim V is a girl of Dutch nationality who officially lived in the Netherlands but fell prey to a loverboy of originally Antillean nationality. We found multiple indications in filed suspicious activity reports that referred to elements of the human trafficking model. The lattice of suspect A and victim V is displayed in Figure 4.11.

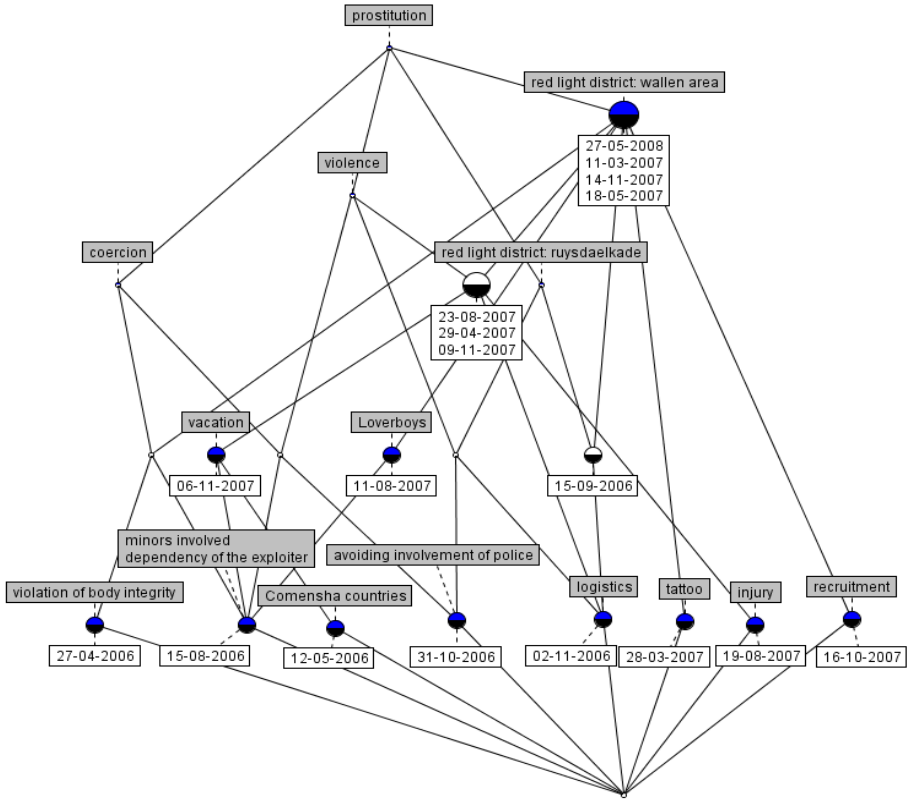


Fig. 4. 11 Profile of loverboy suspect

On 27-04-2006, Suspect A and victim V were noticed for the first time on the streets during a police patrol. They had a serious argument with each other and suspect A took the cell phone with force out of V's hand. When the police intervened they claimed nothing happened. In the police station she declared that she works voluntarily in prostitution although her words were not convincing to the officer. On 15-08-2006 an Amsterdam citizen sent an email to the police about young Antillean men who constantly surveillance some women in the red light district. Amongst

other suspect A brings food and drinks to the women who are not allowed to leave their rooms. On 31-10-2006 during a police patrol, victim V was noticed while she got out of a car and quickly ran inside. The driver of the car was suspect A. She told the police later on that she was brought to and picked up every day at this apartment by her boyfriend suspect A. The police noticed her dismayed and timid attitude and asked again if she was forced to work in prostitution. In a non-convincing way she responded that she did her job voluntarily. On 15-09-2006, suspect A had to stay in jail for 6 hours because of illegal weapon possession. When the police asked about his income he told he earned good money thanks to his girlfriend who works in prostitution. On 2-11-2006, officers noticed the car of victim V was parked on the road and two Negroid men were inside. The driver, suspect A got out of the car and yelled to the girl he was picking up at her apartment that she had to hurry up. The whole scene looked very intimidating to the police and it turned out the girl was victim V. Suspicious was that the car was registered on the name of V while V had no driver license. On 28-03-2007, victim B came to the police office to ask if she was allowed to work with a badly damaged id-document or if she had to wait for a new one. She mentioned that suspect A was her ex-boyfriend and that she and victim V were the victim of extortion but she did not dare to make an official statement to the police. Afterwards, the police checked a home where they found 2 women: victim V and B. Victim V had a big tattoo on her right shoulder and a smaller tattoo on her upper arm. On 19-08-2007, suspect A was involved in a knifing incident in the red light district between 3 men and one of these men got seriously injured. This man wanted sex with victim V but suspect A did not allow this because of the man's ethnicity, which caused the fight. On the camera surveillance videos, victim V was observed to accompany suspect A all the time. On 16-10-2007, officers observed that suspect A walked over the streets said hi to all women who passed by.

4.2.6 Discussion

Human-centered data mining focuses on making the human expert efficiently interact with the data by supporting him instead of trying to replace him. We wanted to help him in the laborious task of searching through the police reports and coming up with potential suspects but did not want to decide for him who should be investigated. The main goal of our semi-automated KDD in unstructured text approach is the active involvement of the human expert who steers the knowledge discovery process, sifts through the data and is supported in his decision making by visualizations that make the massive amounts of data that used to numb domain experts accessible again.

Our semi-automated approach works as follows. First, early warning, indicators are used to extract a pool of potential suspects. These early warning indicators are cheap and reliable indicators that may indicate involvement of a person in illegal activities but may result in some false positives remaining. They serve to reduce the search space effectively without losing suspects. Then, in the reduced search space, concept lattices based on early and late indicators are created. The presence of a (combination of) late indicator(s) is a strong hint that a person might be involved in illegal activities. Sometimes also a combination of early indicators is an interesting

situation for further analysis. The concept lattice visualization allows the human expert to zoom in on aspects of the reduced search space and interactively explore the data. He can steer the KDD process and the lattice partial ordering gives him clues on where to look first.

Police officers were found to be particularly interested in the following aspects of the FCA technique:

- Summarization of conceptual structure of data in one picture: the lattice of section 4.2.5.1 was used to showcase this appealing aspect of FCA. The overload of reports was turned into an intuitively analyzable artefact.
- An effective means to zoom in and out of the data: from the lattice in section 4.2.5.1, multiple persons were picked out and analyzed in detail in the subsequent sections.
- Intuitive visualization with a partial ordering of the persons based on the indicators observed. Police officers were guided by FCA is partial ordering when analyzing the lattice in section 4.2.5.1. Analysis indeed revealed they had more evidence to start a case against suspects lower in the lattice than suspects higher in the lattice.
- Conceptual relationships between individual documents, persons, timestamps, etc. became visible whereas they often stay hidden when individual documents are analyzed one by one: the lattice of section 4.2.5.2 was used to showcase how a criminal network operating in Amsterdam was exposed. Multiple independent observations contained indications that illegal network activity was performed around one central location.
- Visualization of temporal evolution of a person: the lattices in section 4.2.5.3 and 4.2.5.5 showed the evidence that became available over time against a human trafficking and loverboy suspect. Section 4.2.5.4 showed how a woman was first a victim and later on became a human trafficking suspect.

The literature on data mining describes many fully automated approaches for thesaurus building, classification, visualization, etc. Fully automated approaches have proven their usefulness for the analysis of certain crimes and criminals such as the identification of a serial killer's living address (ViCLAS system[11]). The algorithm is based on a domain with clear underlying rules and concepts takes as input a carefully prepared large amount of structured information about the suspect (over 260 attributes). The powerful pattern matching and computational capabilities of the computer clearly outperform the human expert in this task.

Unfortunately, in complex domains such as the domain described in this paper, it is very difficult if not impossible to be successful with pure automated analysis techniques. Many of these automated techniques may have serious drawbacks for complex domains with one or more of the following properties:

Black-box classification is not acceptable: police officers need insight into the reasons behind a decision, behind an assigned label, etc. Each decision to label a

[11] Violence Crime Linkage Analysis System, Royal Canadian Mountain Police, http://www.rcmp-grc.gc.ca/fs-fd/viclas-salvac-eng.htm

suspect should be grounded in evidence and be accompanied by a detailed report of the indications observed. False positives and false negatives are unacceptable given the severity of the crimes in which the persons are potentially involved and the penalties they may receive.

Texts are short, of equal length and written by authors with different writing styles: This makes it impossible and useless to apply term extraction techniques such as frequency analysis, etc. The terms we obtained through software packages such as DataDetective and Clementine were not satisfying either. Advanced NLP techniques were tried out in the past but failed because of the short textual reports. Relationship between persons, documents and networks play an essential role but are hard or impossible to automatically distill from the texts, etc. An essential element to the success of a text mining approach is a high-quality thesaurus. We choose for a semi-automated thesaurus-building-approach and complemented it with following automated methods to maintain quality: word stemming, synonym lists, spelling checking, etc. We also use Named Entity Recognition for extracting license plates, suspect names, etc.

Contexts of words and phases are essential for interpretation of the data: The interpretation of words, phrases, etc. is often strongly dependent of the context in which they are used. For example, during a police patrol, an officer checks a new prostitute and asks her about the scars on her legs. He wrote down that she told that during her childhood she was sexually abused and beaten but then suddenly their conversation was interrupted by the pimp who brought her food. The attributes “sexually abused”, “pimp”, “bring food”, “scars” may lead to a false positive although this document alone is far from sufficient to start a forced prostitution case. Moreover, multiple persons are mentioned in many reports and their roles such as suspect, victim or both are difficult to distill from these reports, even with advanced NLP instruments. Also some attributes should be solely attributed to one person but often it is impossible to automatically infer to which one. Human decision making remains necessary.

Only little information is available per person and the target group is a small fraction of the total population. The information we have is naturally incomplete since the reports written by officers describe only a part of the reality, namely that part observed by them during their work. The police does only have information about fragments of these persons’ lives based on which they decide whether or not this person might be interesting. Given the incompleteness of the information, one should take caution with fully automated decision making and leave this critical task to specialized and trained police officers who can judge whether or not sufficient evidence is available and slightly vary their decision criteria bared on their years of experience in the field. The focus of our approach lies on the development of an early warning system that helps to reduce the pool of potential suspects, gather all information about them in one visual picture that supports the officers in efficient decision making on which case should or should not receive special attention.

There have been no labels assigned to individuals or reports: our data did not contain any labeled individuals. Moreover, the target group is a small fraction of the total population. Training an automated classifier became impossible. To identify

phrases referring to forced prostitution during thesaurus construction we had to rely on expert knowledge.

The underlying concepts of the domain are unclear: the conceptual relationships between persons, documents, locations, etc. were of significant importance and had to be made visible to officers since they are essential to decision-making. Many visualization techniques such as Self Organizing Maps only give a distribution of the persons, documents, etc. but the relations between them are not explicitly shown.

A potential issue and avenue for future research is scalability of the approach. Lattices are only readable for a certain amount of concepts. Therefore, in each lattice we must limit the member of attributes and/or objects. This was however not a serious problem in our case since we were working with a stationary dataset, in which only a small part of the individuals was of interest. For other types of crimes such as credit card fraud detection, where we are dealing with massive amounts of fast changing data, FCA should be complemented with other visualization techniques such as ESOM. Another issue is the potential evolvement over time of the terms and phrases used by officers to describe their observation. Our thesaurus may become incomplete and maintenance methods should be developed to keep our system up to date on the long term.

4.2.7 Conclusions

Textual documents contain a lot of useful information that is rarely turned into actionable knowledge by the organizations that own these data repositories. The Amsterdam-Amstelland Police Department dispose of a large amount of such textual reports that may contain early warning indicators that can help to proactively identify persons involved in illegal activities. Since the observations of one suspect are typically made by different officers who are not aware of each others work, spread over multiple databases, etc. automated analysis techniques such as FCA can be of significant importance for police forces who are interested in the proactive identification of perpetrators. FCA is one of the few techniques that can be used to interactively expose, investigate and refine the underlying concepts and relationships between them in a large amount of data. In this paper we described our successful application of FCA to find suspects of human trafficking and forced prostitution in the Amsterdam-Amstelland Police Department district. From 266,157 observational reports we distilled multiple suspicious cases of which 3 have been described in this paper. For each of these persons and networks we composed a document containing all the indicators and evidence available and sent this to the Public Prosecutor. Permission to use special investigation techniques was obtained by the anti-human trafficking team based on these documents. For each case we exposed, phone-taps, observation teams, etc. indeed confirmed the suspect's involvement in human trafficking and forced prostitution. We believe that in making the shift from reactive police work, where action is only undertaken when a victim comes to talk directly to the police, to the pro-active identification of suspect's, FCA can play an important role.

CHAPTER 5

Concept Relation Discovery and Innovation Enabling Technology (CORDIET)

Concept Relation Discovery and Innovation Enabling Technology (CORDIET), is a toolbox for gaining new knowledge from unstructured text data. At the core of CORDIET is the C-K theory which captures the essential elements of innovation. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artefacts in the analysis process. The user can define temporal, text mining and compound attributes. The text mining attributes are used to analyze the unstructured text in documents, the temporal attributes use these document's timestamps for analysis. The compound attributes are XML rules based on text mining and temporal attributes. The user can cluster objects with object-cluster rules and can chop the data in pieces with segmentation rules. The artefacts are optimized for efficient data analysis; object labels in the FCA lattice and ESOM map contain an URL on which the user can click to open the selected document.

5.1 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 (Collier 2006, Rattcliffe 2008). 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.

In this chapter we report on our recently finished and ongoing research projects on concept discovery in law enforcement and the CORDIET tool that is being developed based on this research. At the core of CORDIET is the Concept-Knowledge (C-K) theory which structures the KDD process. For each of the 4 transitions in the design square functionality is provided to support the data analyst or domain expert in exploring the data. First, the data source and the ontology containing the attributes used to analyze these data files should be loaded into CORDIET. In the ontology, the user can define temporal, text mining and compound attributes. The text mining attributes are used to analyze the unstructured text in documents, the temporal attributes use these document's timestamps for analysis. The compound attributes are XML rules based on text mining and temporal attributes. The user can cluster objects with object-cluster rules and can chop the data in pieces with segmentation rules. After the user selected the relevant attributes, rules and objects, the analysis artefacts can be created. The tool can be used to create

FCA lattices, ESOMs and HMMs. The artefacts are optimized for efficient data analysis; object labels in the FCA lattice and ESOM map contain an URL on which the user can click to open the selected document. Afterwards the knowledge products such as a 27-construction for a human trafficking suspect can be deployed to the organization.

Section 5.2 shortly describes the analysis artifacts used in this research. Section 5.3 discusses the data sources from which our datasets were distilled. Each of these datasets contained police reports from domains such as domestic violence, human trafficking and terrorism and these application domains are discussed in section 5.4. Section 5.5 discusses the overall system architecture of CORDIET and section 5.6 describes in detail the functionality of the tool. Section 5.7 showcases some data analysis scenarios. Finally, section 5.8 presents the main conclusions of this chapter.

5.2 Data analysis artefacts

In this section we briefly describe the data analysis and visualizations artefacts that can be created with the CORDIET software. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artefacts in the analysis process.

5.2.1 Formal Concept Analysis

Formal Concept Analysis (FCA), a mathematical unsupervised clustering technique originally invented by Wille (1982) 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 (Stumme et al. (1998), Poelmans et al. (2010f)). 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 analyzing and refining these underlying concepts and strongly engages the human expert in the data discovery exercise. The main goal is to make previously inaccessible information available for practitioners easy to interpret visual display. 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 (Poelmans et al. 2009). Traditional FCA is mainly using data attributes for concept analysis. We also used process activities (events) as attributes (Poelmans et al. 2010b). Typically, coherent data attributes were clustered to reduce the computational complexity of FCA.

5.2.2 Temporal Concept Analysis

Temporal Concept Analysis (TCA) is an extension of traditional FCA that was introduced in scientific literature about nine years ago (Wolff 2005). 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. 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.

5.2.3 Emergent Self Organising Maps

Emergent Self Organizing Maps (ESOM) (Ultsch 2003) 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 (Ultsch 2004). 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 (Ultsch et al. 2005). In the traditional SOM, the number of nodes is too small to show emergence.

5.2.4 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical technique that can be used to classify and generate time series. A HMM (Rabiner 1989) 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 labeled according to the observation symbol probability.

5.3 Data sources

In this thesis three main data sources have been used. The first data source was the police database “Basis Voorziening Handhaving” (BVH) of the Amsterdam-Amstelland Police Department. Multiple datasets were extracted from this data source, including the domestic violence, human trafficking and terrorism dataset. The second data source was the World Wide Web, from which we collected over 700 scientific articles. The third dataset consist of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008.

5.3.1 Data source BVH

The database system BVH is used by all police forces of the Netherlands and the military police, the Royal Marechaussee. This database system contains both structured and unstructured textual information. The contents of the database are subdivided in two categories: incidents and activities. Incident reports describe events that took place that are in violation with the law. These include violence, environmental and financial crimes. During our research we analyzed the incident reports describing violent incidents and we aimed at automatically recognizing the domestic violence cases.

Activities are often performed after certain incidents occurred and include interrogations, arrestment, etc., but activities can also be performed independent of any incident, such as motor vehicle inspections, an observation made by a police officer of a suspicious situation, etc. Each of these activities performed are described in a textual report by the responsible officer. We used the observations made by police officers to find indications for human trafficking and radicalizing behavior.

In the year 2005, Intelligence Led Policing was introduced at the Amsterdam-Amstelland Police Department, resulting in a sharp increase in the number of filed activity reports describing observations made by police officers, i.e. from 34817 in 2005 to 67584 in 2009. These observational reports contain a short textual description of what has been observed and may be of great importance for finding new criminals. The involved persons and vehicles are stored in structured data fields in a separate database table and are linked to the unstructured report in a separate database table using relational tables. The content of all these database tables is then used by the police officer to create a document containing all the information. We however did not use these generated documents because it is possible that the information in the database tables is modified afterwards without updating the generated documents.

Therefore, we wrote an export program that automatically composes documents based on the most recent available information in the databases. These documents are stored in XML format and can be read by the CORDIET toolset. The structure of the input data is described in section 5.5.

Before our research, no automated analyses were performed on the observational reports written by officers. The reason was an absence of good instruments to detect the observations containing interesting information and to analyze the texts they contain. Only on the structured information stored in police databases, analyses were performed. These include the creation of management summaries using Cognos information cubes, geographical analysis of incidents with Polstat and data mining with DataDetective (Van de Veer et al. 2009).

5.3.2 Data source scientific articles

For the survey of FCA research articles, we used the CORDIET toolset. Over 700 pdf files containing articles about FCA research were downloaded from the WWW and automatically analyzed. The structure of the majority of these papers was as follows:

1. Title of the paper
2. Author names, addresses, emails
3. Abstract and keywords
4. The contents of the article
5. The references

During our research we used parts 1, 2 and 3. Parts 2 and 3 to detect the research topics covered in the papers. Part 1 was used for doing a social analysis on the authors of the papers i.e. which research groups are working on which topics, etc.

During the analysis, these pdf-files were converted to ordinary text and the abstract, title and keywords were extracted. The open source tool Lucene was used to index the extracted parts of the papers using the thesaurus. The result was a cross table describing the relationships between the papers and the term clusters or research topics from the thesaurus. This cross table was used as a basis to generate the lattices.

We only used abstract, title and keywords because the full text of the paper may mention a number of concepts that are irrelevant to the paper. For example, if the author who wrote an article on information retrieval gives an overview of related work mentioning papers on fuzzy FCA, rough FCA, etc., these concepts may be irrelevant however they are detected in the paper. If they are relevant to the entire paper we found they were typically also mentioned in title, abstract or keywords.

One of the central components of our text analysis environment is the thesaurus containing the collection of terms describing the different research topics. The initial thesaurus was constructed based on expert prior knowledge and was incrementally improved by analyzing the concept gaps and anomalies in the resulting lattices. The thesaurus is a layered thesaurus containing multiple abstraction levels. The first and finest level of granularity contains the search terms of which most are grouped together based on their semantical meaning to form the term clusters at the second level of granularity.

The term cluster “Knowledge discovery” contains search terms “data mining”, “KDD”, “data exploration”, etc. which can be used to automatically detect the presence or absence of the “Knowledge discovery” concept in the papers. Each of these search terms were thoroughly analyzed for being sufficiently specific. For example, we first had the search term “exploration” for referring to the “Knowledge Discovery” concept, however when we used this term we found that it also referred to concepts such as “attribute exploration” etc. Therefore we only used the specific variant such as “data exploration”, which always refers to the “Knowledge Discovery” concept. We aimed at composing term clusters that are complete, i.e. we searched for all terms typically referring to for example the “information retrieval” concept. Both specificity and completeness of search terms and term clusters was analyzed and validated with FCA lattices on our dataset.

5.3.3 Data source clinical pathways

The third dataset consist of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008. They all followed the care trajectory

determined by the clinical pathway Primary Operable Breast Cancer (POBC), which structures one of the most complex care processes in the hospital. Before the patient is hospitalized, she ambulatory receives a number of pre-operative investigative tests. During the surgery support phase she is prepared for the surgery she will receive, while being in the hospital. After surgery she remains hospitalized for a couple of days until she can safely go home. The post-operative activities are also performed in an ambulatory fashion. Every activity or treatment step performed to a patient is logged in a database and in the dataset we included all the activities performed during the surgery support phase to each of these patients. Each activity has a unique identifier and we have 469 identifiers in total for the clinical path POBC. Using the timestamps assigned to the performed activities, we turned the data for each patient into a sequence of events. These sequences of events were used as input for the process discovery methods. We also clustered activities with a similar semantical meaning to reduce the complexity of the lattices and process models. The resulting dataset is a collection of XML files where each XML corresponds with exactly one activity.

5.4 Application domains

5.4.1 Domestic violence

In 1997, the Ministry of Justice of the Netherlands made its first inquiry into the nature and scope of domestic violence (Keus et al. 2000). 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 (Poelmans et al. 2009). 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 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 3.17 in section 3.8, one in the middle and two on the left of the map.

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 folk's home? They have no family ties with each other, however there is a clear dependency relationship between them.
- Faulty case labeling: 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.

5.4.2 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 (United Nations 2004). 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" (Equality division 2006). The Amsterdam-

Amstelland Police Department mainly focuses on fighting forced prostitution and sexual exploitation of women.

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 (Poelmans et al 2010c). 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 created with CORDIET is displayed in Figure 5.34 in section 5.7.3.1.3. 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.

CORDIET was used here complementary to some existing systems at the Amsterdam-Amstelland Police Department. The Amazone database contains a list of suspects and potential suspects and the information available about them in police databases. A person found with CORDIET can be added to this list and automatically an email is sent to interested police officers in case this person is observed again. The text mining attributes from the CORDIET ontology can also be used by TopicView, which automatically retrieves all documents from the BVH database (and other police data sources) and generates hypotheses from these data. These hypotheses may include relations between suspects, roles and activities performed by suspects, etc. and can be validated by police officers. These associations between persons and certain attributes can be used by CORDIET to create FCA input files.

5.4.3 Terrorist threat assessment

In the terrorist threat assessment case study (Elzinga et al. 2010), 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 (AIVD 2006). 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, jihadization 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 5.1 shows an example of a TCA lattice for a newly found suspect who went through all 4 phases.

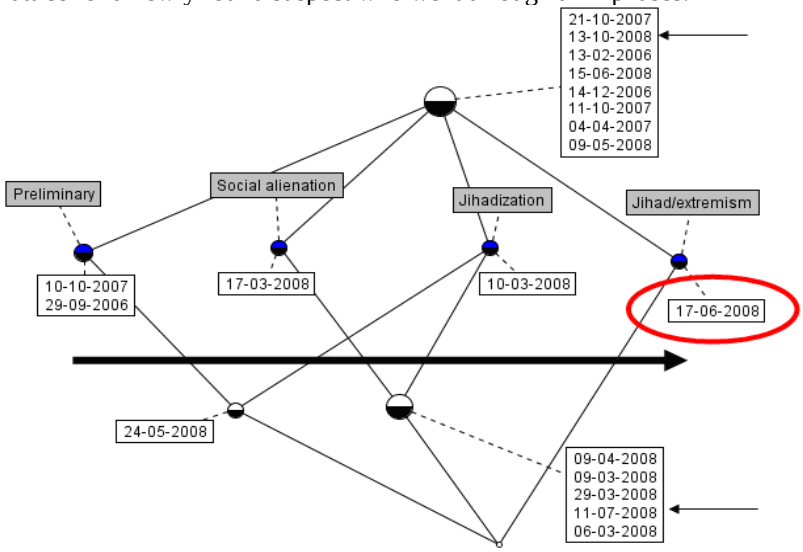


Fig. 5. 1 Found suspect who went through all 4 phases

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).

5.4.4 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 (Poelmans et al 2010d). 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 Department 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.

5.5 CORDIET system architecture and business use case diagram

5.5.1 Business use case diagram

In chapter 3 we instantiated the C- K design theory with FCA and ESOM and showed it was an ideal framework to structure the KDD process on a conceptual level as multiple iterations through a design square. The C-K theory is also at the core of CORDIET. For each C-K phase there are use cases that describe the functionality of the phases. The results of the use cases of a previous phase serve as input for the use cases of the next phases. The business use case model in Figure 5.2 clearly shows this C-K inspired architecture of CORDIET.

The first C-K space, “start investigation”, aims at transforming existing knowledge and information into objects, attributes, ontology elements etc. (conceptualization). The second C-K phase, "compose artefact”, will create artefacts from the data that visualize its underlying concepts and conceptual relationships (concept expansion). The third C-K phase, “analyze artefact”, is about distilling new knowledge from these concept representations. The fourth and last C-K phase is about summarizing this newly gained knowledge and feeding it back to the domain experts who can incorporate it in their way of working. After this final step, a new C-K iteration can start based on the original information and/or newly added knowledge. Iterating though the design square will stop when no new knowledge can be found anymore. In section 5.6 we will describe the use cases from the business use case diagram in more detail.

Fig. 5. 2 Business use case of CORDIET.

5.5.2 The software lifecycles of CORDIET

The architecture of the CORDIET software underwent some serious changes during the development of this thesis. During the first stage of this PhD we were working on the domestic violence data and CORDIET consisted of an FCA, ESOM component and a commercial text mining tool was used to index the documents. Our own programming took care of the documents extraction from the database and the conversion of the data to be used as input for the artefact creation components. This

first version had its limitations and was seriously modified for the terrorism and human trafficking research. Amongst others, indexing of documents was done with Lucene.

A separate RDBMS database was used for the maintenance of the ontology with an ERD model. The latest version used a topic map for maintaining the ontology and the open source topic map editor “ontopoly”. This latest version will be described in detail in this chapter.

5.5.3 The development of an operational version of CORDIET

The Katholieke Universiteit Leuven and the Moscow Higher School of Economics decided to jointly develop an operational software system based on the latest version of CORDIET toolbox. This system will be a user friendly application making visualizations such as FCA, ESOM and HMM available to its users. This version of the toolbox will be based on distributed web service architecture. Web services are a well standardized, easy to access and flexible piece of technology that can be adapted for different languages and environments.

As a consequence, all input/output activities are represented as XML. Figure 5.3 shows the general architecture of the new version of CORDIET.

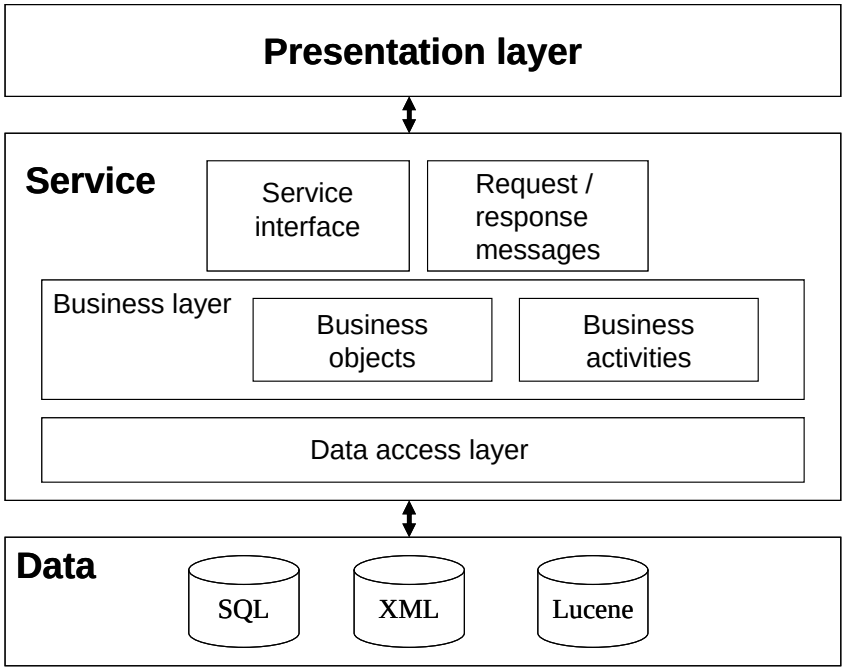


Fig. 5. 3 A representation of the CORDIET web service oriented architecture

5.5.3.1 Presentation layer

The presentation layer is the graphical user interface where the interactions of the user with the system are handled.

5.5.3.2 Service

The service layer will be the core of CORDIET. The service interface takes care of the I/O activities with the presentation layer and accessing of the data through the data access layer.

5.5.3.3 Business layer

The business layer is divided into two sections, the business objects and the business activities which refer to the different activities within the C/K cycle.

5.5.3.4 Data access layer

The data access layer is used to access the data sections: the relational database, the XML data and the Lucene indexes.

5.5.3.5 Data

The data sets consist of a relational database (PostgresSQL), a dataset with XML files and a Lucene index. The data-indexer component reads the XML files from a selected dataset, parses the XML into the SQL database and generates the Lucene index.

5.5.3.6 User interface

CORDIET will use two types of main visualizations. The master mode will mainly be used by domain experts who have limited knowledge of data analysis. The user will be able to load a profile for each of the four C-K transition steps, this profile contains all information the tool needs to automatically complete the step in the data analysis. This profile has been prepared by a data analyst.

The user can go to the advanced mode. In the advanced mode, he can fully edit an existing or create a new profile. In the advanced mode, a graph-like display will be used to create, modify and compose different attributes.

5.5.3.7 Language module

Different languages including English, Dutch and Russian should be supported. The user must be able to choose between these languages. The version of Lucene indexer of documents used has a large variety of analyzers like Russian Analyzer, Dutch Analyzer, German Analyzer etc. The default Analyzer is English.

5.6 CORDIET functionality

5.6.1 K->C phase: start investigation

Each investigation with CORDIET starts by choosing, loading and/or adding the dataset and the ontology to be used. The following sections will describe in detail the structure of these two important files and the semantics of their elements. Figure 5.4 shows the business use case of the K->C phase.

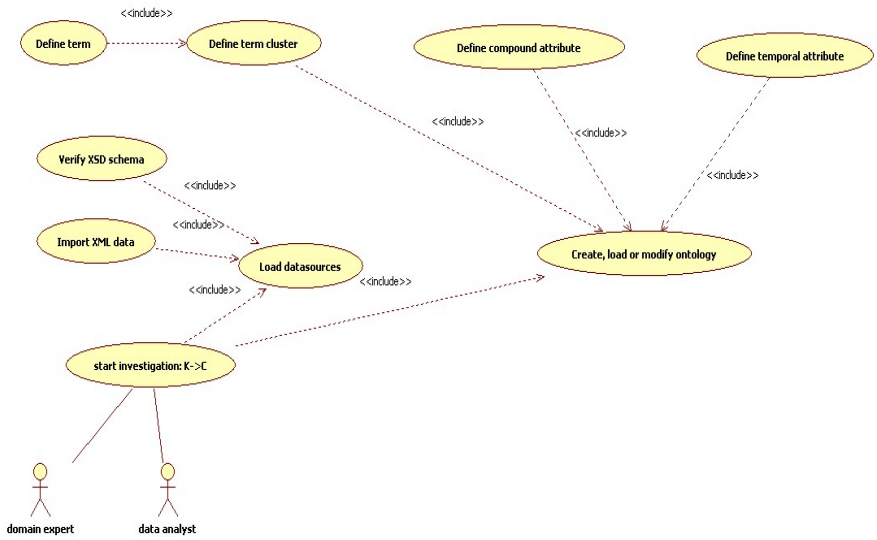


Fig. 5. 4 Start investigation: K->C

5.6.1.1 Load data sources

Data files used as input for the CORDIET software package should be in XML format. This XML file has an identifier and a number of structured data fields. For example an XML input file corresponding to a police report, contains the name of the suspect, the location of the incident and the textual report. Each of these fields has a value. The XML document also has a timestamp. For police reports this is the time of reporting the incident. Finally the XML document contains the unstructured text. For a police report this is the statement made by the victim or the observations made by the police officer. Each XML document contains only 1 data document, for example one police report or one patient and all the activities performed to this patient. For our dataset of 4814 domestic violence reports, we transform these reports in 4814 XML documents and store them into a PostgreSQL database.

5.6.1.2 PostgreSQL database:

With the PostgreSQL database a data ontology is associated which contains the structure of the input data so that the data can be read and stored into the database. This file is an XSD file and used to verify the well-formedness of the XML files using a SAX or DOM parser. Figure 5.5 shows the XSD scheme of the data sources and table 5.1 describes the X ML input source.

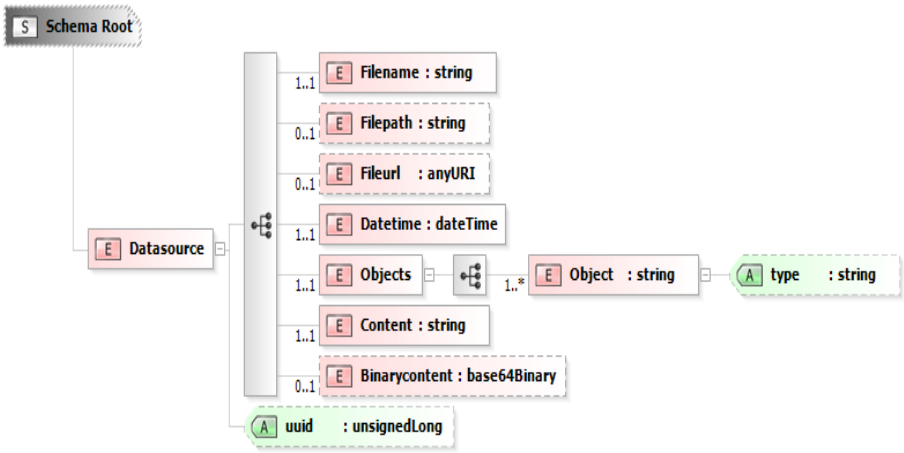


Fig. 5. 5 XSD scheme of the XML input source

Table. 5.1 Description of the XML input source of CORDIET

Element	Attribute	Description
Datasource		The element containing the information about the data. For each XML document there is exactly 1 data source element**.**
	UUID	Unique identifier used to communicate between the rdbms and the Lucene index.
Filename		Name of the file without path name.
Filepath		Full name of the file including path name.
Fileurl		URI with which the file can be accessed from an application or web service.
Datetime		Timestamp which can be used to execute and verify temporal rules
Objects		Collection of objects which will be indexed by Lucene as separate fields.
Object		The object itself with its value
	Type	Type of the object used to create fields within the Lucene index but also used as an option for the user to cluster objects.
Content		The unstructured text contained in the document.
Binarycontent		The original document.

The parsing procedure goes as follows. The unique identifier of the XML document is retrieved. It is verified if the database table already contains the XML file. If the document was already added, the record is updated; otherwise a new record is created. The timestamp of the file is retrieved and stored, in the database. The object node list is retrieved from the XML file, the element <object> is retrieved, the value of the attribute "type" is retrieved, and the value of the object is retrieved. A new

record is created in the data source table. The following attributes of the record are inserted: datasource_id, unique identifier UUID, filename, document path, document url, XML document. Then a new Lucene document is created.

5.6.1.3 Lucene:

Lucene index is used to index and optionally store the documents in the index. The field “content” will contain the unstructured text of the XML input file. The Lucene index stores for each term where it appears in which documents. Lucene allows to quickly search and find out in how many for example "Person" fields or "content" fields the name “Jan Janssen” appears. When filling the database, the documents are also stored in the Lucene index. Lucene has many interesting options for storing documents; options should be compared for optimizing performance. Timestamps and data are stored in Lucene without using the Analyzer, where the value will be stored as a single term. This means fields like "datetime" can be used for defining temporal rules by using the [… TO ...] operator-which returns all documents between two dates.

5.6.1.4 Create, load or modify ontology

The ontology is stored in XML format and should be loaded into the database. The XML syntax of the attributes, rules, etc. should be translated to the Lucene syntax. Table 5.2 gives the XML tags that can be used to create and compose ontology elements.

Table. 5.2 Description of the ontology XML tags

Tag	**Attribute**	***Create ontology element tags***
<searchterm>		This tag is used to define a new search term.
	proximity	If two or more words are contained in the search term, it can be specified they must occur within a specific distance from each other. This can be implemented in Lucene as follows to search for “bed” and “kitchen” within 5 words from each other in a document, we "bed bathroom "~ 5.
<term>		This tag is used to define a new term composed of a list of search terms
	name	Name of the term
	id	Identifier of the term
<termcluster>		This tag is used to define a new term cluster composed of a union of terms
	name	Name of the term cluster
	id	Identifier of the term cluster
<compoundattribute>		This tag can be used to create a new compound attribute
	name	Name of the compound attribute
	id	Identifier of the compound attribute

<temporal attribute>		This tag can be used to create a new temporal attribute
	name	Name of the temporal attribute
	id	Identifier of the temporal attribute
		Compose ontology elements tags
<union>		Tag to compose an ontology element such as a term, from other ontology elements, such as search terms. The attribute is true if at least one of the elements within the tag is present.
	occursmin	Indicate how many of the elements inside the tags must be present for the attribute to be true
<intersection>		Tag used to indicate that all ontology elements inside these tags must be true, for the attribute to be true
<xmlrule>		Tag used to indicate that a xml rule is defined here
<and>		Tag used to interconnect two items that jointly must be evaluated to true
<or>		Tag used to interconnect two items of which one should be evaluated to true
<not>		Tag used to indicate that the negated element should not be true
<cartesian>		Tag that can be used to make a cartesian product of a literal and the contents of a term (cluster)
		Create rules tags
<segmentationrule>		Tag used to create a new segmentation rule
	name	Name of the segmentation rule
	id	Identifier of the segmentation rule
<objectclusterrule>		Tag used to create a new object-cluster rule
	name	Name of the object-cluster rule
	id	Identifier of the object-cluster rule
<classifierrule>		Tag used to create a new classifier rule
	name	Name of the classifier rule
	id	Identifier of the classifier rule
<foreach>		
<objecttype>		This tag refers to which input data field and which Lucene field an operation should be performed
<color>		For a classifier rule, to specify the visualization color of cases

The PostgreSQL database not only contains an ontology defining the data

architecture but also a domain ontology defined by the user containing terms, term clusters, temporal rules, compound attributes, etc. Text mining attributes, temporal attributes and compound attributes can be added and removed using the respective ontology maintenance modules. The value of specific data fields in the documents can also be used as attributes. If certain terms should only be searched for in a certain Lucene data field, this can be indicated with a compound attribute.

5.6.1.5 Text mining attributes

A text mining attribute can be a term or a term cluster. A term is an array of search terms. A term cluster is a list of terms. For example the term cluster "family" consists of the terms “mother”, “father”, “uncle”, etc. The term '”father” consists of the search terms “my father” “my dad”, “my daddy”, etc.
These are 2 examples of text mining attributes. The XML syntax is translated to Lucene syntax and applied on the Lucene index.

5.6.1.6 Temporal attributes

A temporal attribute consists of a name and an XML rule that uses timestamps available in the data. It uses the timestamps of the police reports. A list of examples will be given in the temporal attributes section. A XML syntax should be introduced for defining these rules. It is possible to use the date field in Lucene. A temporal rule language should be defined for working with these dates. Complex rules should be transformed to operations on these date fields.
Temporal rule examples:

1. Find all criminals that were seen 4 times or more by the police between January 2009 and June 2009. This rule can be used to find unknown repeat offenders.
2. Find all victims from domestic violence that were reported in general reports 2 times or more within a time span of 6 months. This rule can be used to find domestic violence cases where the victim does not want to make a statement against the perpetrator.

5.6.1.7 Compound attributes

Compound attributes have a name and XML rule. This XML rule uses text mining attributes. Again a XML syntax should be defined.
Compound attributes examples:

1. All documents mentioning a term referring to a violent incident and a term referring to a person from the domestic sphere of the victims .
    - This rule uses the text mining attributes "violence" and "person from domestic sphere". The text mining attribute ”violence” contains terms such as “beat”, “kick”, “scratch”, “strangle”, etc. Terms such as “beat” contain search terms such as “beaten”, “heated”, “beaten up”, etc.
    - The compound rule then indicates that attributes “violence”, “person from domestic sphere” and the temporal attribute must be true and present for the document.

2. A variation of example 1: all documents that contain “violence” and not “person of domestic sphere” attribute.
3. All documents with domestic violence label.
    - This rule used the value of the object type of “projectcode” which consists a list of values from “HG1.1” to “HG1.14”. The compound attribute can be used as classifier when training datasets with ESOM

5.6.2 C->C phase: compose artefact

In the C->C phase the user can select and adjust the parameters needed for generating the desired artefact, an FCA lattice, an ESOM map or a Hidden Markov Model. Figure 5.6 shows the business use case of the C->C phase.

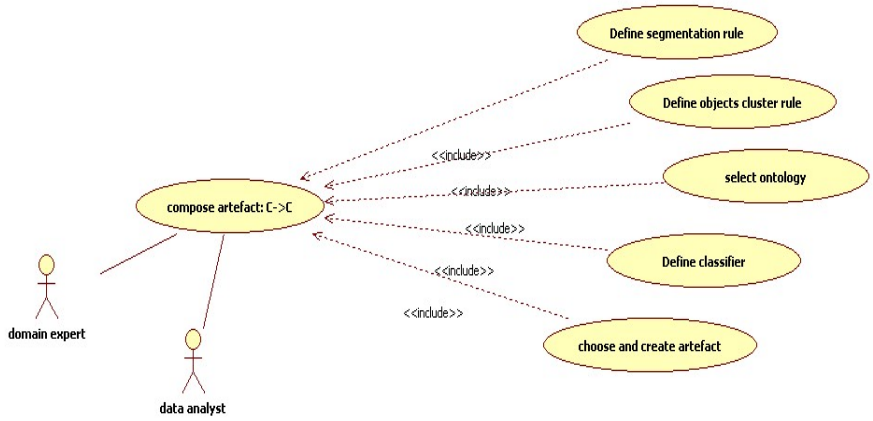


Fig. 5. 6 Compose artefact: C->C

5.6.2.1 Select ontology

The user selects an ontology containing the desired attributes for building the artefact. The selected ontology is stored in XML format and is than loaded into the database. The user has the option to use the entire ontology or make a selection of some ontology elements. A XML syntax should be defined. We will now shortly describe the ontology we created for the domestic violence investigation area. This ontology contains text mining attributes like “domestic sphere” which contains search terms related to all members of the domestic sphere. The ontology also uses compound attributes like “acts of violence within domestic sphere” which is composed of a cartesian product of the text mining attribute “acts of violence” and text mining attribute “domestic sphere”.

5.6.2.2 Define rules

Segmentation, object cluster and classifier rules are stored in XML format and should be loaded into the database. The XML syntax of the attributes, rules, etc. should be translated to the Lucene syntax. Table 5.x in the previous sections gives

the XML tags that can be used to define the rules.

5.6.2.2.1 Segmentation rules

Segmentation rules have a name and XML rule. This XML rule uses text mining attributes and values of object tags. Segmentation rule examples:

1. All documents with observation date in the year 2009 and events observed in the red light district should be retrieved.
    - This rule uses the object type “observationdate” and applies the range from 20090101 to 20091231.
    - The rule also uses the text mining attribute “red light district” with all search terms containing references to the red light district area. The search terms vary from street names, names of bridges to names of sexclubs.
2. All documents of the social network of suspect A should be retrieved.
    - This rule uses the object type “suspect” from the fields of the Lucene index and verifies for each document if in this field matches the exact value of the name of suspect A.
3. All documents of a suspicious pub or coffeeshop should be retrieved.
    - This ruse used the object type “location” from the fields of the Lucene index where the pub or coffeeshop is located and verifies for each document if this field matches the exact value of the location of the pub or coffeeshop.

5.6.2.2.2 Object cluster rules

Object cluster rules have a name and a XML rule. The rule uses the index fields of Lucene and depends on the number of different object types from the XML data files. Object cluster rule examples

1. Document level.
    - The individual documents are not clustered and attributes are assigned to individual documents. This rule is the default rule
2. Date level
    - This rule clusters the documents based on their time stamp and attributes are assigned to such clusters if at least one of the documents in this cluster has the attribute.
3. Person level
    - This rule clusters the documents based on the person involved in the crime.
4. Location level
    - This rule clusters the documents based on the location of the crime scene.

5.6.2.2.3 Classifier rules

Classifier rules have a name and a XML rule. The rule uses only compound attributes and a color tag. It generates a true if the requirements of the classifier rule are met or false if the requirements are not met. Classifier rule examples:

1. Labeled domestic violence classifier.

- This rule is composed of a compound attribute with object type “projectcode” and consists of a list of values from “HG 1.1” to “HG 1.14” and a optional color tag “red” which displays all domestic violence cases as red dots in a generated ESOM map.

2. Prostitution classifier
    - This rule is composed of a compound attribute with one text mining attribute, “prostitution” containing all search terms related to prostitution.

5.6.3 Choose and create artefact

5.6.3.1 C->K phase: analyze artefact

In the C->K phase the user can detect objects of interest, detect anomalies and detect new knowledge concepts by analyzing the artefacts. Figure 5.7 shows the business use case of the C->K phase.

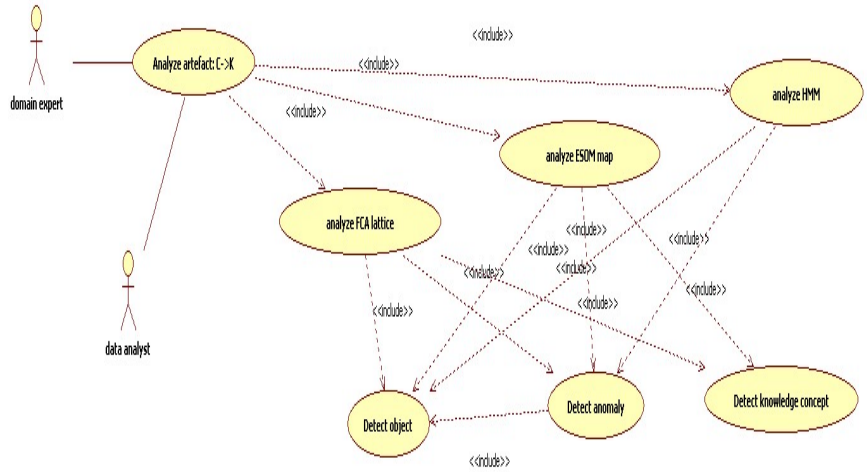


Fig. 5. 7 Analyze artefact: C->K

5.6.3.1.1 Detect object of interest

Depending on which artefact is analyzed in combination with the selected object-cluster rule, different kinds of objects can be detected. Examples are

1. Documents.
    - Using an FCA lattice, individual documents can be selected and inspected to gain knowledge about the concept to which the document belongs. In case of human trafficking, documents with evidence can be found.
2. Persons
    - A FCA lattice containing, all documents referring to a selected person can give insights in the profile of the person. In for

example the case of human trafficking whether he or she has the role of suspect, victim or both.
- Using Hidden Markov Model process models of clinical pathways and individual patients can be visualized.

3. Companies
   - An FCA lattice containing all documents referring to a selected company, like a pub, can give insight in possible illegal activities committed by clients of the pub. An example is a recently closed pub which was used as a meeting point for prostitution. Clients of prostitution could pick up a prostitute, use their services and bring her back.

5.6.3.1.2 Detect anomaly

Depending on which artefact is analyzed different kinds of anomalies can be detected. Examples are:

- Using an FCA lattice, concepts missing subconcepts or having conflicting subconcepts can be selected, the documents belonging to the concept can be inspected. In case of domestic violence, wrongly classified documents can be found where no violence is involved (missing subconcept) or a domestic violence case with an unknown suspect (conflicting subconcept).
- Using an ESOM map, outliers in the toroid map might give indications to wrongly classified documents. After inspecting these documents, this might lead to a new concept which detects wrongly classified documents. Examples are a new text mining attribute or a new compound attribute which conflicts with the definition of domestic violence.

5.6.3.1.3 Detect knowledge concept

Depending on which artefact is analyzed different kinds of knowledge concepts can be detected. Examples are:

- Using an FCA lattice, combinations of concepts may lead to new classification rules. Examples are the combination of suspect and victim living at the same address and the address is not associated with an organization like the Salvation Army.
- Using an ESOM map, individual documents can be selected that might lead to new attributes that can be used to create an FCA lattice in which new concepts emerge. Examples are new search terms of an existing text mining attribute, a new text mining attribute or a new compound attribute.

5.6.3.2 K->K phase: deploy knowledge product

In the K->K phase the objects of interest, anomalies, new knowledge concepts that were detected during the C-> K phase can be deployed to the organization. Figure 5.8 shows the business use case of the K->K phase.

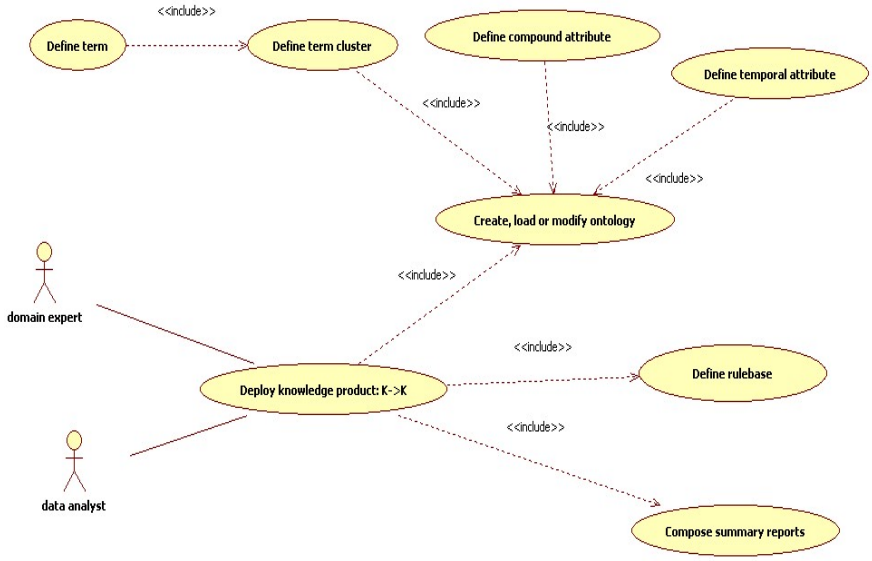


Fig. 5. 8 Deploy knowledge product: K->K

Examples of deploying knowledge concepts are:

- Adjusting an existing rule base by adding new classifier rules. Examples are adding a new rule to the human trafficking rule base to detect general reports in which women in a car do not have an ID-paper with them. This is one of the signals of human trafficking.
- Generate an official document with all detected general reports with signals of human trafficking with respect to one or more suspects and victims and using it to get permission from the public prosecutor to start an investigation after the suspects.
- Generate an official document with all detected general reports with signals of dealing hard drugs in a coffee shop and using it to close down the coffee shop by the council of Amsterdam.

5.7 Data and domain analysis scenarios

In this section the functionality of CORDIET will be explained and demonstrated with one data and two domain analysis scenarios. Section 5.7.1 describes the functionality of CORDIET. Section 5.7.2 demonstrates how CORDIET is applied to construct an ontology for domestic violence from the original definition. Section 5.7.3 demonstrates how CORDIET is applied to find new victims of human trafficking. Section 5.7.4 demonstrates how CORDIET is applied to analyze the workforce intelligence of clinical pathways of breast conserving surgery.

5.7.1 The functionality of the CORDIET toolbox

CORDIET is a multi user system with a stand alone java client environment and two tomcat web applications, one for the ontology and one for the highlighter and the rule base. The CORDIET toolbox is shown in Figure 5.9 and the functionality will be described during the C.K transitions of the data and domain analysis scenarios.

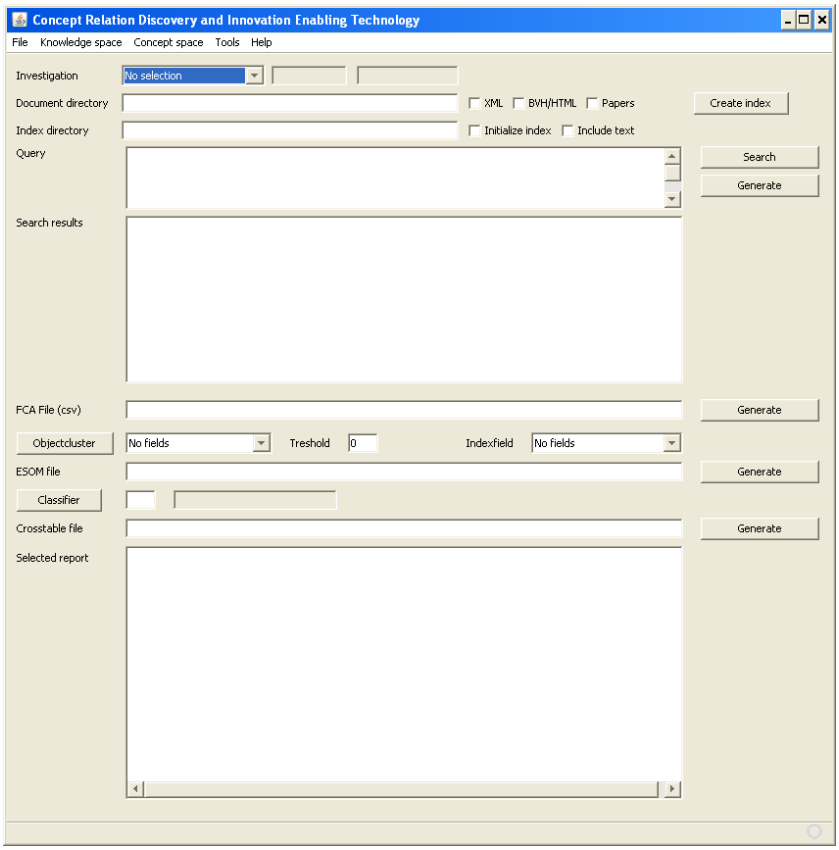


Fig. 5. 9 The CORDIET toolbox

The CORDIET toolbox has three pull down menus.

1. The knowledge space with one K->C and two K->K transitions
2. The concept space with the C->K transitions.
3. The tool menu, modules to export and examine the results of the index and ontology.

The main screen supports the K->C load data source transition and the C->C transitions for generating the input files for the FCA, ESOM and VENN artefacts. The artefacts can be activated by the concept space pull down menu.

5.7.1.1 Knowledge space options

A screenshot of the pull down menu with the knowledge space options is shown in Figure 5.10. The options will be discussed more in detail in the next sections when the various C/K transitions will be showcased.

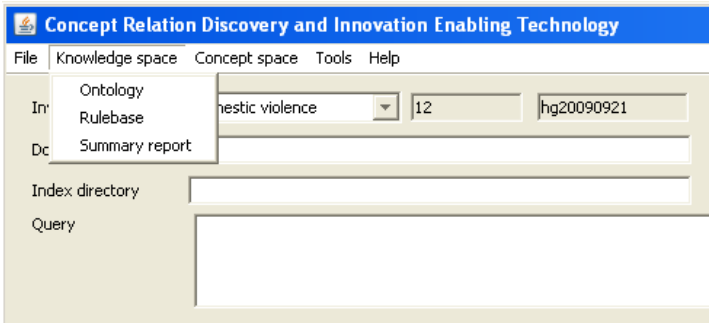


Fig. 5. 10 Pull down menu with the knowledge space options

5.7.1.1.1 Ontology

The ontology option activates a web based application where ontologies can be created and maintained.

5.7.1.1.2 Rule base

The rule base option generates a Prolog file and one input file for the commercial thesaurus application we used in the first version of the CORDIET toolbox and is used by the project “text mining by fingerprints”. The Prolog file consists of all possible predicates, where each text mining and compound attribute is transformed to a predicate and is added to the rule base which is used by the classifier application for detecting domestic violence cases missing a domestic violence label.

5.7.1.1.3 Summary report

This option uses a FCA input file with the filename and file path as object cluster rule to read the documents and generated a three column report with the relevant information for i.e. a 27-construction document.

5.7.1.1.4 Concept space options

Figure 5.11 shows the pull down menu with the concept space options

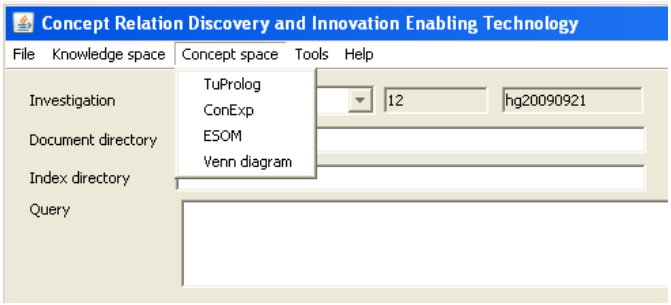


Fig. 5. 11 Pull down menu with concept space options

The pull down options will be described in the next sections.

5.7.1.1.5 TuProlog

This option activates the TuProlog IDE where the Prolog rules are developed and tested. Appendix E shows an example of the tuProlog IDE.

5.7.1.1.6 ConExp

This option activates the ConExp application to analyze the generated FCA input files.

5.7.1.1.7 ESOM

This option activates the ESOM application to train the generated ESOM input files and analyze the toroid maps.

5.7.1.1.8 Venn Diagramm

This option activates the Clustermap application to analyze the generated FCA input files with Venn diagram.

5.7.1.1.9 Tool menu options

Figure 5.12 shows the pull down menu with the tool menu options.

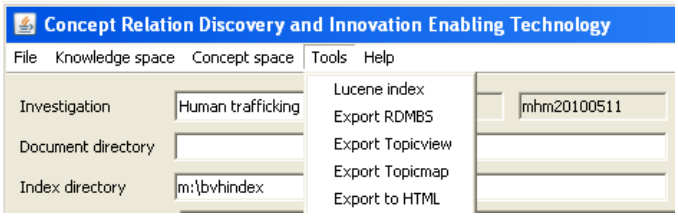


Fig. 5. 12 Pull down menu with tool options

5.7.1.1.10 Lucene index

This option activates the Lucene index toolbox, Luke[12]. Luke is an open source initiative and a handy development and diagnostic tool, which works with Lucene search indexes and allows the user to display and modify their contents in several ways (browse documents, search, delete, insert new, optimize indexes, etc). An example of the Lucene index with a BVH XML dataset is shown in Figure 5.13

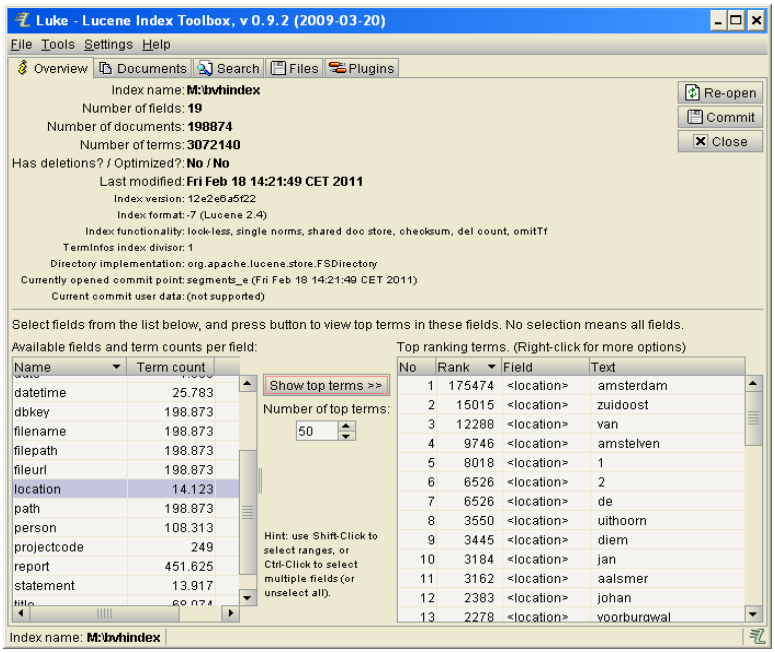


Fig. 5. 13 An example of browsing an index with Luke

[12] http://sourceforge.net/projects/luke/

With Luke it is possible to simulate queries with different language analyzers and get an overview of top terms in the index. The terms “Amsterdam”, “amstelveen”, “uithoorn”, ”diem:” (i.e. Diemen) and “aalsmer” (i.e. Aalsmeer) belong to the most frequent terms of the index and gives an indication of the distribution of the reports over the five communities from the Amsterdam-Amstelland Police Department. Luke also offers the opportunity to repair indexes and commit the changes. This can be useful to delete documents with specific properties, which are responsible for outliers. Instead of using a segmentation rule each time, the documents with the outliers can be deleted with Luke.

5.7.1.1.11 Export RDBMS

The new version of CORDIET, which is jointly under development with the Katholieke Universiteit Leuven and the Moscow Higher School of Economics, will use a PostGreSQL RDMBS to store the ontology and the XML datasets. This option exports the ontologies in a SQL file with insert-statements for the PostGresSQL database. When the new version becomes fully operational, all defined ontologies from this thesis can be reused.

5.7.1.1.12 Export Topicview

This option generates a topic map file based on the topicmap ontology of Topicview. Topicview is a person monitoring system which makes intensive use of the text mining attributes. At this moment the developed ontology of Muslim fundamentalism is fully operational by the terrorism intervention team of the Amsterdam-Amstelland Police Department and will soon be operational for the National Police Service Agency. Topicview is connected to several data sources, where the BVH is one of them. The same reports we used in our investigation of Terrorist Threat Assessment from chapter 4 are automatically imported when a suspect is activated in Topicview. The text mining attributes are generated as hypothesis and offered to the members of the intervention team. The intervention team validates the found textmining attributes of each suspect or possible suspect and accepts or rejects the hypothesis.

5.7.1.1.13 Export Topicmap

This option generates the ontology in a Topic map format which can be explored by web application with a topic map engine. Appendix F shows screenshots with examples of the exported Topicmap from the FCA literature study.

5.7.1.1.14 Export to HTML

For documentation purposes it is necessary to have the ontology in a readable format. This option is also used to generate the excerpts of the thesauri from Appendix A, B and C.

5.7.2 Data analysis scenario “Create an ontology and a rule base for Domestic Violence”

In this section we will show how CORDIET is used to create a domestic violence ontology and a rule base for qualifying domestic violence cases. We will show how the process goes through the various C/K iterations and how the ontology and rule base are constructed.

5.7.2.1 K->C, prepare the datasets and create the ontology

This transition used two options of CORDIET, first the option within the main screen to prepare the datasets and second the pull down option “ontology” of the knowledge space.

5.7.2.1.1 Prepare the datasets

To prepare the dataset the user is offered to enter the directory where the input documents are stored, the type of input document, the directory for the Lucene index and the option to initialize the Lucene index. Figure 5.14 shows an example of loading a XML dataset from BVH

Fig. 5. 14 An example of loading a XML dataset from BVH

CORDIET is designed to read three different file formats. The first format is BVH/HTML. We started the domestic violence investigation with datasets of generated HTML reports from the BVH databaset which are used for the project “text mining with fingerprints” (Elzinga 2006). The structured BVH information, like persons, locations and dates are stored in the header with meta tags. The second format is XML as described in section 5.6.1.1. The XML format has the advantage of flexibility. If an investigation need more structured data, like forensic traces, it can be parsed by CORDIET. The structured data are added as Lucene fields to the Lucene document and can be used in the ontology. This XML format is applied for the datasets of the clinical pathways of the breast cancer patients. The third format is the scientific papers, which are offered in a flat file format. The scientific papers were available in PDF format. To parse the file into the necessary structure of title, authors, abstract with keywords and the contents, the PDF files needed to be converted to flat files first.

The datasets are generated by a parameterized export from the BVH system with the choice to structure the information into HTML or XML format and stored in one or more directories. The CORDIET parses the files and creates the Lucene index. Each file corresponds with one Lucene document and each structured data from the file corresponds with a Lucene field which is stored in the Lucene document. The Lucene index can be used by more than one investigation. The first time when an

index is created, the “initialize index” checkbox is selected. When selecting more than one data source, the “initialize index” checkbox should be deselected. The checkboxes “include text” and “include terms” are optional checkboxes when the Lucene index itself is need to be analyzed. But these options are not needed throughout the various C/K iterations and do have a heavy impact on the performance of the system.

5.7.2.1.2 Create a new ontology.

In this section we will showcase how the domestic violence ontology is stepwise constructed by using the ontology option from the pull down menu. This option will redirect the user to the web application Ontopoly, an open source topic map[13] editor[14].

At the core of the each ontology is the definition of the problem area and in this example the definition of domestic violence employed by the police organization of the Netherlands, which is as follows:

> *“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 (Keus 2000, Van Dijk 1997)”*

Starting from this definition initial text mining attributes can be constructed:

- acts of violence
- partner members
- ex-partner members
- family members
- relative members
- family friend members

It should be noted that a report is always written from the point of view of the victim and not from the point of view of the officer. A victim always adds “my”, “your”, “her” and “his” when referring to the persons involved in the crime. Therefore, the report is searched for terms such as “my dad”, “my mom” and “my son”. These terms are grouped into the compound attribute “family members”. The initial ontology is composed of one termcluster, acts of violence, and five compound attributes, which each is composed of a cartesian product with the termcluster “my-his-her” and the corresponding termcluster, partner, ex-partner, family, relative and family friend. One compound attribute is added to the ontology, labeled as domestic violence. Most of the domestic violence reports with statements of the victim are labeled. This can be used when analyzing the FCA lattices and expanding the

[13] http://www.ontopia.net/section.jsp?id=tm-intro
[14] http://www.ontopia.net/download.jsp

concept space. Figure 5.15 shows the initial ontology with the text mining attributes started from the definition of Keus (2000) and Van Dijk (1997).

Fig. 5. 15 The initial ontology

The act of violence term cluster consists of one or more terms. Each term consists of a list of one or more search terms which is used by querying the reports. Figure 5.16 shows the act of violence term cluster with its terms.

Fig. 5. 16 Termcluster “acts of violence”

Figure 5.17 shows the compound attribute “family members”.

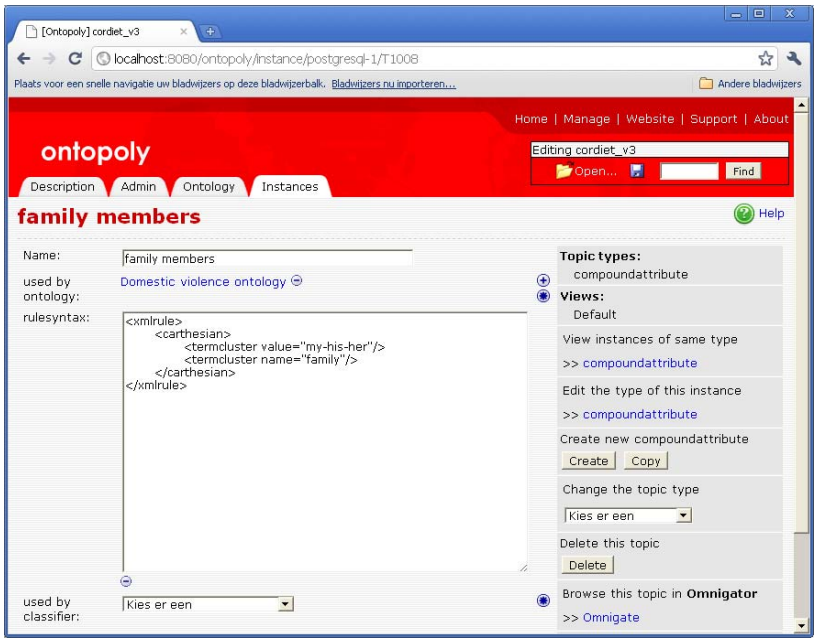


Fig. 5. 17 Compound attribute “family members”.

Figure 5.18 and Figure 5.19 show the two termclusters used by the compound attribute “family members” and Figure 5.20 shows the list of search terms belonging to term “child”.

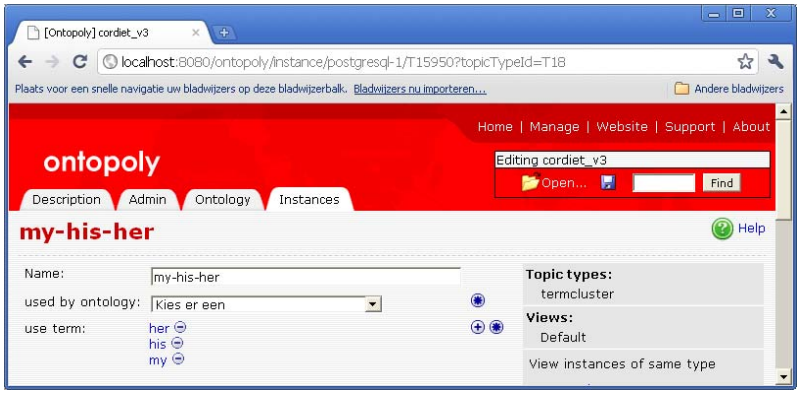


Fig. 5. 18 Termcluster “my-his-her”.

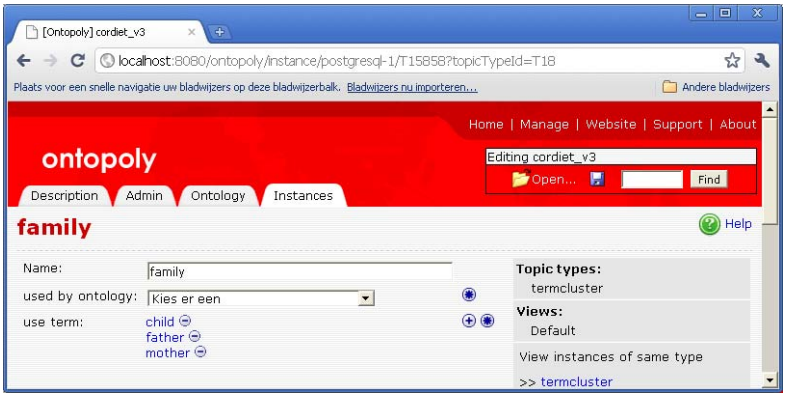


Fig. 5. 19 Termcluster family

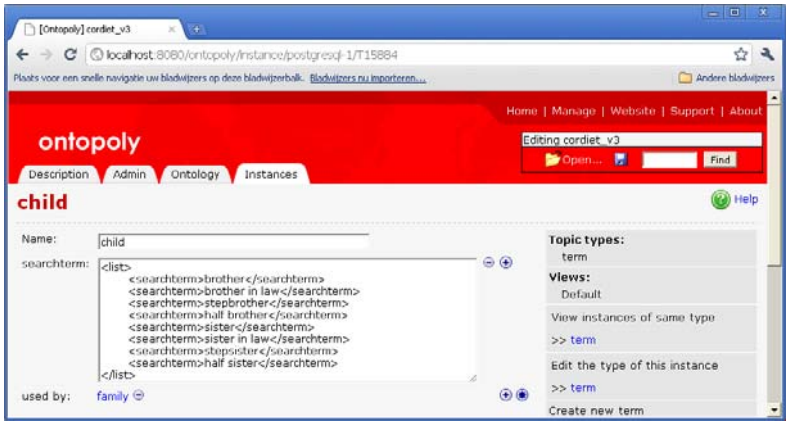


Fig. 5. 20 Term “child”

If a query is executed in CORDIET, all compound attributes are parsed into Lucene queries. The example below is an excerpt of a cartesian product of the compound attributes “my family” with the termclusters “my-his-her” and “child”.

> *("my brother") OR ("my stepbrother") OR ("my half brother") OR ("my brother in law") OR ("his brother") OR ("his stepbrother") OR ("his half brother") OR ("his brother in law") OR ("her brother") OR ("her stepbrother") OR ("her half brother") OR ("her brother in law") OR ("my sister") OR ("my stepsister") OR ("my half sister") OR ("my sister in law") OR ("his sister") OR ("his stepsister") OR ("his half sister") OR ("his sister in law") OR ("her sister") OR ("her stepsister") OR ("her half sister") OR ("her sister in law")*

In the new version of CORDIET, the ontology will be constructed in a user friendly way with a visual editor using drag and drop options to select search terms into text mining attributes and a rule editor with users supported actions, like intersection, union, etc. operators. The rule will be stored in XML. The current toolbox uses a built in XML editor which validates the XML and used termclusters and objects before storing it into to the ontology.

5.7.2.2 C->C: compose artefact

The input files which are needed for the FCA, ESOM and Venn artefacts are generated by selecting the options from the main screen.

5.7.2.2.1 Select the ontology and rules

The artifacts are generated from the main window. When creating a new artifact the user should define the path- and filename and where the artefact input files should be created. The default file format is “csv”, a flat file with separators. This file is used in both generating the FCA lattice as generating a Venn diagram. Both artifacts will be showcased. Next the user should define which object cluster rule and which Lucene index field should be selected. Figure 5.21 shows an example of the creation of a FCA input file.

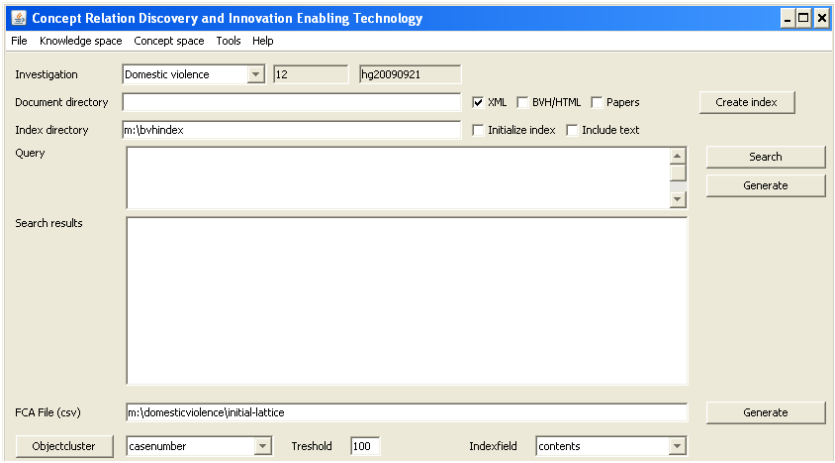


Fig. 5. 21 Create FCA input file

Segmentation rules in the toolbox version are implemented by invoking the built-in, Prolog based, rule base. Text mining attributes are implemented as rules within the rule base. In our case we use one segmentation rule with one compound attribute: “labeled domestic violence”. Assigning the threshold to a non-zero value, the built-in rule base is invoked to evaluate the documents. To generate the FCA input file, an existing Lucene index must be selected, a filename must be entered in the text field “FCA file (csv)”, an object cluster rule and the Lucene index field with the unstructured text must be selected. When all required fields are entered and all

required selections are made, activating the button next to the FCA (file) input field generates the desired FCA input file and is available for the analyzing the results.

5.7.2.3 C->K analyze the artefacts

Activating the rule base by entering a treshold value of 100 has resulted in 1979 reports with statements and labeled as domestic violence. We can investigate the initial ontology with both a Venn diagram and a FCA lattice. We will show that FCA lattices outperform Venn diagrams in comprehensibility when the number of attributes in the diagram and /or lattice increases.

5.7.2.3.1 Analyze the initial results with a Venn diagram

Venn diagram's are very handy when verifying the completeness of the definition working with a small number of text mining attributes. The used Venn diagram software was available as open source tool[15] during our investigations, but is unfortunately only available as a commercial library package now. This tool has a user-friendly interface. By activating the checkboxes in the left panel of the tool; the Venn diagram is automatically drawn. When the number of objects within the intersection is low, the individual objects can be selected and shown by the web based highlighter application. The user can state simple questions like: "do all domestic cases have an act of violence?". Figure 5.22 shows the results of the Venn diagram in activating two checkboxes in the left panel "acts of violence" and "labeled domestic violence".

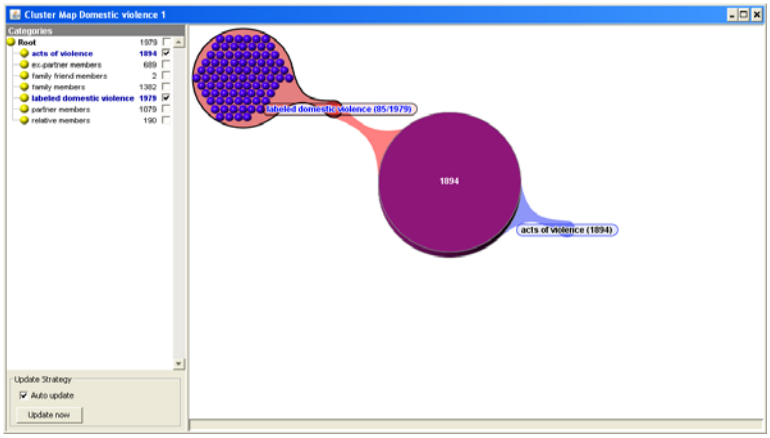


Fig. 5. 22 Venn diagram of the intersection of labeled domestic violence with acts of violence

The example in Figure 5.22 shows 85 domestic violence cases which do not have an act of violence and can be selected and shown by the highlighter. In the same way intersections of members of the domestic sphere can be validated against the labeled domestic violence cases. The Venn diagram gives the user a quick insight in

[15] http://www.aduna-software.com/technology/clustermap

the quality of the ontology, where in this case the definition turns out not complete. But if we want to investigate if there are cases without any persons of the domestic sphere, the Venn diagram becomes very hard to analyze, even with a low number of 7 different attributes. Figure 5.23 shows a Venn diagram with the text mining attributes of the domestic sphere and the labeled domestic violence attribute.

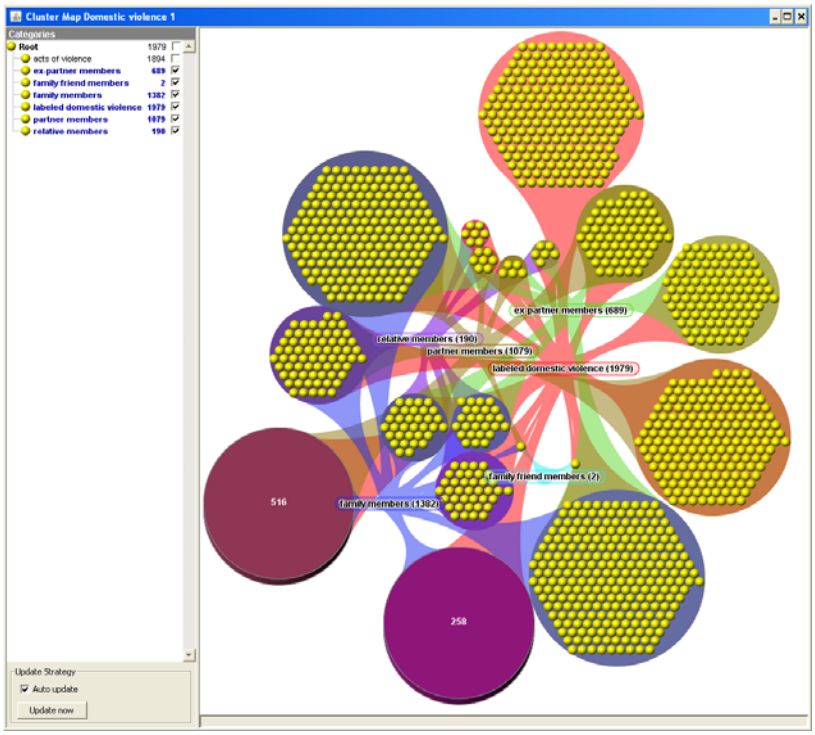


Fig. 5. 23 Venn diagram with intersection of the members of the domestic sphere with labeled domestic violence cases

5.7.2.3.2 Analyze the initial results with FCA lattices

FCA lattices can handle the complex situation with combining objects with a larger number of text mining attributes more effective as we will demonstrate by the next example. The same artefact file is used as input for generating a FCA lattice by Conexp (Yevtushenko 2000) an open source tool. Conexp[16] is integrated in the CORDIET toolbox and used to explore the FCA lattice. We choose the option Conexp from the concept space pull down menu and the FCA lattice is shown in Figure 5.24. In the FCA lattice screen we selected the option to show the own object count, which gives an optimal insight in the gaps of the definition. Figure 5.24 is more comprehensible than Figure 5.23. It is almost impossible to detect the cases

[16] http://conexp.sourceforge.net/index.html

which met none of the text mining attributes in Figure 5.23, as in Figure 5.24 these cases are visible on the top of the lattice.

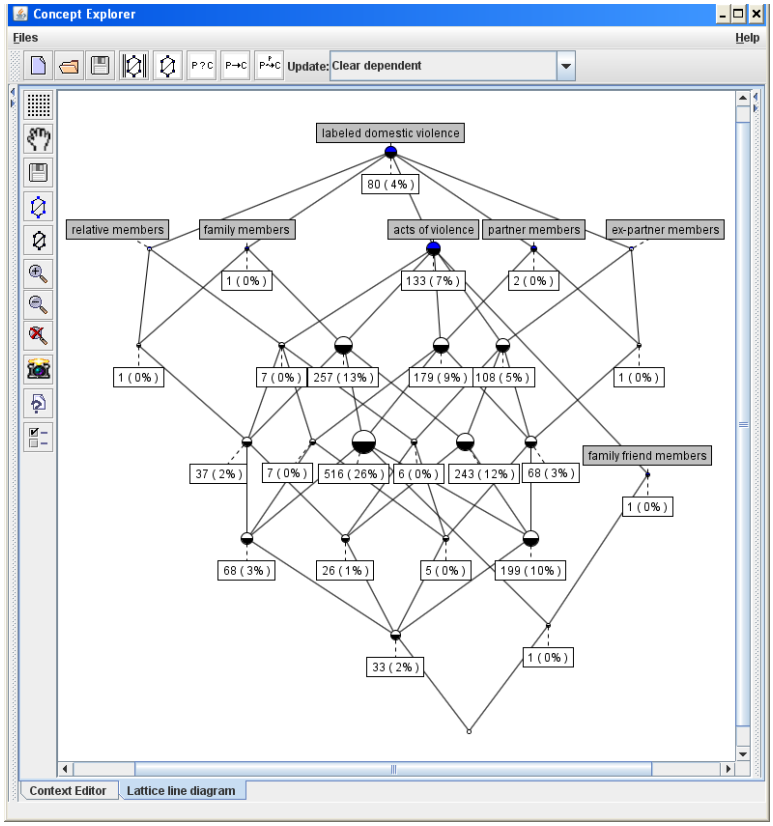


Fig. 5. 24 FCA lattice of the initial ontology with domestic violence cases

The top of the lattice shows 80 labeled domestic violence cases which do not meet the definition as formulated in the beginning of the section. At the same time there are 133 cases of domestic violence with acts of violence, but without members of the domestic sphere.

5.7.2.3.3 Validate the ontology using FCA lattice

Analyzing the lattice from Figure 5.24 shows the following differences which can be analyzed in detail by inspecting the documents:

1. 133 cases of act of violence without mentioning someone of the domestic sphere.
2. 5 cases with no acts of violence but containing a member of the domestic sphere
3. 80 cases with no acts of violence and no members of the domestic sphere.

By activating the object option “show multi labels”, the individual documents can be visualized and the documents can be selected. Figure 5.25 shows an example with two documents of family members with acts of violence. In this example all attributes are deselected except “family members” and “acts of violence”.

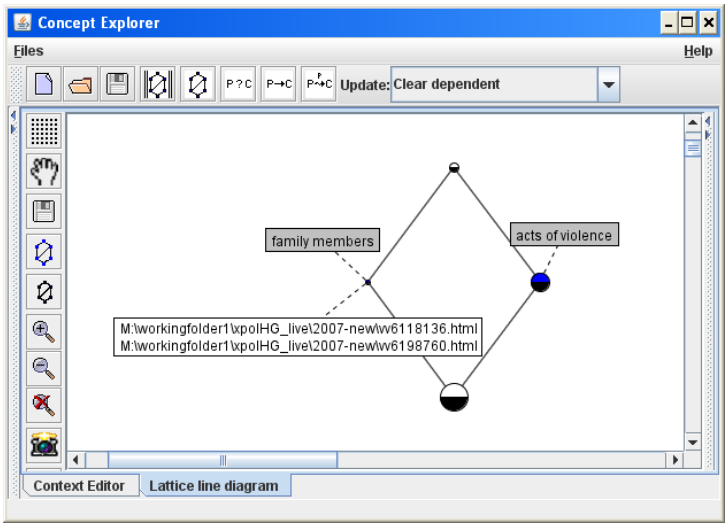


Fig. 5. 25 Showing the documents of family members without an act of violence

Double clicking a document from the lattice with only members of the domestic sphere and no acts of violence executes the highlighter application from CORDIET and shows the documents with the textmining attributes and the corresponding search terms.

Verklaring

Ik kom hier om aangifte te doen tegen mijn nicht mevrouw [redacted] vandiefstal van een armband en een oorbel. Zij heet volledig [redacted], zij is op 6 januari jarig maar ik weet niet hoe oud zij is.Wat is nu het geval.Ik heb in oktober 2006 mijn nicht die op datmoment zonder woon- of verblijfplaats hier ten lande was toegestaanom voor enkele maanden bij mij in de woning te komen wonen.Tot 7 maart 2007 was alles in orde. Op 7 maart 2007 ontdekte ik'savonds dat een oorbel van mij weg was. Deze oorbel is eigendom vanmijn dochter genaamd [redacted] geboren op [redacted]/[redacted]/19[redacted] en nogsteeds thuis wonend.Die oorbel lag op de kaptafel in mijn dochters slaapkamer. Zij hadhem zondag 4 maart 2007 daar neergelegd en gezien. Zij had namelijkdie oorbel toen net schoongemaakt. Toen wij die oorbel mistten zijnwij gaan kijken of er nog meer weg was.Toen bemerkten wij dat ook een brachelet van mijn dochter met daaraandrie muntjes weg was. Wij hebben overal gezocht in de woning maarkonden zowel de brachelet als de oorbel nergens vinden.De brachelet en de oorbel zijn beide van surinaams goud. Van debrachelet heb ik een foto. Van de oorbel niet, die had ik namelijk injanuari 2007 pas gekregen en nog nooit gedragen. Ik had die oorbelaan mijn dochter gegeven.Het was een ronde oorbel en klein van stuk.Ik houd namelijk niet van kleine oorbellen en daarom had ik die aanmijn dochter gegeven.Waarom verdenk ik mijn nicht van deze diefstal, omdat wij met vierpersonen, ik mijn dochter en zoon en mijn nicht, in het huis wonen ende spullen er al jaren, althans die armband en andere sieraden, daarliggen zonder dat zij gestolen

Fig. 5. 26 An excerpt from a statement without an act of violence

The example from Figure 5.26 shows the complexity of interpreting domestic violence cases. In this case a niece of the victim has stolen a golden bracelet; no threatening from the suspect has been made to the victim, no real act of violence has taken place. This had led us to the conclusion this case is wrongly labeled as domestic violence.

5.7.2.4 K->K: deploy new knowledge

After reading 20 more cases, we discover many of them refer to theft, pick pocketing and burglary and all 20 are wrongly classified as domestic violence. This discovery led to a new text mining attribute: ”theft and burglary”. This text mining attribute is also added as a rule to the rule base to recognize wrongly classified domestic violence cases. Another discovery is made by finding wrongly labeled documents referring to missed or stored ID-documents. Another text mining attribute is defined, “missing ID-documents” and another rule to recognize wrongly labeled domestic violence cases.

To store the rules, a new rule base is created and the rules are added to the rule base. In the next C/K iterations the rule base is used in combination with the highlighter application and the results are shown in the browser. Appendix E gives a more detailed description of the rule base application with the highlighter functionality.

The used rule syntax is Prolog and the used Prolog rule engine is tuProlog[17] Text mining attributes are used as predicates and clauses to evaluate the rule base. When a document is matched against the ontology, during the evaluation the detected text mining attributes are added as so-called facts to the knowledge base of Prolog and the rule base is evaluated by backtracking, starting from the first rule in the rule base. Evaluating the labeled domestic violence case from Figure 5.26 with the ontology returns three text mining attributes. “theft and burglary”, “family members” and “labeled domestic violence”. Adding these attributes as facts to the rule base will return a truth value after evaluating the first rule. Figure 5.28 of section 5.7.2.5 gives another example of a highlighted, wrongly labeled domestic violence case where the text mining attributes, the applied rule and the search terms are highlighted in the document.

Although the definition of domestic violence refers to the different members of domestic violence, it turns out to be more efficient to group all members into one compound attribute, “domestic sphere members”. This reduces the number of attributes and which makes the lattices more readable.

5.7.2.5 Start a new C/K iteration

With the new text mining attributes and the new rules, a new C/K iteration can be started and the result is shown in the next lattice.

[17] http://alice.unibo.it/xwiki/bin/view/Tuprolog/

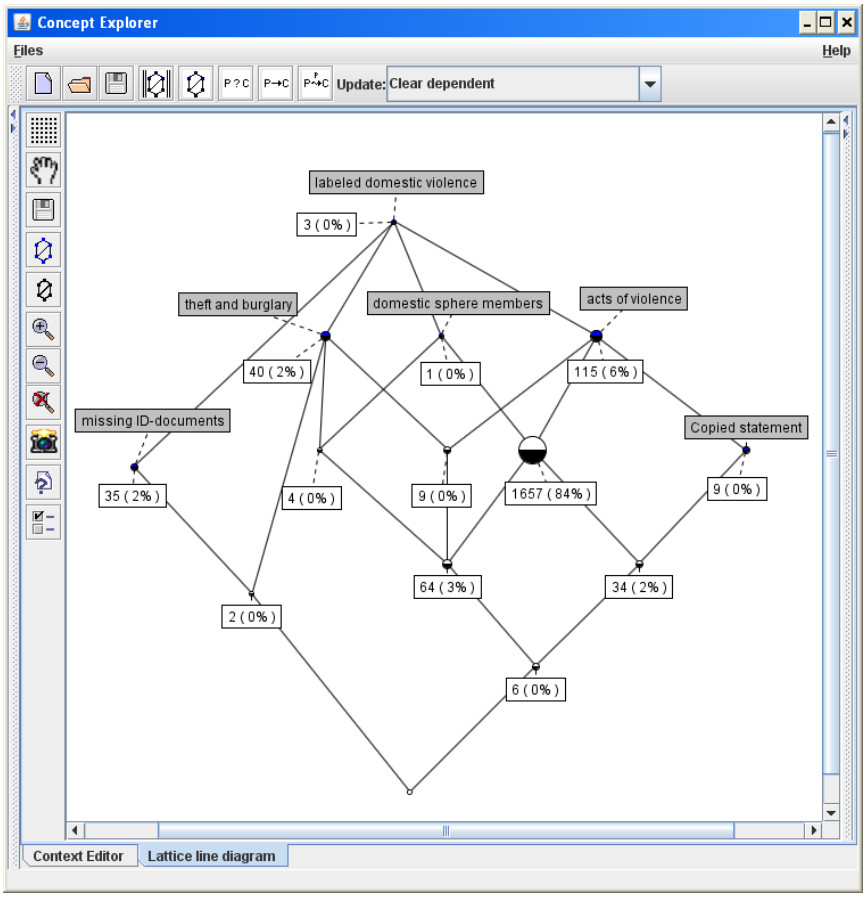


Fig. 5. 27 Second lattice with new ontology elements

The lattice of Figure 5.27 show we have reduced the number of 133 cases labeled as domestic violence and containing a term from the acts of violence cluster without mentioning a member of the domestic sphere to 115 cases. By adding new terms based on the knowledge obtained about a missing ID document and/or the knowledge of a theft or burglary in the cases we found. These cases are wrongly classified and this now visible in the lattice. Figure 5.28 shows an example of a wrongly labeled domestic violence case.

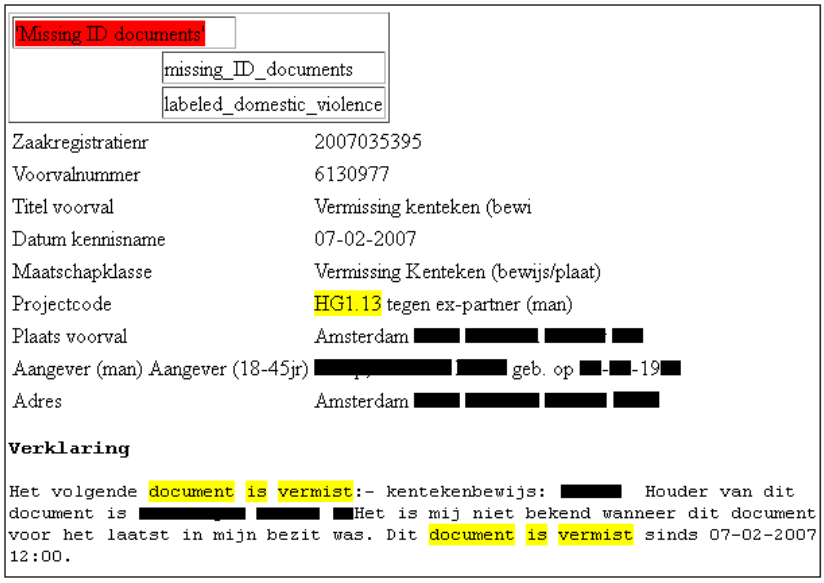


Fig. 5. 28 A wrongly classified domestic violence case

The table in the header of Figure 5.28 presents left the applied rules and right the applied text mining attributes. The color is red, defined in the rule, which alerts the user to a fault. The rule base is used in a standalone application as quality instrument, to recognize both faulty labeled cases and missing labeled cases (Elzinga et al 2009). Appendix E gives a description of the stand alone classification application with examples.

5.7.2.6 Validate the ontology using ESOM toroid map

Although the C/K iterations with FCA can help the user to find new terms, termclusters and compound attributes, the FCA lattice in Figure 5.29 shows that the found text mining attributes are not complete for recognizing domestic violence.

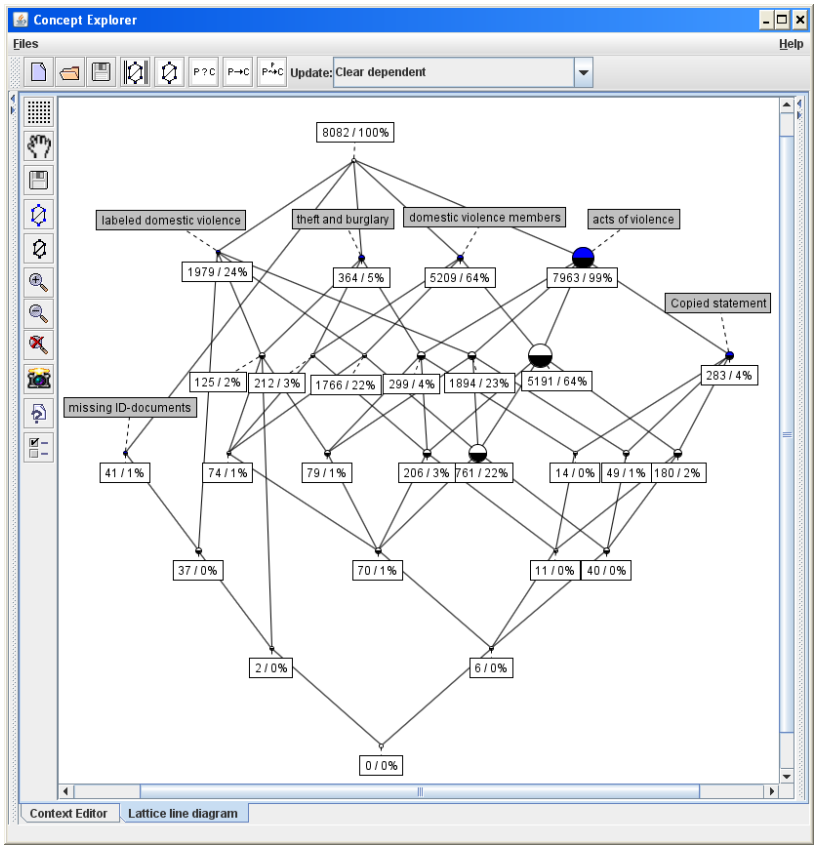


Fig. 5. 29 FCA lattice of all acts of violence of 2007

To visualize how the combination of “domestic sphere members” and “acts of violence” can be used to classify cases as domestic violence, we can make a selection of three attributes from the lattice by deselecting the remaining attributes from the lattice. The result is shown in Figure 5.30.

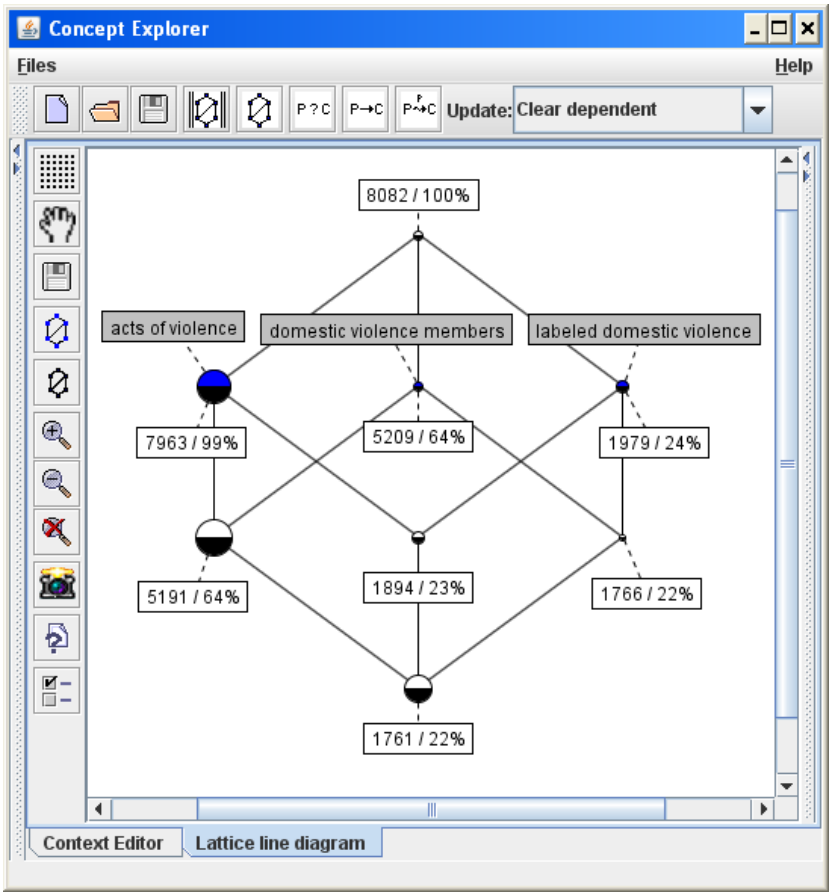


Fig. 5. 30 Lattice with selected concepts

The FCA lattice from Figure 5.30 shows that 1761 cases out of 1979 do have the two attributes, domestic violence members and acts of violence. What does give a good translation of the definition into text mining attributes, but at the same time there are 5191 (64%) out of 8082 cases with acts of violence and domestic sphere members which are not labeled as domestic violence. Conclusion is that the definition of domestic violence is not sufficient for classifying purposes. We need more additional text mining attributes and will demonstrate how ESOM, Emerging Self Organizing Maps, can be used as a catalyst to find new text mining attributes.

5.7.2.7 C->C: compose the ESOM input files

In CORDIET the name of the ESOM input files are filled in and generated by pushing the button next to the “ESOM file” field. The threshold parameter is reset to zero and the classifier rule is filled in. Figure 5.31 shows the CORDIET screen with the fields that were entered.

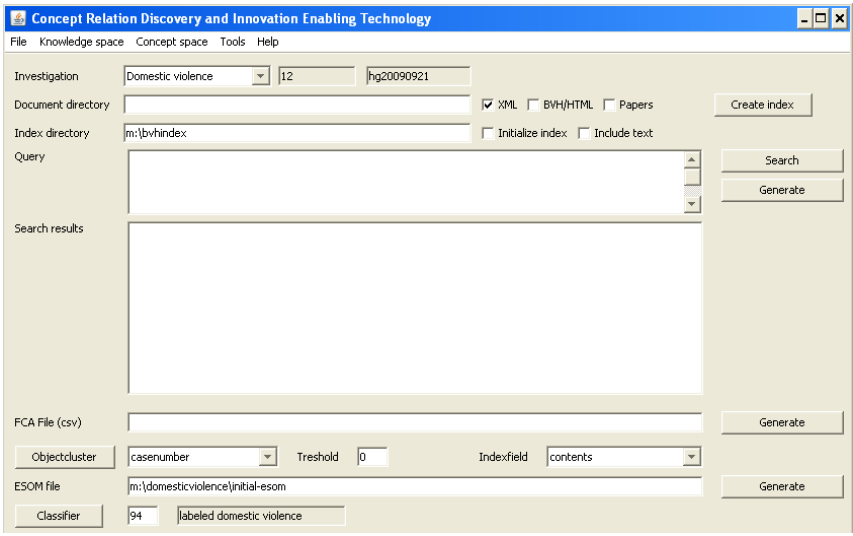


Fig. 5. 31 Compose the ESOM input files

The ESOM tool needs two input files, a cross table with documents as objects and the terms as attributes and a classification file where each document is classified if it has the value of labeled as domestic violence. To generate the ESOM input files, an existing Lucene index must be selected, a filename must be entered in the text field “ESOM file”, the object cluster rule “casenumber” and the Lucene index field with the unstructured text must be selected. For the classification file a classifier needs to be entered, in our case it is the ID of the text mining attribute “labeled domestic violence”. When all fields are entered and all selections are made, activating the button next to the ESOM file input field generates the desired ESOM input files and is available for the analyzing the results.

The initial cross table contains 250 attributes and 8082 objects. Appendix I shows examples of both the cross table and the corresponding classification table. After the generation of the input files is finished, the pull down menu option “ESOM” is activated. The ESOM tool appears with the data files loaded. We start the training with selecting the option “training” from the ESOM application.

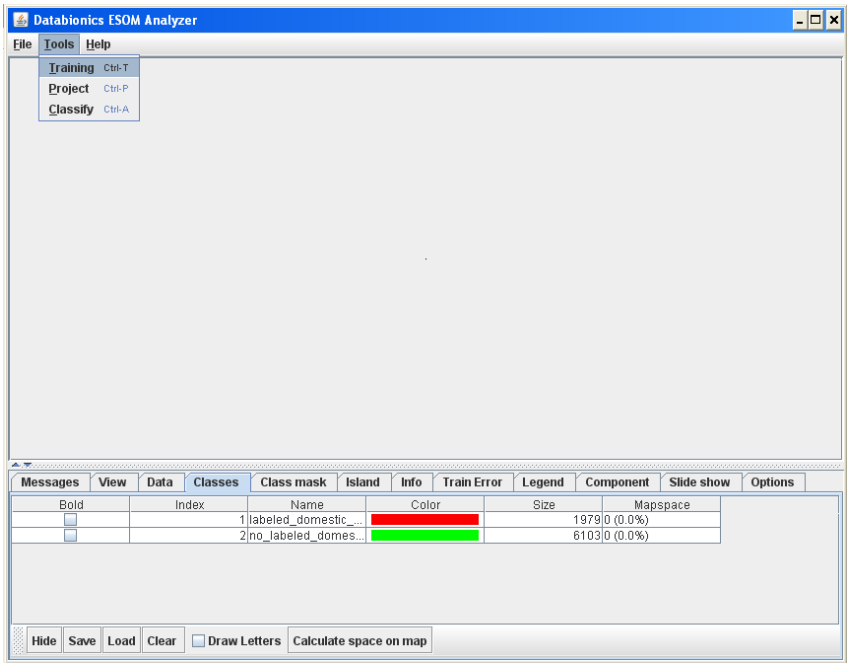


Fig. 5. 32 Start the training of the ESOM toroid map

Figure 5.32 shows the screen of the ESOM tool. Both the data cross table and the classification table are loaded into ESOM. The classes-tab of ESOM shows 1979 labeled domestic violence cases and 6103 unlabelled domestic violence cases. To generate the ESOM toroid map, we choose the option “training”. We use the default training parameters of ESOM which results in a map with 50 rows and 82 columns. One of the standard parameters is the search method, the standard best match method. We have chosen for this option, because an exhaustive search over the whole map is needed. The result of the training is a map with best matches. The cases (or objects) are clustered close to the best match and when selecting neurons in the map with the data tab activated, the corresponding objects will be shown.

5.7.2.8 C->C: Analyze the results of the ESOM map

The result of the training is shown in Figure 5.33. Each red dot is a best match with labeled domestic violence cases and each green dot is a best match for unlabeled domestic violence cases. We are interested in outliers within this map and are looking for a cluster of red dots within a cluster of green dots or vice versa.

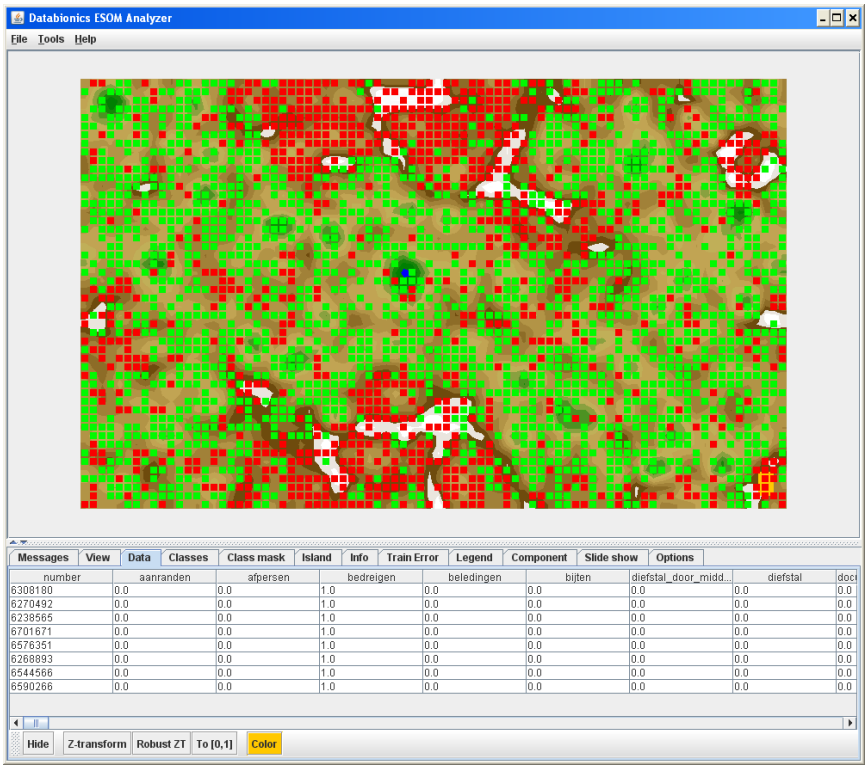


Fig. 5. 33 ESOM map based on the initial definition of domestic violence

Figure 5.33 shows in the lower right corner a group of red dots surrounded by green dots. Analysis of these reports might give us indications of new term clusters, terms or search terms. We make a selection of the red dots. The selected area in the ESOM map is colored yellow and the result with case numbers and attributes is presented in the data tab of ESOM. Each case number is investigated in detail and has led to three new text mining attributes which will be discussed in the next section:

1. legal proceedings against someone of the domestic sphere.
2. acts of violence committed by someone of the domestic sphere
3. relationships

5.7.2.9 K->K and K->C: update the ontology

The new text mining attributes gains new insight in the domestic violence area and results in an update of the ontology.

5.7.2.10 C->C and C->K: compose new FCA input files and analyze the FCA lattices

Using a new filename for the FCA input file with a treshold value of 0 (we want to select all cases), a new FCA input file is generated and the ConExp application is started. Figure 5.34 shows the result of adding the new text mining attributes to the ontology.

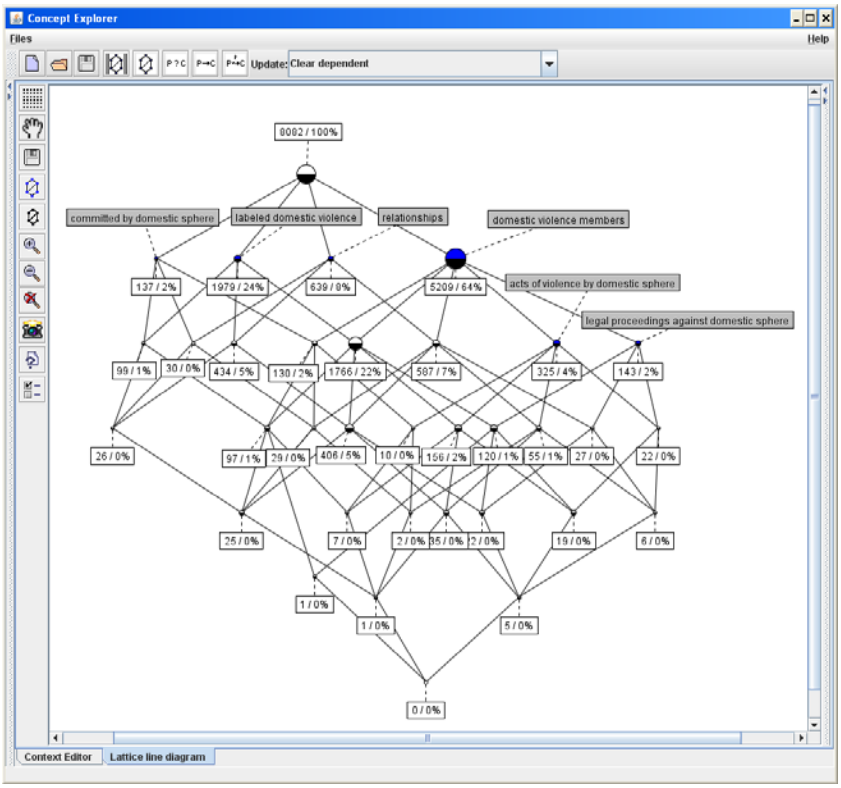


Fig. 5. 34 Result of adding new attributes to the ontology

The lattice from Figure 5.34 shows 143 cases of “legal proceedings against domestic sphere” of which 23 are not labeled as such. An in-depth investigation of the 23 cases revealed, these cases should be labeled as domestic violence. The same applies for “committed by domestic sphere”. Of the 137 cases, 38 were investigated and all cases should be labeled as domestic violence. To show how the FCA lattice helps the user to visualize these cases, Figure 5.35 shows the “committed by domestic sphere” example by deselecting all attributes and the option “show own objects”.

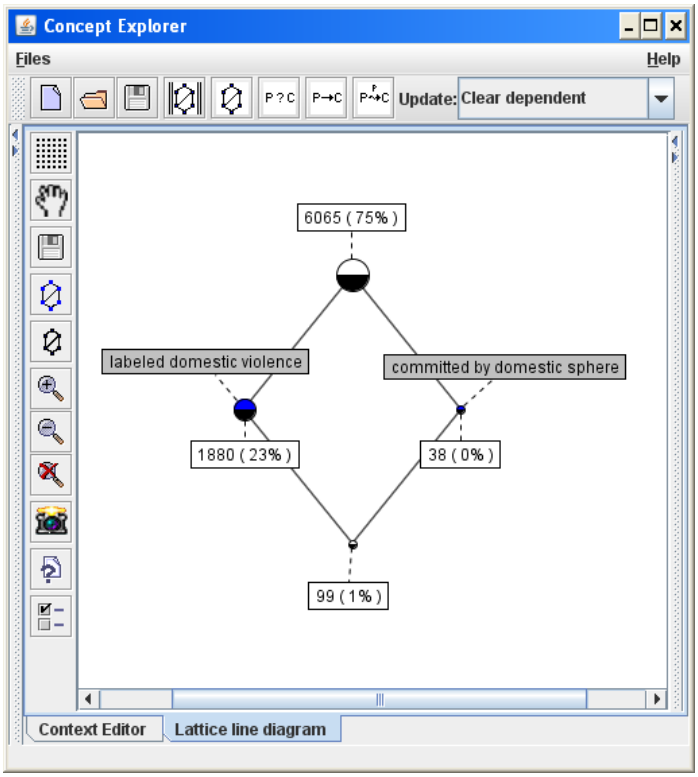


Fig. 5. 35 Showcasing reduced lattice with option “show own objects”

The next step in this lattice is to visualize the 38 cases of Figure 5.35. The option “show multi label” is selected. The case numbers become visible and can be investigated in detail.

5.7.2.11 K->K: deploy new knowledge

We discovered that when the two text mining attributes, “legal proceedings against domestic sphere” and “committed by domestic sphere” are detected in a document, this document always turned out to be domestic violence and we decided to add these attributes as rules to the rule base to detect unlabeled domestic violence cases with a accuracy of 100%.

5.7.3 Domain analysis of human trafficking.

In chapter 4 several cases of human trafficking were discussed. In this section the discovery of a new loverboy suspect case using CORDIET will be demonstrated. As described in chapter 4 there are five main groups of signals which might give indications of human trafficking. For a detailed overview of the signals, see chapter 4. In this section we propose a four step investigation process. First indentify all possible unknown suspects and victims. Second investigate the suspect or victim in

detail. Third validate the connected attributes to the person and fourth, collect all documents with all validated information of the person. In this chapter we will create a new ontology based on a set of the found attributes from chapter 4.

5.7.3.1 Identify possible suspects and or victims

In this chapter we will demonstrate how the CORDIET toolbox helps the domain analyst to find new unknown suspects and/or victims of human trafficking. First we introduce a new ontology where the human trafficking indicators are clustered into a small group of compound attributes. This ontology we will use to detect new suspects and / or victims. To profile the detected suspect or victims, we use the original ontology with all detailed human trafficking attributes. The reason why we have chosen for this two step process is that the number of possible concepts of a FCA lattice can become too large. Using 266,157 general reports and all human trafficking attributes will result in 26,995 concepts. Calculating the resulting lattice does not only require huge CPU capacity but also delivers an unreadable lattice.

One of the real life human trafficking suspects we detected in chapter 4 has used the rule base with a treshold value to select only suspects or victims from the former central and eastern European region who were regularly seen in the red light district. This has reduced the number of possible objects and the number of associated reports. In this chapter we will propose a new method how we detect new possible suspects and victims out of all 266,157 general reports based on the idea of clustering the attributes from chapter 4 into a smaller number of compound attributes. We will name this the signals ontology

5.7.3.1.1 K->C: Create the signals ontology

The first step is to create a new ontology based on clustering the 37 defined human trafficking signals into 5 compound attributes and two termclusters. The compound attributes are constructed based on the guidelines of the attorney generals. One adjustment is made with the guideline “working under bad circumstances”. One of the termclusters in this guideline is composed of extracting all persons with relevant antecedents. This termcluster has been split up into two new termclusters: ‘known suspects” and “known victims”. These termclusters are automatically generated by importing all unique named persons with human trafficking antecedents directly from the BVH database and loaded into the termclusters of the ontology. The result of the new ontology is shown in Figure 5.36.

Fig. 5. 36 Human trafficking ontology with signals

5.7.3.1.2 C->C: compose the FCA lattices

We want to generate a lattice with persons as objects. In this situation we want to cluster the reports by the persons involved and select from the object cluster rule “person”. A name for the FCA lattice is entered and the button “generate” next to the text field of “FCA file (csv)” is submitted. The result is a cross table with persons as objects and the ontology elements as attributes. After the file is generated, the lattice can be generated by activating Conexp from the concept space pull down menu. The lattice is ready to be analyzed.

5.7.3.1.3 C->K: analyze the FCA lattices

The result of generating the FCA lattice is shown in Figure 5.37. In the lattice of Figure 5.37 all attributes and all persons and the “show object count” option is selected.

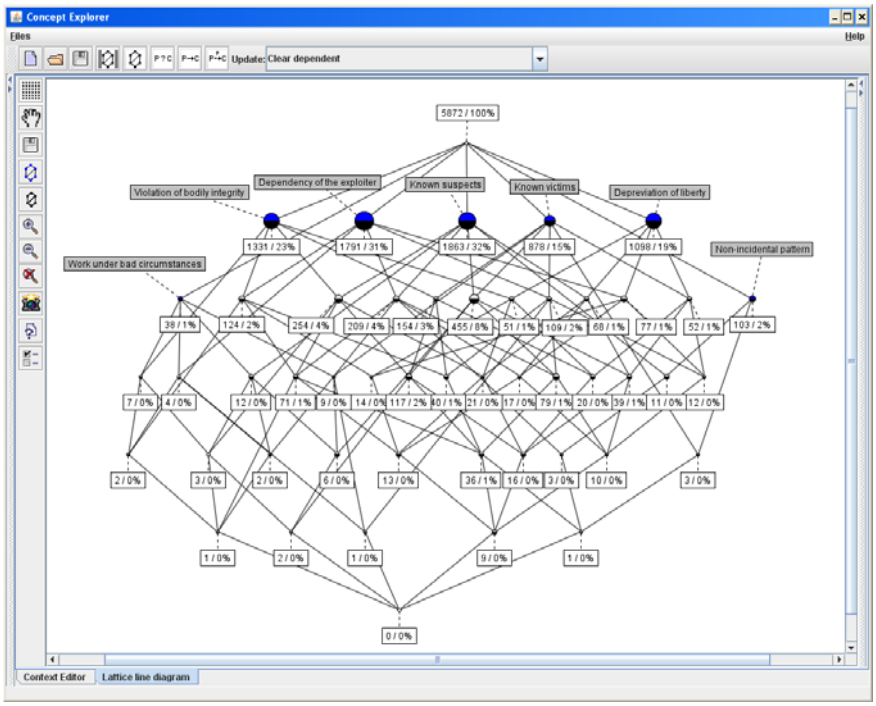


Fig. 5. 37 Lattice with ontology of human trafficking signals

The lattice in Figure 5.37 shows the influence of two top concepts within the lattice: “known suspects” and “known victims”. Out of 5872 persons, who meet at least one of the six human trafficking indications, we find 1862 suspects and 878 victims. This is the result of proactively reporting of police officers because those persons are known to them. In our case we are interested in possible new suspects or victims. The left most nodes take our attention at first sight. To visualize those two nodes more in detail, the option “show own objects count” is selected and three attributes are deselected: “dependency of the exploiter”, “Deprivation of liberty” and “Non-incidental patterns. Figure 5.38 shows the resulting lattice.

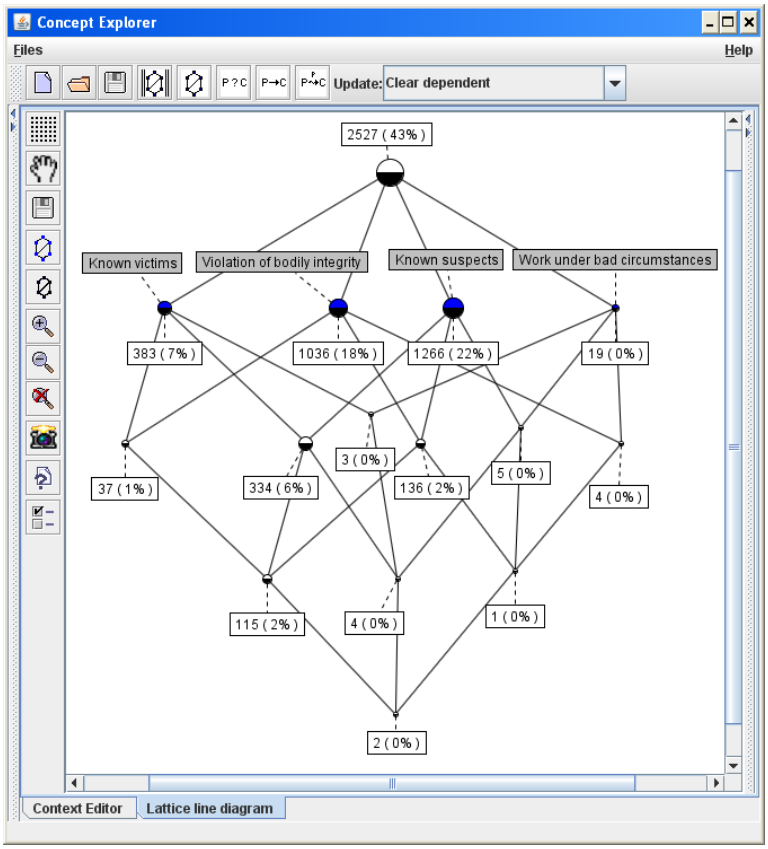


Fig. 5. 38 Lattice with “show own objects count” option

Figure 5.38 shows two interesting concepts on the right:

- One concept with four persons, who do not have relationships with either a known suspect or a known victim and meet at least two attributes
- One concept with one person who is unknown, meets at least two attributes, but has a relationship with a known suspect.

Figure 5.39 shows the same lattice with the “show multi labels” option. The names of the persons become visible.

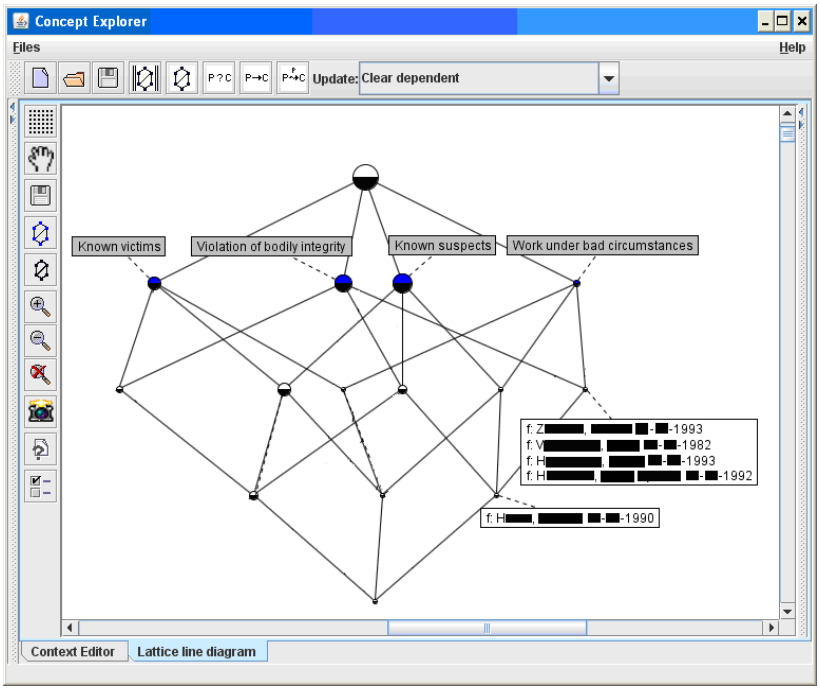


Fig. 5. 39 Lattice with show labeled objects.

All five detected persons from Figure 5.39 are young women in the age from 15 to 18 years (observation dates were in 2008) old. The women have Dutch origins and combined with their age, this might give an indication of a loverboy situation.

The 5 persons are investigated with CORDIET by using the query functionality. The used query syntax is Lucene. One of the Lucene index fields is “person” and using this field, we can query the person very accurately. The query results are shown in the search results text area. From this text area the case numbers can be double clicked and the highlighter method of CORDIET is activated. The corresponding document is read from the directory, the search terms are highlighted and the result is shown in the selected report text area.

Querying the first of the 4 women shows one report where all four women were involved in a fight on a fair. In this case the indications gave false positives, because of the names they called after each other and the threats they made were put in the wrong context. The last person gives us serious indications of a loverboy situation. The query gives us three documents from November and December 2008. Figure 5.40 shows the query, the results of the query and the first document with highlighted relevant terms in the selected report text area.

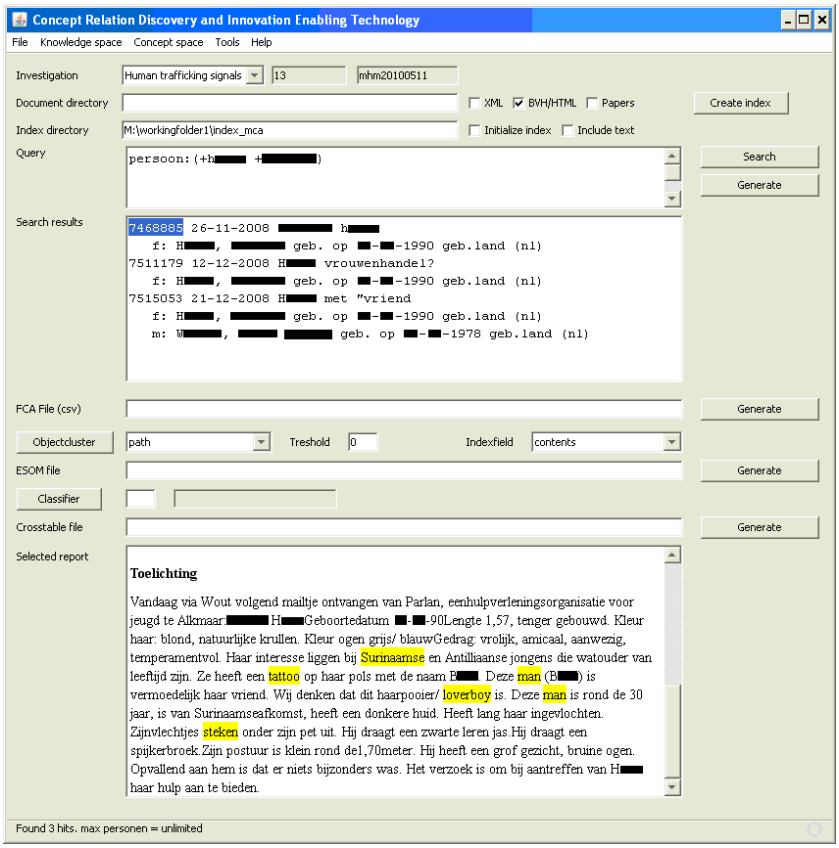
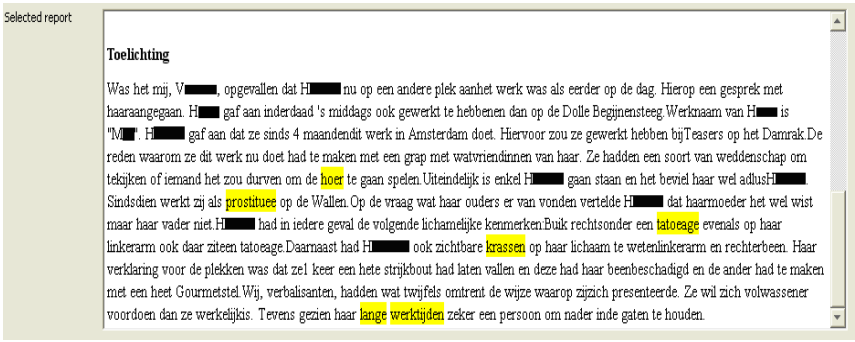


Fig. 5. 40 Possible loverboy victim

Selecting the first report (November 26th 2008) in the search results text area gives a report with a suspicious loverboy situation signaled to the police from a youth aid organization in Alkmaar. In this report the youth aid organization does only have a presumption of a loverboy situation of a man B which first name has been tattooed on her wrist. A detailed description of the person is made.

The second document (December 12th 2008) is a report of an observation made by a police officer who works in the red light district. He observes H. working as a prostitute behind the window and he makes a chat with her. The observed suspicious facts are reported in a general report. The result is shown in Figure 5.41.

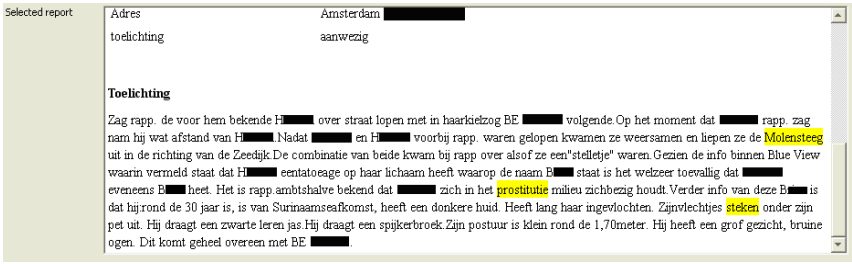

Selected report

Toelichting

Was het mij, V[redacted], opgevallen dat H[redacted] nu op een andere plek aanhet werk was als eerder op de dag. Hierop een gesprek met haaraangegaan. H[redacted] gaf aan inderdaad 's middags ook gewerkt te hebbenen dan op de Dolle Begijnensteeg.Werknaam van H[redacted] is "M[redacted]". H[redacted] gaf aan dat ze sinds 4 maandendit werk in Amsterdam doet. Hiervoor zou ze gewerkt hebben bijTeasers op het Damrak.De reden waarom ze dit werk nu doet had te maken met een grap met watvriendinnen van haar. Ze hadden een soort van weddenschap om tekijken of iemand het zou durven om de hoer te gaan spelen.Uiteindelijk is enkel H[redacted] gaan staan en het beviel haar wel adlusH[redacted] Sindsdien werkt zij als prostituee op de Wallen.Op de vraag wat haar ouders er van vonden vertelde H[redacted] dat haarmoeder het wel wist maar haar vader niet.H[redacted] had in iedere geval de volgende lichamelijke kenmerken:Buik rechtsonder een tatoeage evenals op haar linkerarm ook daar ziteen tatoeage.Daarnaast had H[redacted] ook zichtbare krassen op haar lichaam te wetenlinkerarm en rechterbeen. Haar verklaring voor de plekken was dat ze1 keer een hete strijkbout had laten vallen en deze had haar beenbeschadigd en de ander had te maken met een heet Gourmetstel.Wij, verbalisanten, hadden wat twijfels omtrent de wijze waarop zijzich presenteerde. Ze wil zich volwassener voordoen dan ze werkelijkis. Tevens gezien haar lange werktijden zeker een persoon om nader inde gaten te houden.

Fig. 5. 41 Observation of abnormal injuries and long working days

The report from Figure 5.41 shows four suspicious facts reported by the police officer. First an unbelievable story why she works as a prostitute: a bet between girl friends if someone dares to work as a prostitute. Second the tattoos of which one tattoo is mentioned in the document of Figure 5.40 and a new one on her belly. Third the injuries, she has scratches on her arm (possible from a fight) and burns on her leg. According to the victim, she has dropped a hot iron on her leg and had an accident with a gourmet set. Fourth is the observation of making long working days. The third document in Figure 5.42 (December 21st 2008) shows an observation of the victim walking with the possible suspect.


Selected report

Adres Amsterdam [redacted]

toelichting aanwezig

Toelichting

Zag rapp. de voor hem bekende H[redacted] over straat lopen met in haarkielzog BE [redacted] volgende.Op het moment dat [redacted] rapp. zag nam hij wat afstand van H[redacted] Nadat [redacted] en H[redacted] voorbij rapp. waren gelopen kwamen ze weersamen en liepen ze de Molensteeg uit in de richting van de Zeedijk.De combinatie van beide kwam bij rapp over alsof ze een"stelletje" waren.Gezien de info binnen Blue View waarin vermeld staat dat H[redacted] eentatoeage op haar lichaam heeft waarop de naam B[redacted] staat is het welzeer toevallig dat [redacted] eveneens B[redacted] heet. Het is rapp.ambtshalve bekend dat [redacted] zich in het prostitutie milieu zichbezig houdt.Verder info van deze B[redacted] is dat hij:rond de 30 jaar is, is van Surinaamseafkomst, heeft een donkere huid. Heeft lang haar ingevlochten. Zijnvlechtjes steken onder zijn pet uit. Hij draagt een zwarte leren jas.Hij draagt een spijkerbroek.Zijn postuur is klein rond de 1,70meter. Hij heeft een grof gezicht, bruine ogen. Dit komt geheel overeen met BE [redacted].

Fig. 5. 42 Possible loverboy suspect

In the report from Figure 5.42 the police officer reports he saw the victim and a man walking close to each other. The police officer knows the man as being active in the world of prostitution. When this man saw the officer, he immediately took some distance of the victim. As soon they have passed the officer, they walk close together and into a well known street where prostitutes work behind the windows. The first name of the person B. is the same name which is tattooed on the victim's wrist, and the description of the person is about the same as described by the youth aid organization from Alkmaar.

The three reports together give serious presumptions of B. being a loverboy and H. being the victim. The next step is investigating B. by query the person in CORDIET. We need more serious loverboy signals for B.. Querying B. with CORDIET gives 12 general reports. Investigating these documents shows he frequently visits the red light district and has strong relationships with other pimps.

One the other pimps is the suspect from the loverboy case from chapter 4. We used the detailed ontology with the human trafficking signals from chapter 4. Figure 5.43 shows the lattice with the profile of B.

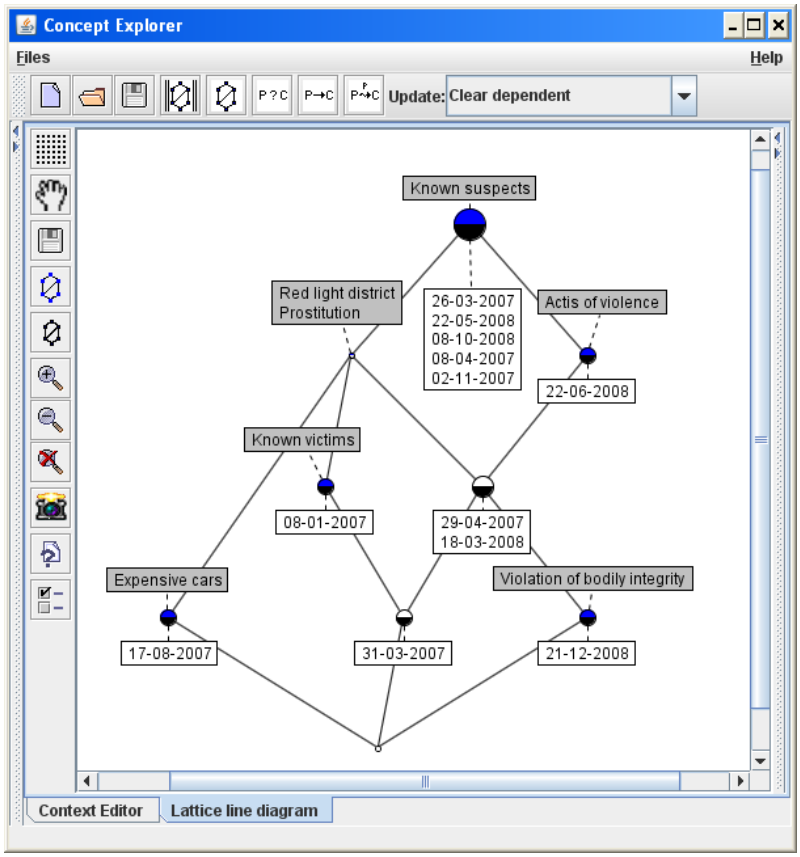


Fig. 5. 43 Lattice of loverboy suspect B.

Figure 5.43 shows six observations where B. is seen in the red light district. From the six observations, four are violence related, including the suspicious burn wounds of the victim H (observation date 21-12-2008). The other violence related observations are situations of fights with customers who are unwilling to leave or to pay. Those violence observations are related to pimps who want to protect there prostitutes. In the Netherlands prostitution is legal, so each prostitute has the right to ask the police to protect her. The violence observations against the customers strengthen our suspicion of B. being the pimp of H. and combined with the burn wounds we have enough indications for making a 27 –construction report.

5.7.3.1.4 K->K: Creating a 27-construction report

Using the 12 general reports a summary report can be created. Not all 12 documents do have significant signals. The summary report will be used by the police investigator to select the relevant documents and turn it into an official 27-construction document.

To construct the summary report, we first have to create a cross table with the relevant reports. We use the FCA cross table file layout to select the reports. A query is formulated in the query text area with the person Brian as query. Path is selected as object cluster, because we need the names of the reports. Pushing the generate button below the search button generates the cross table based on the query. The cross table is ready for generating the summary report. Figure 5.44 shows the selections made in de toolbox and executing the pull down option summary report.

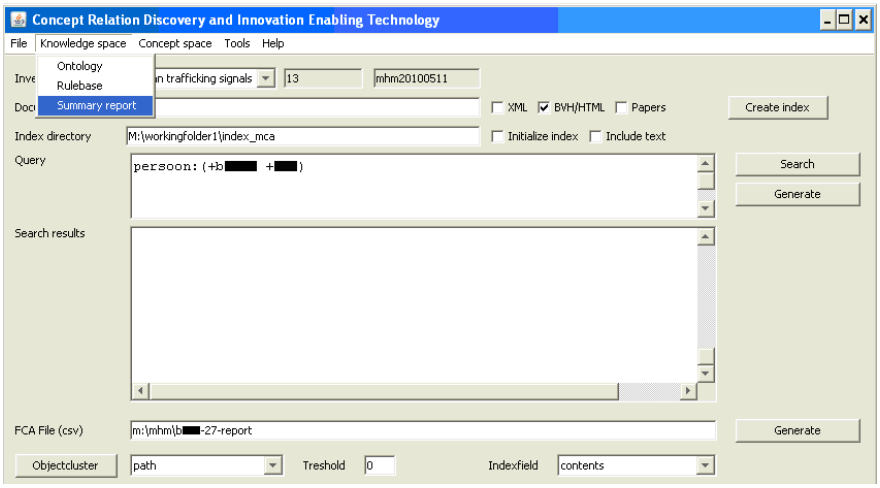
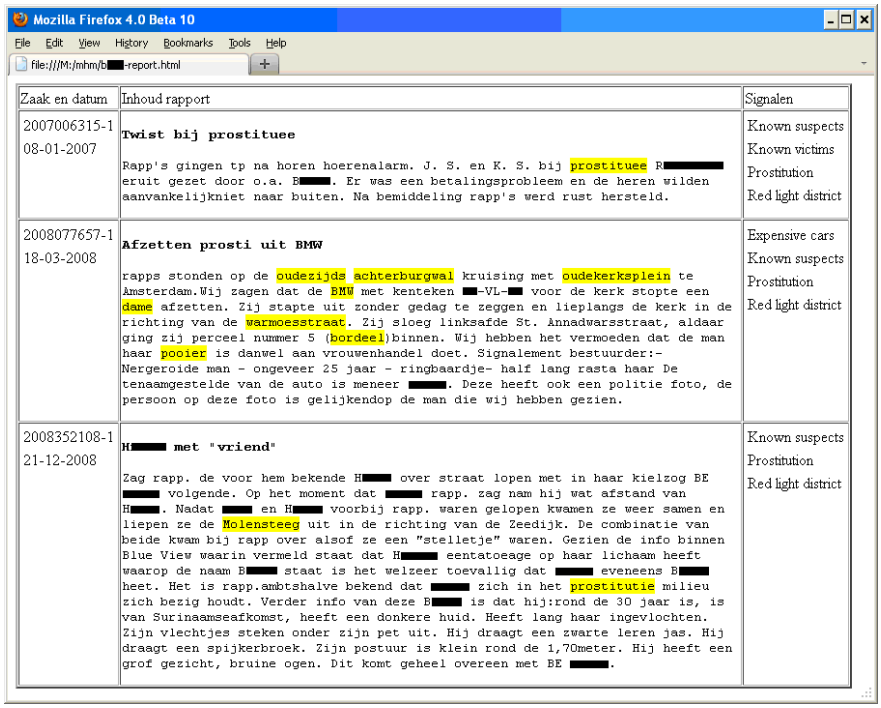


Fig. 5. 44 Create a summary report of suspect B.

After the cross table has been generated, the menu option of the summary report can be activated. The cross table will be read by CORDIET and all reports are transformed into a HTML document with a table with three columns. The first column consists of the case number and date of the observation, the second column consists of the unstructured part of the observation with highlighted terms of the human trafficking signals and the third column consists of the found human trafficking signals or text mining attributes. When the summary report is generated, CORDIET will open a web browser and show the report. Figure 5.45 shows an excerpt of the result.



Zaak en datum	Inhoud rapport	Signalen
2007006315-1 08-01-2007	**Twist bij prostituee** Rapp's gingen tp na horen hoerenalarm. J. S. en K. S. bij prostituee R[redacted] eruit gezet door o.a. B[redacted]. Er was een betalingsprobleem en de heren wilden aanvankelijkniet naar buiten. Na bemiddeling rapp's werd rust hersteld.	Known suspects Known victims Prostitution Red light district
2008077657-1 18-03-2008	**Afzetten prosti uit BMW** rapps stonden op de oudezijds achterburgwal kruising met oudekerksplein te Amsterdam.Wij zagen dat de BMW met kenteken [redacted]-VL-[redacted] voor de kerk stopte een dame afzetten. Zij stapte uit zonder gedag te zeggen en lieplangs de kerk in de richting van de warmoesstraat. Zij sloeg linksafde St. Annadwarsstraat, aldaar ging zij perceel nummer 5 (bordeel)binnen. Wij hebben het vermoeden dat de man haar pooier is danwel aan vrouwenhandel doet. Signalement bestuurder:- Nergeroide man - ongeveer 25 jaar - ringbaardje- half lang rasta haar De tenaamgestelde van de auto is meneer [redacted]. Deze heeft ook een politie foto, de persoon op deze foto is gelijkendop de man die wij hebben gezien.	Expensive cars Known suspects Prostitution Red light district
2008352108-1 21-12-2008	**H[redacted] met "vriend"** Zag rapp. de voor hem bekende H[redacted] over straat lopen met in haar kielzog BE [redacted] volgende. Op het moment dat [redacted] rapp. zag nam hij wat afstand van H[redacted]. Nadat [redacted] en H[redacted] voorbij rapp. waren gelopen kwamen ze weer samen en liepen ze de Molensteeg uit in de richting van de Zeedijk. De combinatie van beide kwam bij rapp over alsof ze een "stelletje" waren. Gezien de info binnen Blue View waarin vermeld staat dat H[redacted] eentatoeage op haar lichaam heeft waarop de naam B[redacted] staat is het welzeer toevallig dat [redacted] eveneens B[redacted] heet. Het is rapp.ambtshalve bekend dat [redacted] zich in het prostitutie milieu zich bezig houdt. Verder info van deze B[redacted] is dat hij:rond de 30 jaar is, is van Surinaamseafkomst, heeft een donkere huid. Heeft lang haar ingevlochten. Zijn vlechtjes steken onder zijn pet uit. Hij draagt een zwarte leren jas. Hij draagt een spijkerbroek. Zijn postuur is klein rond de 1,70meter. Hij heeft een grof gezicht, bruine ogen. Dit komt geheel overeen met BE [redacted].	Known suspects Prostitution Red light district

Fig. 5. 45 An excerpt of the summary report of B.

5.7.4 Analyze the workforce intelligence of clinical pathways

Poelmans (2010d) demonstrated the power of the combination of FCA and Hidden Markov Models. In this section we will demonstrate how CORDIET is applied to read the data sources and generate the FCA lattices. HMM is not integrated in the current version of CORDIET, but the statistical environment of Matlab is used to generate and visualize the model. In the new version of CORDIET HMM will be integrated.

5.7.4.1 Data sources.

Our dataset consists of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008. They all followed the care trajectory determined by the clinical pathway Primary Operable Breast Cancer (POBC), which structures one of the most complex care processes in the hospital. The treatment of breast cancer consists of 4 phases in which 34 doctors, 52 nurses and 14 paramedics are involved. Figure 5.46 contains a high-level summary of the breast cancer care process. Before the patient is hospitalized, she ambulatory receives a number of pre-operative investigative tests. During the surgery support phase she is prepared for the surgery she will receive, while being in the hospital. After surgery she remains hospitalized for a couple of days until she can safely go home. The post-operative activities are also performed in an ambulatory fashion. Every activity or treatment

step performed to a patient is logged in a database and in the dataset we included all the activities performed during the surgery support phase to each of these patients.

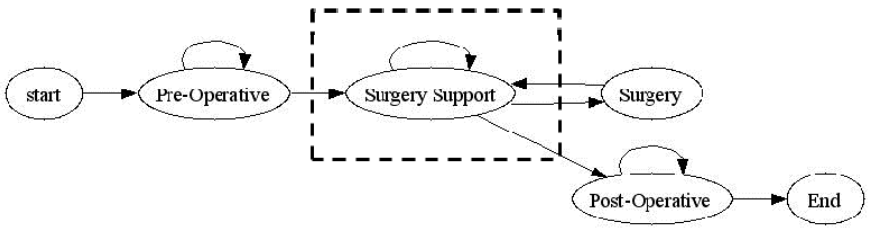


Fig. 5. 46 Breast cancer care process

Each activity has a unique identifier and we have 469 identifiers in total for the clinical path POBC. Using the timestamps assigned to the performed activities, we turned the data for each patient into a sequence of events. These sequences of events were used as input for the process discovery methods. We also clustered activities with a similar semantical meaning to reduce the complexity of the lattices and process models.

The datasets with the activities from the hospital are stored in two flat files with comma separated valued. We transformed both flat files into the XML format as formulated in section 5.6.1.1. This resulted in 33,180 activities. Of the 148 patients from one patient we do not know what surgery she has had, because the surgery record of this patient is missing.

5.7.4.2 Ontology for workflow intelligence

We developed two ontologies for the clinical pathways, one to gain insight in the process complexity and one to gain insight in the length of stay of each patient with respect to each surgeon. The first ontology is fully designed in Dutch. 469 activity codes and 4 surgery codes are grouped into 56 clusters. The second ontology consists of 11 surgeons, 4 surgery codes and 3 temporal attributes. Figure 5.47 shows the second ontology

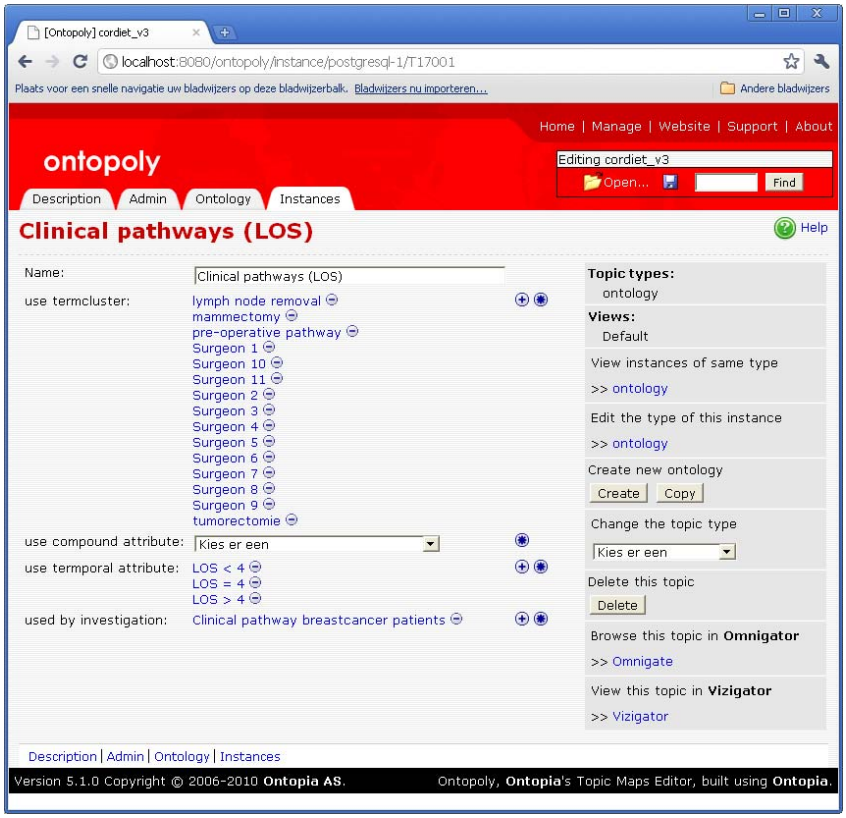


Fig. 5. 47 Workflow intelligence ontology

During the clinical pathway investigation we developed and applied the temporal attributes in CORDIET. Temporal attributes are applied on objects with date time properties. Figure 5.48 shows an example of patients with a length of stay less than 4 days.

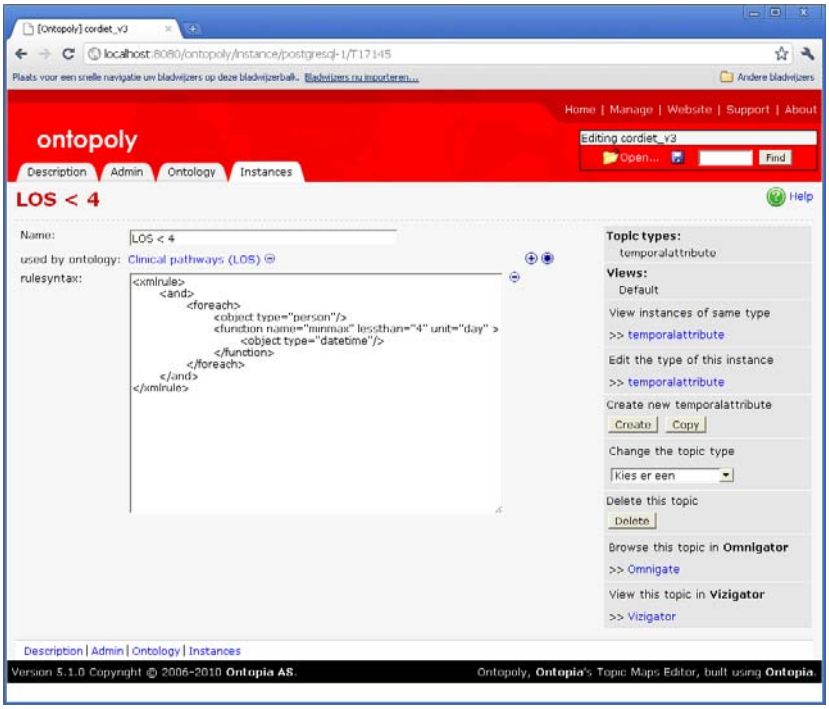


Fig. 5. 48 Temporal attribute of patients with a length of stay less than 4 days

As temporal attributes are applied on date variables, in this case all documents of a person are collected; the difference of the lowest and highest is calculated. If the difference is less than 4, the temporal attribute gets a value of "1" otherwise a value of "0".

5.7.4.3 Process variations

There are five types of breast cancer surgery: mastectomy, breast conserving surgery, lymph node removal and the combination of either mastectomy or breast conserving surgery with lymph node removal. For each of these surgery types, we extracted the corresponding patients in the dataset and constructed a process model and a FCA lattice for in-depth analysis of the characteristics of these groups. Mastectomy surgery consists of completely removing the breast and during breast conserving surgery only the tumor is removed. First we show a FCA lattice with the 147 patients and the surgery they have undergone, the result is shown in Figure 5.49.

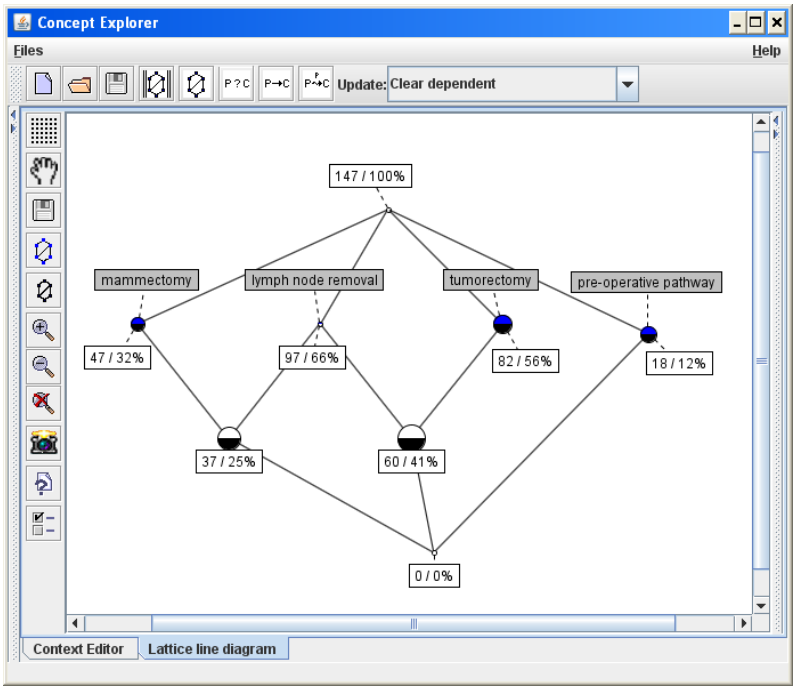


Fig. 5. 49 The 147 patients

Table 5.3 shows the distribution of patients by breast surgery where the numbers are derived from the lattice in Figure 5.49.

Table. 5.3 Distribution of patients by breast surgery

Pre-operative pathway	18
Mastectomy without lymph removal	10
Mastectomy with lymph removal	37
Tumorectomy without lymph removal	22
Tumorectomy with lymph removal	60

Mastectomy surgery consists of completely removing the breast and during breast conserving surgery only the tumor is removed. The process models showed that the complexity of the care process is much larger for the mastectomy patients. Since mastectomy is a more complex surgery type, we expected that the FCA lattices would also be more complex than for breast conserving surgery. Surprisingly we found out that this was not true. The complexity of the lattice was larger for the breast conserving surgery patients and we found that this was due to the less uniform structure of this care process, in which for many patients some essential case interventions were missing. Figure 5.50 contains the interventions performed to the 60 patients receiving breast-conserving surgery with lymph node removal.

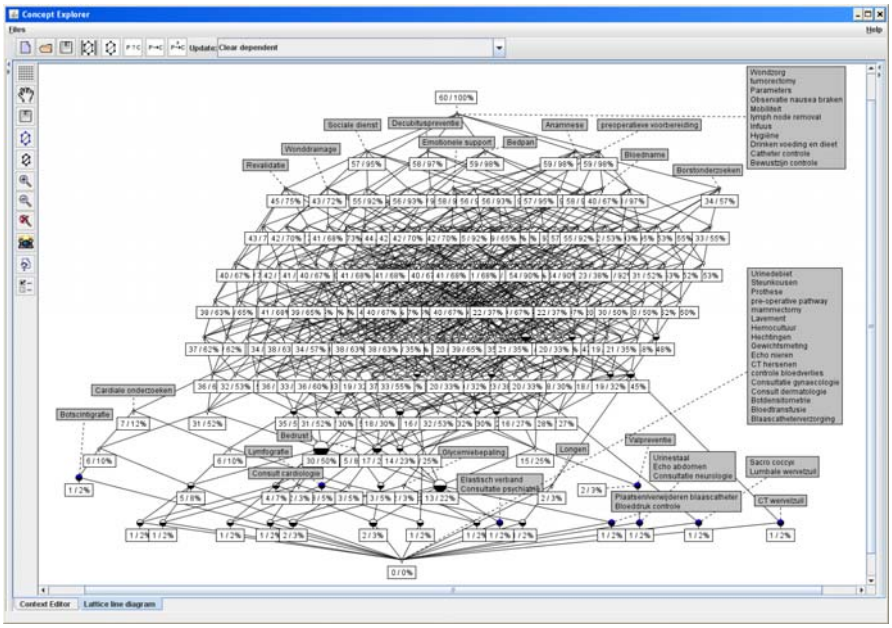


Fig. 5. 50 60 patients with breast conserving and lymph node removal

The Conexp tool has an option to zoom in, we zoomed in on the top level and the result is shown in Figure 5.51.

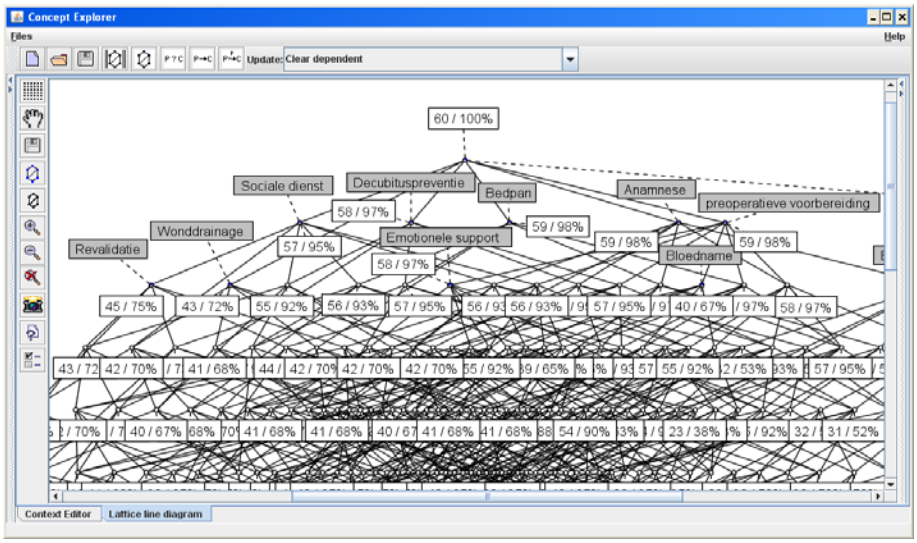


Fig. 5. 51 Outliers of patients with breast conserving and lymph node removal

The lattice of Figure 5.51 shows that 3 of these patients did not receive a consultation from the social support service (“Sociale dienst”). 15 patients did not have an appointment with a physiotherapist and did not receive revalidation therapy (“Revalidatie”). 1 patient did not receive a pre-operative preparation (“preoperatieve voorbereiding”) and 2 patients were missing emotional support before and after surgery (“Emotionele support”).

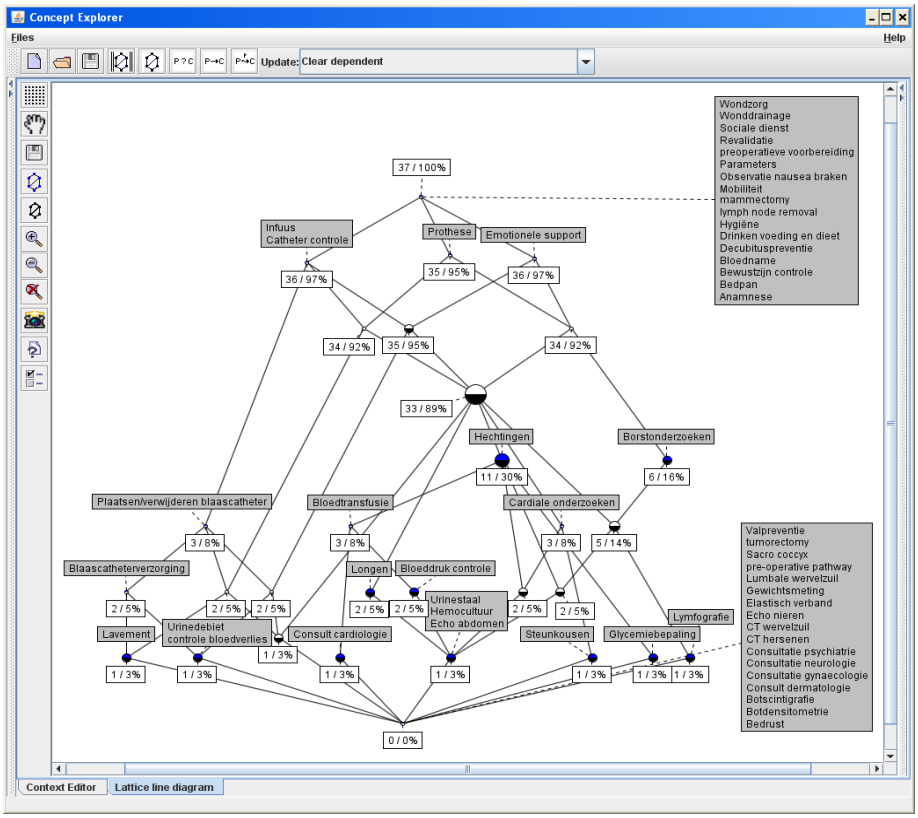


Fig. 5. 52 37 patients receiving mastectomy surgery with lymph node removal

The FCA lattices from Figure 5.52 with the 37 breast patients with mastectomy surgery shows a more structured and less complex clinical pathway than the breast conserving surgery from Figure 5.51.

5.7.4.4 Analyzing the workflow intelligence

We also made a lattice for each type of surgery in which we used as attributes the names of the surgeons and the length of stay of the patients in the hospital. We used temporal attributes to calculate the length of stay of each of the patients and looked at how many patients stayed longer, equal or shorter than this calculated time of' stay. Figure 5.48 in section 5.7.4.2 shows the temporal attribute of a length of stay less than 4 days.

Figure 5.53 contains the lattice for the 60 patients receiving breast conserving surgery with attributes length of stay and doctor performing the operation.

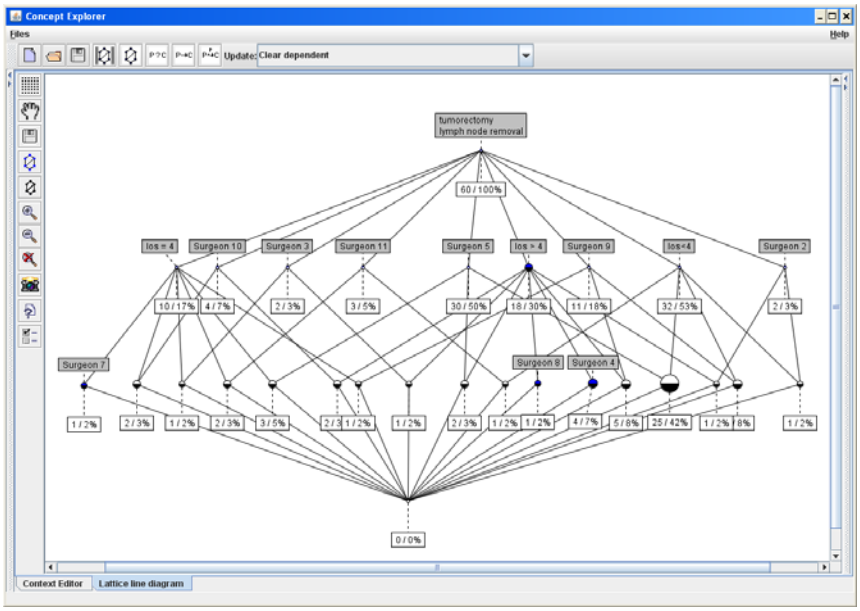


Fig. 5. 53 Length of stay of 60 breast conserving surgery

From this lattice we can conclude for the breast conserving surgery with lymph node removal that 25 patients with a time of stay smaller than 4 days were treated by "surgeon 5", whereas almost all patients treated by the other doctors had a longer time of stay. Poelmans et al (2010d) extracted these subsets of patients and constructed a process model for the groups of patients with a time of stay smaller than 4 days, equal to four days and larger than 4 days. This way, we were able to extract some best practices that could be used to improve the care provided to all patients. Figure 5.54 contains the HMM process model extracted from the datasets with the 10 breast-conserving surgery patients with a length of stay in the hospital of 4 days (the average length of stay). This process model was chosen because of its simplicity in comparison with the other models and since it most closely resembles the standard care process as perceived by the domain experts.

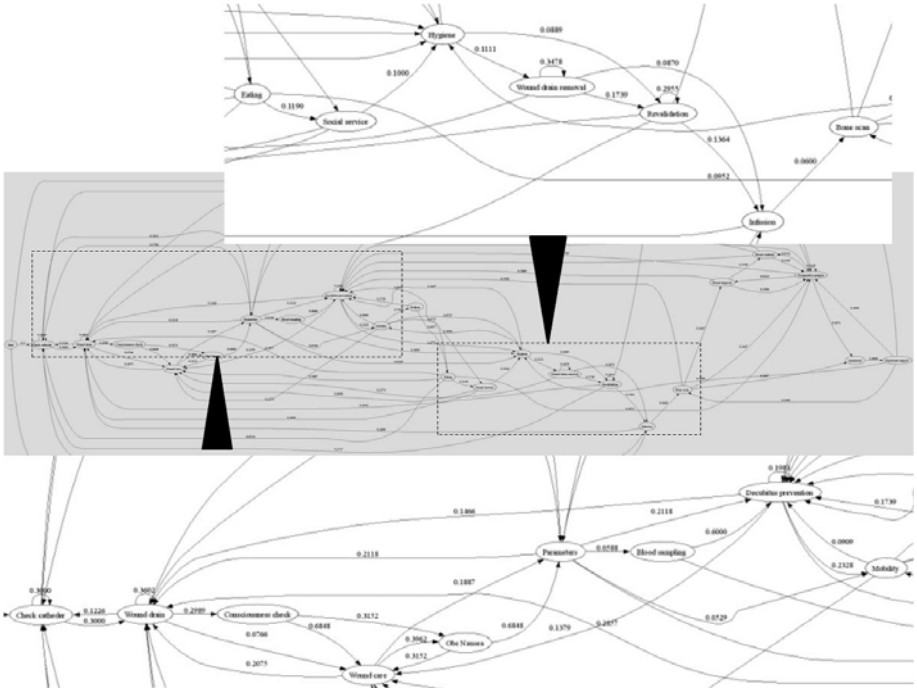


Fig. 5. 54 Process model for 10 breast-conserving surgery patients with length of stay of 4 days

5.8 Conclusions

In this chapter we described the toolbox, Concept Relation Discovery and Innovation Enabling Technology (CORDIET), for gaining new knowledge from unstructured text data. This toolbox has been embedded within the C-K theory, which captures the essential elements of innovation. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artefacts in the analysis process.

At the core of the CORDIET toolbox is the business use case where the C-K transitions are mapped on the functionalities of the toolbox. The C-K functionalities are described in detail and demonstrated with real life cases. CORDIET in its current version has become a very powerful toolbox for mining all general reports of the BVH database of the past 5 years. The Katholieke Universiteit Leuven and the Moscow Higher School of Economics decided to jointly develop an operational software system based on the latest version of CORDIET toolbox. This system will be a user friendly application making visualizations such as FCA, ESOM and HMM more uniformly available to its users where as the current toolbox makes use of integrated open source packages with different user interfaces.

CHAPTER 6

Thesis conclusions

6.1 Thesis conclusions

In this thesis, we investigated the possibilities of using FCA in knowledge discovery. The main theme was applying FCA for concept discovery and representation in various domains such as scientific literature review, text mining, temporal data mining and process mining. 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.

We developed and implemented for the thesis a toolset, Concept Relation Discovery and Innovation Enabling Technique (CORDIET), based on FCA and C-K theory for analyzing police data. The CORDIET data and process discovery environment was designed to cope with the vastly growing amount of structured and unstructured (often textual) information. The architecture of CORDIET takes its roots in three case studies with the Amsterdam-Amstelland Police Department, which were started in 2007. The first project aimed at automatically identifying domestic violence in police reports and was awarded with the best paper award at the Industrial Conference on Data Mining (ICDM) in 2009. Subsequent law enforcement projects dealt with discovering and profiling criminals involved in human trafficking or terrorism-related activities from massive amounts of observational police reports and analyzing chat conversations of arrested pedophiles to identify networks of child abusers. In 2010, the healthcare case study in which the analysis of patient-activity data revealed serious unknown shortcomings in the care process of breast cancer patients and received the best paper award at ICDM in 2010. Given the nature of the research domains we dealt with and the necessity of an expert human being who can be held accountable for each decision being made; we adopted a human-centered KDD approach. At the core of the KDD approach is the C-K design square. Each of the activities belonging to one of the four C-K transitions is implemented and provided to the user. The tool consists of a main window and four components corresponding to each of the four transitions. Functionality includes text mining support such as indexing police reports using Lucene with a thesaurus, FCA lattice visualization and highlighting the selected report with the search terms of the ontology. The goal of CORDIET is not to replace the human expert but to offer the expert with an ergonomic and powerful data analysis toolkit which can be significantly speed up and improve the quality of his or her work.

In chapter 2 we gave an overview of the literature on FCA, covering over 700 papers. Using CORDIET and a thesaurus containing terms and phrases referring to research topics in the FCA community we explored these papers. We built multiple FCA lattices and analyzed them in detail. Data mining and knowledge discovery, information retrieval and ontology engineering were some of the most prominent

research topics. Also, multiple authors expanded FCA with fuzzy theory or rough set theory and for temporal or triadic data. By using FCA to characterize the literature on concept analysis, we not only gained insight into the main research topics but also discovered multiple gaps in the literature which we tried to fill in this thesis. In chapter 2, FCA was found to be an interesting Meta technique for exploring large amounts of text which was further investigated in Chapter 3.

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. In chapter 3 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 process was framed in C-K theory and interpreted as multiple successive iterations through the design square. There was a continuous process of iterating back and forth between analyzing the FCA and ESOM artefacts, selecting reports for in-depth manual inspection, gaining new knowledge and beginning a new knowledge creation cycle. Using FCA we analyzed and ESOM a large set of unstructured text reports from 2007 indicating incidents in the region of the Amsterdam-Amstelland Police Department. We not only uncovered the true nature of domestic violence but also found multiple anomalies, faulty case labeling, and 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 consists 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.

In Chapter 4, we analyzed FCA's applicability to data with an inherent time dimension. We twice made a combination of FCA and Temporal Concept Analysis. We used FCA to distill potential terrorist’s suspects from observational police reports and for suspicious persons a detailed profile was constructed with TCA. Our text analysis method was based 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 Amsterdam-Amstelland Police Department. These subjects are currently being monitored by police authorities. We also used this combination of FCA and TCA 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.

In Chapter 5, we described the CORDIET toolbox, with the C/K theory at its core. The functionality of CORDIET is displayed in a UML business use case. We

demonstrated the toolset at hand of four real life cases by iterating the C/K transitions. We showed the human role is important when moving through the C/K transitions. Each transition needs human interaction for validating the information and making decisions moving to the next C/K transition.

6.2 Future work

6.2.1 Terrorist threat assessment.

Many general reports related to terrorist activity have not been labeled as such by police officers. We want to find all relevant reports since each one may contain crucial information. Using an incremental learning algorithm we plan to build a classifier to automatically label cases. Since there are only few reports labeled as terrorism-related we first construct this model on a small partition of the dataset. The assigned labels are manually verified and a new model is built on a larger training set. The same procedure is repeated until a scalable and operationally useable classification model is obtained. We also intend to use TopicView as a means to validate some of the relationships between police reports and indicators. TopicView will amongst others be used to scan general police reports and incoming email messages on terrorist activity and will offer interesting relationships to the analyst for further investigation. The analyst can confirm or decline these associations and build an FCA model on these manually validated data.

6.2.2 Soloist threateners threat assessment.

On April 30th 2009 Karst Tates drove his car into a crowd, killing 7 people and wounding 10 others[18]. He aimed to kill members of the royal family, and died one day later because of his injuries suffered during the attack. Karst operated alone. People like Karst are called soloist threateners. Soloist threateners are obsessed by their ideas, which mostly are focused on members of the royal family and members of the Parliament. The DKDB, a department of the National Police Service Agency, is responsible for protecting the threatened public persons and is monitoring the known soloist threateners. This department has several problems for which CORDIET may provide a solution:

- Identify new soloist threateners
- Actualize the information about (potentially) soloist threateners
- Risk assessment: How dangerous is this person to our society

The DKDB currently has to collect information from several sources manually and the national search database, Blueview, plays a central role in this process. CORDIET can be used to actualize the available information and identify new soloist threateners. Future research will consist of extending FCA with risk assessment parameters.

[18] http://nl.wikipedia.org/wiki/Aanslag_tijdens_Koninginnedag_2009

6.2.3 Human trafficking.

One of the first steps in future research will be expanding and refining the set of terms related to indicators. Using a combination of FCA, ESOM and Natural Language Processing techniques we intend to build a thesaurus capturing the essential concepts underlying the domain, we will complement our research and current analyses with traditional Social Network Analysis and use FCA to characterize the found groups of suspects. The human trafficking team will provide us with a labeled dataset of significant size. After a testing phase in which the practical usefulness of our method is validated, we will embed our analysis approach in daily operational policing practice.

6.2.4 Domestic violence.

Till date, we only performed analyses on reports containing a statement made by the victim to the police. Recently, the criminal code of the Netherlands changed and now allows for proactive searching of suspects. In the future, our analyses will mainly focus on general reports describing observations made by officers. We will also develop a risk assessment model for estimating the probability that a person will become a repeat offender. This model will be based on early warning indicators; some of them were already discovered during the KDD exercise.

6.2.5 Improve the information quality of the BVH system.

The rule base system we developed for to detect unlabeled domestic violence cases, can also been used to detect other cases, like discrimination (race, sex, religion etc) and use of weapons during violence acts. CORDIET can be used to develop an ontology for and a rule base for each of all rules of the in-triage system Trueblue. The rule base application should be redesigned and integrated with the Trueblue. The overall quality of the BVH system would be improved significantly.

6.2.6 Financial Crime Analysis.

Money laundering and financial crime in general are serious problems for the Amsterdam-Amstelland Police Department. Large amounts of transactions, money flows that are only partially visible to law enforcement authorities, etc. made it difficult to detect suspicious behavior. The domain is characterized by vast amounts of data which are rapidly changing on a continuous basis. We will investigate the possibilities of Emergent Self Organizing Maps, process discovery and neural network pattern recognition techniques to gain insight in these data.

6.2.7 Predicting crime careers

At the Amsterdam-Amstelland Police Department 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.

6.2.8 Supporting Large-scale investigation Teams

Despite the pro-active police work, crimes still are committed and for this purpose, the police deploy so-called large-scale investigation team's (TGO's in Dutch). Each TGO starts with collecting all information about the case and uses different information sources, from the own information sources, like the BVH, to the information found on the confiscated computers of the suspects. CORDIET could be used as an instrument to explore the different information sources in a more intelligent way. We will extend CORDIET with interfaces to communicate online with the various information sources and investigate the possibilities of CORDIET.

6.2.9 Intelligence Led Policing and Concept Discovery Toolset.

In cooperation with the Katholieke Universiteit Leuven and the Moscow Higher School of Economics we will redesign and redevelop the CORDIET toolset. The new toolset will consist of a main window and four tales corresponding to each of the four arrows. The main extension of the new version of CORDIET will be the open data connectors, the more user friendly user interface for maintaining the ontology and the rules, replace the current FCA component with an optimized one which can handle larger number of concepts and integrate HMM-components based on the open source statistical environment R[19] or Apache Mahout[20].

[19] http://www.r-project.org/about.html
[20] http://mahout.apache.org/

SAMENVATTING

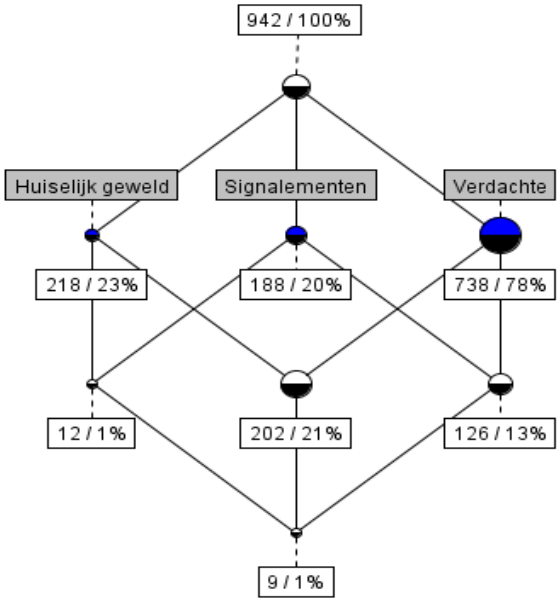
In het kader van de leerstoel "Knowledge Discovery in Databases" hebben de Katholieke Universiteit Leuven en de Regiopolitie Amsterdam-Amstelland de afgelopen jaren een aantal nieuwe analysemethodes ontwikkeld. Deze analysemethodes hebben betrekking op domeinen als huiselijk geweld, mensenhandel en terreur.

De ontwikkeling van de “Informatie of Intelligence Gestuurde Politie” heeft geleid tot een jaarlijkse toename van het aantal aandachtsvestigingen, algemene mutaties en overige meldingen binnen de BVH, het bedrijfsprocessensysteem dat bij alle politiekorpsen in Nederland in gebruik is. Het gaat hier om rapportages met eigen waarnemingen van de mensen op straat (het ‘blauw’) die worden opgeslagen als ongestructureerde tekst binnen de BVH. Tot nu toe werd er relatief weinig gedaan met de mogelijkheden die deze steeds groeiende, ongestructureerde, gegevensverzamelingen bieden om uit deze verzamelingen nieuwe gestructureerde informatie te genereren om het politiewerk beter te ondersteunen. Het hoofddoel van de samenwerking tussen de Regiopolitie Amsterdam-Amstelland en de Katholieke Universiteit Leuven werd het ontwikkelen van een nieuwe, efficiënte en operationeel inzetbare methode om bruikbare kennis uit deze grote hoeveelheden ongestructureerde informatie te onttrekken en toe te passen. Deze methodes moeten leiden tot een betere en snellere herkenning van (nieuwe) potentiële daders en slachtoffers. Voor dit doel is de afgelopen drie jaar gewerkt aan een drietal projecten: huiselijk geweld, mensenhandel (sexuele uitbuiting) en terrorisme (moslim radicalisering). Gedurende dit onderzoek is een toolbox ontwikkeld, Concept Relation Discovery and Innovation Enabling Technology (CORDIET). Aan de basis van deze toolbox ligt de C-K theorie van Hachtuell et al. (1999, 2002 en 2004) welke transitiestappen bevat voor het verkennen van bestaande en ontdekken en toepassen van nieuwe kennis. Belangrijk bij de transitiestappen is de rol van de onderzoeker. Deze moet bij elke stap de waarde van de informatie beoordelen en beslissingen nemen welke informatie meegenomen moet worden naar de volgende transitiestap. De transitieprocessen kunnen gezien worden als kennisexploratie-stappen waarbij elke stap leidt tot het concretiseren en het operationaliseren van de verworven kennis. Deze werkwijze sluit nauw aan bij het proces van informatiegestuurde politie.

2. Huiselijk geweld

Het eerste project ging van start in 2007 en had als doel om een automatische detectie van huiselijk geweld binnen de BVH database mogelijk te maken. De oorspronkelijk uit de wiskunde afkomstige techniek “formele conceptanalyse” (Wille 1982, Ganter et al. 1999) waarin data geanalyseerd worden met behulp van conceptgrafen, werd gebruikt om interactief de onderliggende concepten en eigenschappen van huiselijk geweld (Van Dijk 1997) af te bakenen. De eigenschappen van huiselijk geweld werden weergegeven in de vorm van indicatoren die bestaan uit woorden en/of combinaties van woorden. De open source tool Lucene werd gebruikt om de tekstuele rapporten te indexeren met deze termen

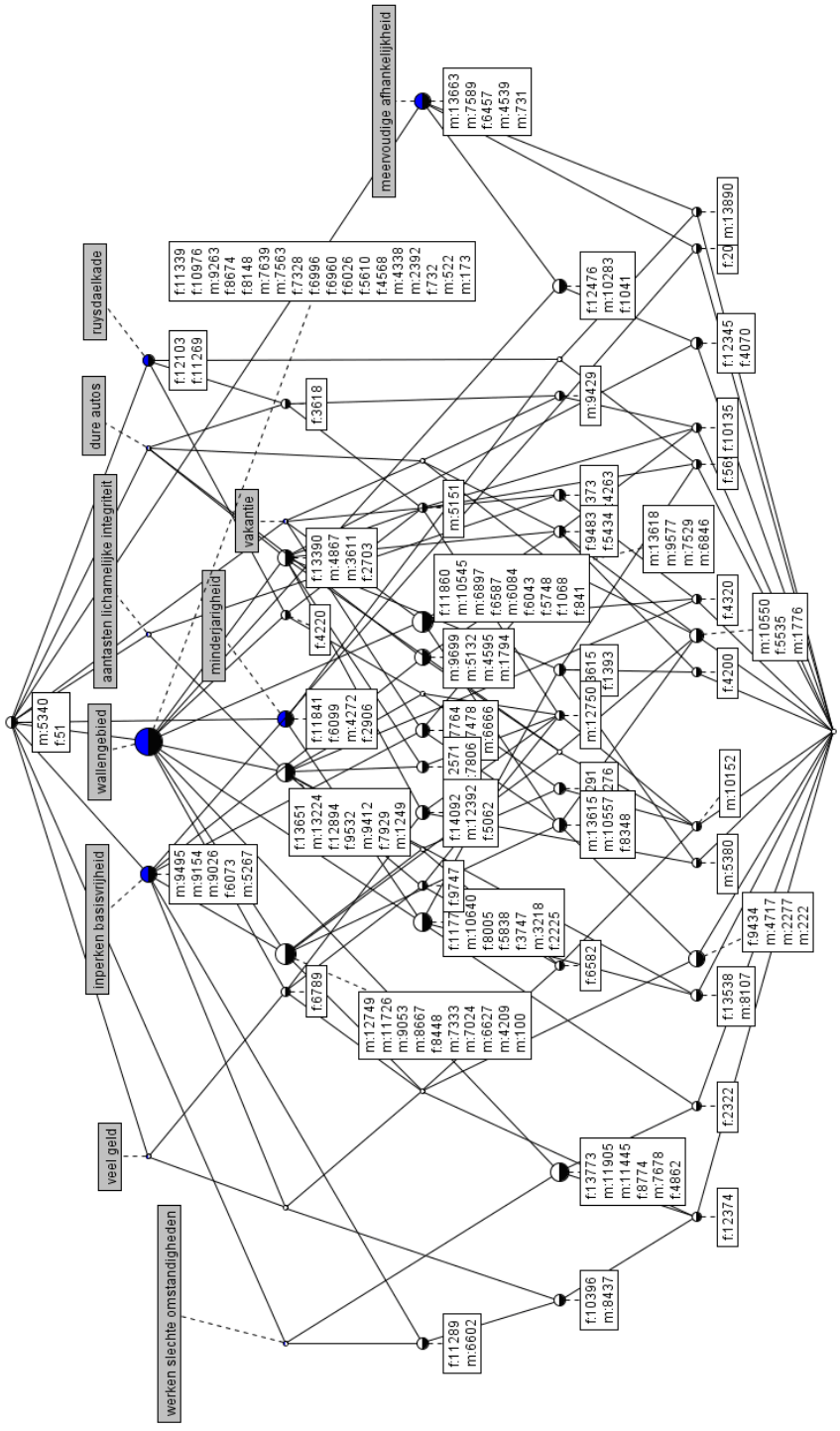
en zinnen. Met behulp van de visualisatie van de conceptgrafen op basis van de indicatoren en BVH-zaken werd het mogelijk kennisregels te ontdekken. Het proces van samenstellen van de indicatoren en kennisregels had tot gevolg dat de definitie van huiselijk geweld verder verfijnd kon worden. Zo konden situaties ontdekt worden die door rapporteurs als verwarrend werden beschouwd. Ook kwamen talloze foutief als huiselijk geweld aangemerkte zaken boven water. Dit onderzoek heeft geresulteerd in een nieuw kennisregel-gebaseerd systeem dat zaken met huiselijk geweld uit de BVH selecteert (Poelmans et al. 2009, Elzinga et al. 2009). Op dit moment wordt binnen de Regiopolitie Amsterdam-Amstelland dit kennisregel-gebaseerde systeem toegepast in combinatie met nTrueblue, het landelijke beheersysteem voor gegevenskwaliteit. Dit kennisregel-gebaseerde systeem kan overigens ook worden toegepast om andere zaken te selecteren, zoals in dit onderzoek is gedaan voor terrorisme en mensenhandel. Onderstaand figuur geeft een voorbeeld van een visualisatie van een “formele conceptanalyse” van mogelijk foutief geclassificeerde zaken van huiselijk geweld in de vorm van een conceptgraaf. De knopen in de graaf geven de concepten weer. Elk concept bestaat uit twee delen: een objecten- en een attributenverzameling. De cijfers in de witte kaders geven het aantal objecten weer dat tot dat concept behoort. De attributen staan vermeld in de grijze kaders. Een concept heeft een attribuut als we vertrekkend van de bijhorende knoop, enkel de lijnen naar boven volgen en bij dit attribuut kunnen uitkomen. De graaf in de onderstaande figuur kunnen we bijvoorbeeld op de volgende manier aflezen. Neem de knoop helemaal onderaan, dit concept bevat 9 politierapporten. Volgen we de lijnen naar boven, dan komen we uit bij de attributen “huiselijk geweld”, “signalementen” en “verdachte”.



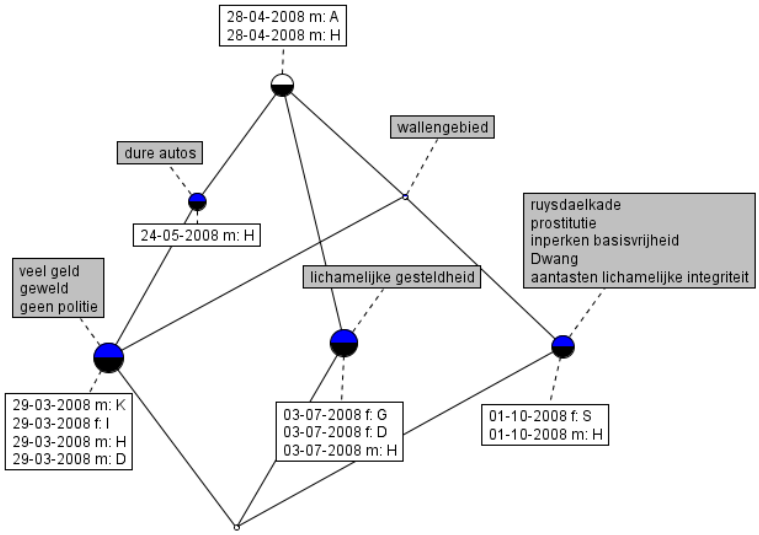
Van de 218 huiselijk geweld zaken zijn er 202 zaken (rechts naar beneden) waarbij een verdachte genoemd wordt. Verder is te zien dat er 9 zaken zijn die als huiselijk geweld gelabeld zijn, waarbij zowel een verdachte genoemd wordt als een signalement aanwezig is. Nader onderzoek leert dat van deze verdachten geen vaste woon- en/of verblijfplaats bekend is en dat een opsporingsbericht is uitgegaan. Dan blijven er nog 3 zaken van huiselijk geweld over waar een signalement voorkomt en geen verdachte wordt genoemd. Al deze 3 zaken bleken foutief als huiselijk geweld aangemerkt te zijn. Uit deze analyse kan een kennisregel afgeleid worden: dat van geweldszaken waarbij een signalement voorkomt, maar er geen verdachte wordt genoemd er met bijna 100% zekerheid gezegd kan worden dat het geen huiselijk geweld kan zijn.

3. Mensenhandel

De volgende stap is het toepassen van de kennisexploratietechniek “formele conceptanalyse” om (nieuwe) potentiële verdachten en slachtoffers te herkennen en te profileren. Het eerste domein was mensenhandel met als motief sexuele uitbuiting van het slachtoffer, een veel voorkomend misdrijf waar de aangiftebereidheid zeer laag ligt (Poelmans et al. 2010a, Highes 2000). Nadat de fase van het samenstellen van de relevante indicatoren is doorlopen, kan met deze methode van een potentiële verdachte of slachtoffer een gedetailleerd profiel gegenereerd worden met daarin de datum van observatie, de indicatoren en de contacten met andere betrokkenen. De eerste stap is het herkennen van potentiële verdachten en slachtoffers. In deze figuur zijn de namen geanonimiseerd en is voor de leesbaarheid een aantal indicatoren weggelaten.


meervoudige afhankelijkheid
m:13663
m:7589
f:6457
m:4539
m:731
ruysdaelkade
dure autos
aantasten lichamelijke integriteit
wallengebied
inperken basisvrijheid
veel geld
werken slechte omstandigheden
minderjarigheid
vakantie
f:11339
f:10976
m:9263
f8674
f8148
m:7639
m:7563
f:7328
f6996
f6960
f:6026
f:5610
f4568
m:4338
m:2392
f:732
m:522
m:173
f:12103
f:11269
f:3618
m:5151
m:5340
f:51
f:11841
f:6099
m:4272
f:2906
f:4220
f:13390
m:4867
m:3611
f:2703
f:11860
m:10545
m:6897
f:6587
m:6084
f:6043
f:5748
f:1068
f841
m:9699
m:5132
m:4595
m:1794
m:12750
3615
f:1393
f:13651
m:13224
f:12894
f:9532
m:9412
f:7929
m:1249
f:14092
2571
7764
m:12392
7806
7478
m:6666
f:5062
m:13615
291
m:10557
276
f:8348
f:1177
f:9747
m:10640
f:8005
f:5838
f:3747
m:3218
f:2225
f:6582
f:6789
m:12749
m:11726
m:9053
m:8667
f:8448
m:7333
m:7024
m:6627
m:4209
m:100
f:11289
m:6602
f:10396
m:8437
f:13773
m:11905
m:11445
f:8774
m:7678
f:4862
f:12374
f:2322
f:13538
m:8107
f:9434
m:4717
m:2277
m:222
m:10152
m:5380
m:9495
m:9154
m:9026
f6073
m:5267
f:12476
m:10283
f:1041
m:9429
f:9483
373
f:5434
4263
m:3618
m:9577
m:7529
m:6846
f:4320
f:4200
m:10550
f:5535
m:1776
f:20
m:13890
f:12345
f:4070
f:56
f:10135

De personen (f = female en m = male) onderin de figuur komen het eerst in aanmerking als potentiële verdachte of slachtoffer aangezien personen lager in de graaf aan meer indicatoren voldoen. Van elke persoon uit deze figuur kan een afzonderlijke analyse gemaakt worden. Een selectie van een van de mannen links onderin de figuur levert de volgende "formele conceptanalyse" op:

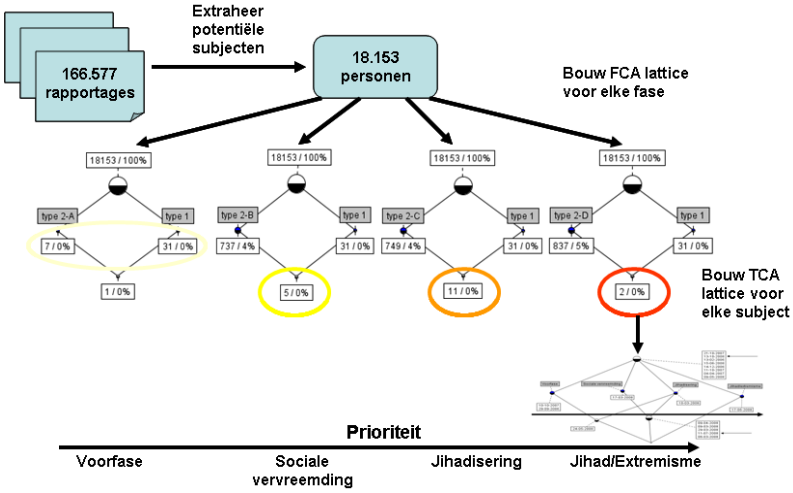


In deze figuur zijn verschillende tijdstippen van de waarnemingen weergegeven bij de indicatoren. De variant van "formele conceptanalyse" die gebruik maakt van temporele gegevens is de "temporele conceptanalyse" (Wolff 2005). Uit de figuur blijkt dat man D (4e links onderin) mogelijk verantwoordelijk is voor de logistiek, omdat deze in een dure auto rijdt waarin de inzittenden gedrag vertonen dat ze liever niet met de politie in contact willen komen. De man H (in alle objecten voorkomend), is de mogelijke pooier, waarbij de vrouw S (1e rechts bovenin) zijn vermoedelijke slachtoffer is, omdat hier sprake is van prostitutie onder dwang. Aan de hand van deze figuur kan in combinatie met de bijbehorende rapporten worden beoordeeld of een *27 constructie*, een document op basis van artikel 273a van het Wetboek van Strafrecht (Staatscourant 2006, 58) omtrent beleidregels opsporing/bevoegdheden mensenhandel, kan worden samengesteld. Dit is een document dat voorafgaat aan eventueel verder strafrechterlijk onderzoek tegen de man H.

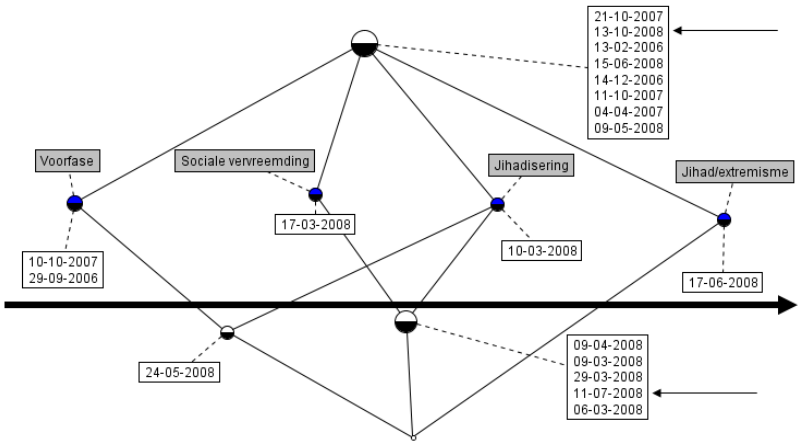
4. Terrorisme

In het laatste project is samengewerkt met het projectteam Kennis in Modellen (KiM) van het Korps Landelijke Politie Diensten (KLPD). Daarbij werd de kennisexploratietechniek ingezet om het moslim radicaliseringmodel van KiM te gebruiken voor het actief opsporen van potentiële terreurverdachten (Elzinga et al. 2010, AIVD 2006). Ook hier bleek het toepassen van de kennisexploratietechniek van de formele conceptanalyse zeer bruikbaar. Waar bij mensenhandel gezocht werd naar profielen in de tijd, is er bij moslim radicalisering sprake van een groeimodel,

waarbij een potentiële verdachte verschillende fasen van radicalisering doorloopt. Het projectteam van KiM heeft op basis van expertonderzoeken een verzameling van 35 indicatoren samengesteld op grond waarvan een persoon in een bepaalde fase kan worden gepositioneerd. Samen met de KLPD is intensief gezocht naar kenmerkende woorden en woordcombinaties die de verschillende indicatoren kenmerken. Het verschil met de voorgaande modellen is dat het KiM-model een extra dimensie toevoegt in de vorm van het aantal verschillende indicatoren waaraan een radicaliseringniveau dient te voldoen.



De analyse is uitgevoerd op de verzameling waarnemingen uit de Basis Voorziening Handhaving (BVH) van de Regiopolitie Amsterdam-Amstelland over de jaren 2006, 2007 en 2008 met als resultaat dat uit 166.577 rapporten 18.153 personen werden gevonden die aan minimaal 1 indicator voldoen. Uit deze 18.153 personen werden 38 personen gevonden die voldeden aan de 1^{e} fase van de radicalisering. Nadere analyse brengt aan het licht dat 19 terecht geselecteerd waren, waarbij 3 personen niet bij de Regiopolitie Amsterdam-Amstelland als zodanig bekend waren, maar wel bij de KLPD. Van deze 19 personen bleken er uiteindelijk 2 te voldoen aan minimale voorwaarden van de extremistische fase. Van een van deze personen is een profiel gemaakt van alle indicatoren verspreid over de tijd.



Uit deze figuur is af te leiden dat de betrokken persoon de extremistische fase heeft bereikt op 17 juni 2008 en na die tijd nog 2 keer is waargenomen door surveillanten (de 2 pijlen rechtsboven en rechtsonder in de figuur) op 11-07-2008 en 13-10-2008.

5. CORDIET

Steeds meer bedrijven beschikken over grote hoeveelheden ongestructureerde gegevens, veelal in tekstuele vorm. De weinige analyse-instrumenten die zich richten op dit probleemgebied bieden onvoldoende functionaliteit voor de specifieke behoeften van veel van deze organisaties. In het kader van het onderzoekswerk verricht in het doctoraatsonderzoek van Jonas Poelmans (Aspirant FWO[21]) werd in september 2010 gestart met de ontwikkeling van de toolset Concept Relation Discovery and Innovation Enabling Technology (CORDIET) in samenwerking met de Moscow Higher School of Economics. Onder toezicht van Prof. dr. Sergei Kuznetsov, drs. Paul Elzinga en dr. Jonas Poelmans werd een projectplan opgesteld, waar 20 master studenten, 2 doctoraatsonderzoekers, 2 postdoctorale onderzoekers en 2 professoren, allen uit Rusland afkomstig, actief aan deelnemen. Het resultaat van deze samenwerking zal de compleet ingerichte toolset CORDIET zijn, waaronder de succesvolle toepassing van deze toolset op ongestructureerde rapportages van de Regiopolitie Amsterdam-Amstelland en medische verslagen van de GZA ziekenhuizen.

Deze toolset zal ingezet worden in de doorlopende projecten voor de proactieve opsporing van mogelijk potentiele verdachten van terrorisme en mensenhandel in de politieregio Amsterdam-Amstelland. In Elzinga et al. (2010) werd al een proof of concept uitgevoerd die de kracht van onze aanpak met conceptgrafen en andere visualisatietechnieken zoals “emergent self organising maps” heeft aangetoond voor de opsporing van individuen die radicaliserend gedrag vertonen. Gedurende dit onderzoek werd een aantal mogelijke verdachten en slachtoffers van mensenhandel

[21] FWO: Fonds voor Wetenschappelijk Onderzoek - Vlaanderen

geanalyseerd en geprofileerd (Poelmans et al. 2010c). Deze toolset biedt de mogelijkheid om veel sneller en gedetailleerder data analyses uit te voeren en relevante personen uit politiegegevens te distilleren. De werkwijze van deze toolset past niet alleen in de filosofie van de Informatie Gestuurde Politie maar past ook binnen een ziekenhuiscontext waar de behandelingsgegevens van borstkankerpatiënten werden geanalyseerd om de verstrekte zorg te verbeteren (Poelmans et al. 2010d). Ook in de GZA ziekenhuisgroep zal deze toolset in een project ingezet worden om de meer dan 43 actieve zorgpaden voor 75 zorgprocessen te verbeteren. Over dit thema is door de Katholieke Universiteit Leuven en de Moscow Higher School of Economics in de zomer van 2011 een workshop georganiseerd met als titel "Concept Discovery in Unstructured Data"[22]. Samen met de Regiopolitie Amsterdam-Amstelland zal worden onderzocht of CORDIET kan worden ingezet voor het voorspellen van criminele carrières van potentiële beroepscriminelen.

De architectuur van de CORDIET toolset bevat 3 lagen. De database laag bevat zowel de data opslag, alsook de ontologie: de tekstdocumenten worden geïndexeerd met Lucene en de ontologie elementen in xml formaat worden vertaald naar Lucene syntax. In de middelste laag worden de "formele conceptanalyse", "temporele conceptanalyse", "emergent self organizing maps", "hidden Markov modellen" en tekstanalysecomponenten gebruikt om visuele modellen te genereren op basis van de data en de ontologie. De derde laag bevat de presentatielaag met de grafische gebruikersinterface.

De grafische gebruikersinterface wordt op een manier ontwikkeld die het toelaat om complexe analyses uit te voeren door mensen met weinig kennis van statistiek en data analyses. In de ontologie kunnen tekstmining attributen gedefinieerd worden om de documenten te analyseren. Temporele attributen kunnen helpen bij het ontdekken van verbanden over de tijd. Samengestelde attributen laten toe om complexe attributen te creëren uit de tekstmining en temporele attributen met behulp van eerste orde logica. Voor deze specifieke ontologische structuren en de bijhorende persistentie (data-opslag) worden nieuwe XML structuren gedefinieerd. Parsers dienen ontwikkeld te worden om de werkomgeving te verbinden aan traditionele data-opslag (SQL databases) en datawarehouse systemen. De modellen gegenereerd met de componenten uit de middelste laag zullen als volgt gebruikt worden:

- "formele conceptanalyse" conceptgrafen: opsporen van verdachten van mensenhandel, terreur, huiselijk geweld etc.
- "temporele conceptanalyse" conceptgrafen: visueel profiel van potentiële daders creëren en interessante patiënten
- "hidden Markov modellen": in kaart brengen van zorgpaden en criminele carrières
- "emergent self organizing maps": in combinatie met "formele

[22] Concept Discovery in Unstructured Data 2011: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-757/

conceptanalyse” de data exploreren

Wij willen nog als bijzonderheid vermelden dat elk van deze vier technieken afzonderlijk weliswaar in één of meerdere statistische omgevingen zijn geïmplementeerd zoals Matlab en SPSS, maar deze technieken zijn nog nooit eerder samen in één omgeving gecombineerd en geïntegreerd. Het gevolg is dat analyses met CORDIET op een veel grotere schaal, veel sneller en efficiënter kunnen worden toegepast,. De user interface maakt mogelijk de ontologie elementen visueel te wijzigen via een graaf. Ook de modellen kunnen eenvoudig gegenereerd en geanalyseerd worden. Bovendien zullen verschillende uitbreidingen voor FCA worden opgenomen, vooral metrieken zoals concept stability, etc.

6. Conclusies

De drie projecten die uitgevoerd zijn in het kader van de leerstoel geven de potentie aan van de kennisexploratietechniek “formele conceptanalyse”. Voornamelijk de intuïtief interpreteerbare visuele voorstelling werd van groot belang gevonden door de informatiespecialisten binnen de politie op zowel de strategische als de tactische en operationele niveaus. Deze visualisatie liet niet alleen toe om interactief de data te verkennen en te analyseren maar ook om de onderliggende concepten van de probleemdomeinen in kaart te brengen. Zo werden onder andere nieuwe concepten, anomalieën, verwarrende situaties en foutief gelabelde zaken ontdekt, maar ook bij de politie niet bekende subjecten die mogelijk betrokken zijn bij mensenhandel of terroristische activiteiten. Ook de temporele variant van de “formele conceptanalyse” bleek van groot nut te zijn bij het profileren van verdachten en hun evolutie over tijd. Niet eerder werden ongestructureerde informatiebronnen ontsloten op een wijze waarop nieuwe inzichten, nieuwe verdachten en slachtoffers zichtbaar werden. Om deze reden zal de “formele conceptanalyse” in de nabije toekomst een belangrijk instrument gaan vormen voor de informatiespecialisten binnen de politie en zal essentieel gaan bijdragen aan de vorming van Intelligence binnen de Nederlandse politie.

DANKWOORD

Ik wil allereerst Guido Dedene en Stijn Viaene bedanken die mij hebben begeleid gedurende het gehele doctorale traject. Onze gezamenlijke bijeenkomsten kenmerkten zich elke keer door stevige wetenschappelijke discussies die mij vele grijze haren hebben bezorgd omdat de finish vaak verder weg leek dan dichterbij. Uiteindelijk is de finish in zicht gekomen met deze thesis als resultaat. Rik Maes is later aangeschoven als co-promotor, waarvoor ik Rik wil bedanken.

Gedurende de meer dan 25 jaar dat ik bij de politie werkzaam ben heeft Hans Schönfeld mijn pad meermalen op belangrijke momenten gekruist door mij nieuwe uitdagingen aan te bieden. Dat heeft in veel gevallen geresulteerd in een nieuwe wending in mijn carrière. Zo ook de kans die Hans mij heeft geboden dit promotiewerk te mogen verrichten. Ik wil Hans hiervoor bedanken. Het was een zware klus, maar ook een rijke levenservaring en het heeft mij vele nieuwe inzichten opgeleverd.

Ik wil Reinder Doeleman bedanken voor de ruimte en de faciliteiten die hij me gegeven heeft om het onderzoekswerk te kunnen verrichten binnen de dienst waar ik werkzaam ben.

Dit promotie-onderzoek had niet tot stand kunnen komen zonder de uitermate vruchtbare samenwerking met Jonas Poelmans. De vele intensieve sessies in Leuven en Amsterdam en gedurende de buitenlandse reizen de afgelopen vier jaar hebben vaak geleid tot nieuwe wetenschappelijke inzichten met als resultaat nieuwe publicaties. Daarnaast heeft Jonas een belangrijke bijdrage geleverd door het reviewen van mijn thesis en het aandragen van waardevolle suggesties om de thesis naar een hoger wetenschappelijk niveau te tillen. Tot slot zijn we samen een uitdagend project begonnen met de Hogeschool van Moskou om de software die ik gedurende mijn onderzoekswerk heb ontwikkeld verder te ontwikkelen naar een professionele en gebruikersvriendelijke versie. Deze versie zal geheel vrij zijn van licenties en het zal een uitdaging worden deze software in te zetten voor de ondersteuning van de Intelligence binnen de Nederlandse politie.

Ik wil een groot aantal collega's bedanken met wie ik vele gesprekken en discussies heb gevoerd. Eveline Vermaat bedank ik voor haar inzicht in de dagelijkse praktijk van de kwaliteitszorg van de gegevens en haar feedback op de resultaten van het kennisregelgebaseerde systeem voor huiselijk geweld. Harold van Gelder bedank ik voor de mogelijkheid om een jaar lang te mogen kijken in de keuken van het team Mensenhandel. Jopie van Louvezijn bedank ik voor de gesprekken die ik met haar heb gehad over de signalen van mensenhandel die in de rapportages van de politieman en -vrouw op straat kunnen worden waargenomen. David Luydens bedank ik die me liet zien hoe de tactische en operationele analisten in ons korps omgaan met informatie en hoe de informatieposities worden opgebouwd voor de verschillende dadergroepen. Ron Boelsma bedank ik voor de kans die hij mij heeft gegeven om de formele conceptanalyse te mogen toepassen op het KiM model (Kennis in Modellen). Ik wil Shanti bedanken die mij meer inzicht verschafte in de gebruikte indicatoren van het KiM model en voor haar bijdrage in

een van de wetenschappelijke publicaties. Paul Bruggink bedank ik voor het ontwerp van de cover van deze thesis.

In hetzelfde jaar dat ik was begonnen met het onderzoekswerk kreeg ik een operatie waarbij een cochlear implantaat in mijn linker oor werd geplaatst. Dit is een stukje techniek waarbij de zenuwcellen in het slakkenhuis rechtstreeks worden gestimuleerd door een uitwendige processor. Er ging een wereld van geluid voor me open. Het communiceren met de directe omgeving verliep daardoor stukken eenvoudiger, waardoor ik veel beter dan voorheen in staat was de vele gesprekken en discussies te volgen. Vooral de gesprekken met mijn Vlaamse begeleiders en de deelnemers op de buitenlandse congressen verliepen stukken beter. Met andere woorden, mijn wereld is dankzij het cochlear implantaat een stuk groter geworden.

Dit onderzoekswerk had nooit tot stand kunnen komen zonder een stabiel thuisfront. Ik wil daarom ook mijn vrouw Mirjam bedanken die mij menige avond, menig weekend en menige vakantie heeft moeten missen. Ook bedank ik mijn dochter Maria die mij regelmatig uit mijn onderzoekswerk haalde om samen naar een film te kijken. Daardoor kon ik me af en toe ontspannen.

Tot slot wil ik mijn ouders Jan en Jannie bedanken die altijd in me geloofd hebben. In de jaren ‘70 en ‘80 was het niet gebruikelijk voor doven of ernstig slechthorenden om hoger onderwijs, laat staan universitair onderwijs, te volgen. De dove of slechthorende leerling of student was in die tijd geheel op zichzelf aangewezen; gebarentolken in het onderwijs kwamen pas veel later. Mijn zus Jennifer bedank ik voor de discussies die we hebben gevoerd over ons werk en er zullen er zeker meer volgen.

PUBLICATIONS

Journal papers published/accepted:

1. Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M. & Dedene G. (2009). Gaining insight in domestic violence with emergent self organizing maps, Expert systems with applications, 36, (9), 11864 – 11874. [SCI 2009 = 2.908]
2. Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M. & Dedene G. (2010) Text Mining with Emergent Self Organizing Maps and Multi-Dimensional Scaling: A comparitive study on domestic violence, accepted for Applied Soft Computing. [SCI 2009 = 2.415]
3. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010) Curbing domestic violence: Instantiating C-K theory with Formal Concept Analysis and Emergent Self Organizing Maps. Intelligent Systems in Accounting, Finance and Management 17, 167-191. Wiley and Sons, Ltd. Doi 10.1002/isaf.319.
4. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010) Formally Analyzing the Concepts of Domestic Violence, Expert Systems with Applications 38, 3116-3130. Elsevier Ltd. doi 10.1016/j.eswa. 2010.08.103 . [SCI 2009 = 2.908]
5. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Van Hulle, M. (2009). Analyzing domestic violence with topographic maps: a comparative study, Lecture Notes in Computer Science, 5629, 246 – 254, Advances in Self-organizing Maps, 7th International Workshop on Self-Organizing Maps (WSOM). St. Augustine, Florida (USA), 8-10 June 2009, Springer.
6. 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, Lecture Notes in Computer Science, 5633, 247 – 260, Advances in Data Mining. Applications and Theoretical Aspects, 9th Industrial Conference (ICDM), Leipzig, Germany, July 20-22, 2009, Springer.
7. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2008). An exploration into the power of formal concept analysis for domestic violence analysis, Lecture Notes in Computer Science, 5077, 404 – 416, Advances in Data Mining. Applications and Theoretical Aspects, 8th Industrial Conference (ICDM), Leipzig, Germany, July 16-18, 2008, Springer.
8. 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): from information to intelligence. 26 - 30 July, Kuching, Sarawak, Malaysia. Springer.
9. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011) A concept discovery approach for fighting human trafficking and forced prostitution. Accepted for the 19th International Conference on Conceptual

Structures, University of Derby, United Kingdom.

10. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011) Text mining scientific papers: a survey on FCA-based information retrieval research. Accepted for Industrial Conference on Data Mining 2011.

Conference proceedings published/accepted:

11. 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.
12. Elzinga, P., Poelmans, J., Viaene, S., Dedene, G. (2009), Detecting domestic violence – Showcasing a Knowledge Browser based on Formal Concept Analysis and Emergent Self Organizing Maps, Proc. 11th International Conference on Enterprise Information Systems ICEIS, Volume AIDSS, pp. 11 – 18, Milan, Italy, May 6-10, 2009.
13. Poelmans, J, Elzinga, P., Van Hulle, M., Viaene, S., and Dedene, G. (2009). How Emergent Self Organizing Maps can help counter domestic violence, World Congress on Computer Science and Information Engineering (CSIE 2009), Los Angeles (USA), Vol. 4, IEEE Computer Society Press ISBN 978-0-7695-3507-4, 126 – 136.
14. 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.
15. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010). Concept Discovery Innovations in Law Enforcement: a Perspective, accepted for Computational Intelligence in Networks and Systems workshop (INCos 2010), Thesalloniki, Greece.
16. Poelmans, J., Elzinga, P., Neznanov, A., Kuznetsov, S., Dedene, G., Viaene, S. (2011) Concept relation discovery and innovation enabling technology (CORDIET), First International Workshop on Concept Discovery in Unstructured Data, Moscow, Russia.

Journal papers submitted:

17. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011) Semi-Automated Knowledge Discovery in Unstructured Text: Identifying and Profiling Human Trafficking. Submitted for Expert Systems with Applications.
18. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010) Formal Concept Analysis in Information Engineering: a Survey, submitted for International Journal of General Systems.
19. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Kuznetsov, S. (2011). Fuzzy and rough formal concept analysis: a survey, submitted for Transactions on Machine Learning and Data Mining.
20. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010) Informatiegestuurd politiewerk: een slimme en nieuwe kijk op de data, submitted for Informatie.

Book chapter:

21. Poelmans J, Van Hulle M, Elzinga P, Viaene S, Dedene G (2008) Topographic maps for domestic violence analysis. Self-organizing maps and the related tools, pp. 136 - 145.

APPENDIX A

Literature survey thesaurus

Appendix A of this thesis contains a small excerpt of the thesaurus used during the literature study. The first column of the following table contains the name of the term clusters, the second column contains the associated search terms separated by a semicolon.

Term cluster	Search terms
Algorithm	algorithm; algorithms
Association rule mining	ARM; association rules; association rule; association rule mining; rule extraction; extraction of rules; closed itemset; closed itemsets; frequent itemsets; frequent itemset; frequent closed itemset; frequent closed itemsets; generators of closed itemset; generators of closed itemsets; CHARM; TITANIC; CLOSET; PRINCE; mingen; classification rules
Classification	classification; document classification; classifier; categorisation; SVM; SVMs; Naive Bayes; ripple-down rules
Galois lattices	galois lattice; galois lattices; galois connection lattice; galois connection lattices
Knowledge discovery	CKDD; KDD; Knowledge discovery; exploratory processes; concept discovery; data mining; datamining; DM; knowledge extraction; machine learning; data exploration; exploring data; exploring knowledge; exploring information; knowledge exploration; exploratory data analysis; information exploration
Ontology	ontology; ontologies; ontology construction; ontological
Scalability	iceberg lattices; iceberg lattice; Iceberg concept lattices; iceberg sets; Handling large formal context; Exploration of a large number of objects; Large contexts; huge database; large database; large databases; pruning strategy; pruning strategies; reduce less useful concepts; reduce the size of large concept lattices; scalability; scalable; large data; alpha lattices
Software mining	AOSD; software; source code; aspect oriented; aspect-oriented; aspect identification; aspect mining; aspect-mining; use case; crosscutting concerns; design of class hierarchies; reengineering class hierarchies; class hierarchies; modularity; modularisation; software evolution; Object-Oriented concept analysis; reversed engineering
Usability	usability testing; novices read line diagrams; usability

testing	evaluation
Web mining	web mining; webmining; world wide web; web document management; query web documents; web-based mail browser; digital library; mining the web; web information; web documents; web search; web-pages; web pages; search results; web data mining; web datamining; web user profiles; virtual surfing trials; internet; weblogs; web logs
Conceptuals	conceptual knowledge processing; conceptual deviation discovery; chianti; concept hierarchies; concept hierarchy; conceptual scaling; conceptual graph; conceptual graphs; conceptual logic; propositional logic; logic programming; concept graphs; concept graph; concept similarity
Information retrieval	manage email; Mail-Sleuth; Mail-Strainer; Conceptual Email Manager; CEM; data retrieval; concept retrieval; conceptual retrieval; retrieval; query; queries; knowledge retrieval; instance retrieval; document retrieval; information retrieval; text retrieval; SNOMED; snomed; IR; lookup method; restructuring help system; CREDO; credo

APPENDIX B

Domestic violence case thesaurus

Appendix B of this thesis contains a small excerpt of the thesaurus used during the domestic violence case study. The first column of the following table contains the name of the term clusters, the second column contains the associated search terms separated by a semicolon.

Term cluster	**Search terms**
Geweldsvormen	slaan; geslagen; sloeg; slaat; schoppen; geschopt; schopte; bedreigen; bedreigd; bedreiging; stompen; gestompt; stompte; knijpen; geknepen; steken; stak; steekt; gestoken; vluchten; gevlucht; terroriseren; terrorisering; geterroriseerd; geterroriseert; stalken; stalking; stalker; gestalkt; krassen; krabben; bekrast; gekrabt; bijten; gebeten; duwen; geduwd; worstelen; worsteling; geworsteld; verkrachten; verkrachting; verkracht; verkrachtte; verkrachte; ….
Huiselijke sfeer	mijn zoon; mijn zoontje; mijn vader; mijn moeder; mijn moeke; mijn broer; mijn broertje; mijn zus; mijn zusje; mijn neef; mijn neefje; mijn nicht; mijn nichtje; mijn ex; mijn ex man; mijn ex-man; mijn exman; mijn ex vrouw; mijn ex-vrouw; mijn exvrouw; mijn ex partner; mijn ex-partner; mijn expartner; mijn ouders; mijn ex vriendin; mijn ex-vriendin; mijn exvriendin; mijn papa; mijn pappa; mijn mama; mijn mamma; mijn neef; mijn nicht; mijn oom; mijn tante; mijn stiefmoeder; mijn stiefvader; mijn stiefzus; mijn stiefzusje; mijn stiefbroer; mijn stiefbroertje; mijn grootmoeder; mijn grootvader; mijn opa; mijn oma; mijn grootouders;…
Aangifte tegen	aangifte tegen mijn zoon; aangifte tegen mijn zoontje; aangifte tegen mijn vader; aangifte tegen mijn moeder; aangifte tegen mijn moeke; aangifte tegen mijn broer; aangifte tegen mijn broertje; aangifte tegen mijn zus; aangifte tegen mijn zusje; aangifte tegen mijn neef; aangifte tegen mijn neefje; aangifte tegen mijn nicht; aangifte tegen mijn nichtje; aangifte tegen mijn ex; aangifte tegen mijn ex man; aangifte tegen mijn ex-man; aangifte tegen mijn exman; aangifte tegen mijn ex vrouw; aangifte tegen mijn ex-vrouw; aangifte tegen mijn exvrouw; aangifte tegen mijn ex partner; aangifte tegen mijn ex-partner; aangifte tegen mijn expartner; aangifte tegen mijn ouders; aangifte tegen mijn ex vriendin; aangifte tegen mijn ex-vriendin; aangifte tegen mijn exvriendin; aangifte tegen mijn papa;…

Gepleegd door	gepleegd door mijn zoon; gepleegd door mijn zoontje; gepleegd door mijn vader; gepleegd door mijn moeder; gepleegd door mijn moeke; gepleegd door mijn broer; gepleegd door mijn broertje; gepleegd door mijn zus; gepleegd door mijn zusje; gepleegd door mijn neef; gepleegd door mijn neefje; gepleegd door mijn nicht; gepleegd door mijn nichtje; gepleegd door mijn ex; gepleegd door mijn ex man; gepleegd door mijn ex-man; gepleegd door mijn exman; gepleegd door mijn ex vrouw; gepleegd door mijn ex-vrouw; gepleegd door mijn exvrouw; gepleegd door mijn ex partner; gepleegd door mijn ex-partner; gepleegd door mijn expartner; gepleegd door mijn ouders; gepleegd door mijn ex vriendin; gepleegd door mijn ex-vriendin; gepleegd door mijn exvriendin; …
Angst voor	bang voor mijn zoon; bang voor mijn zoontje; bang voor mijn vader; bang voor mijn moeder; bang voor mijn moeke; bang voor mijn broer; bang voor mijn broertje; bang voor mijn zus; bang voor mijn zusje; bang voor mijn neef; bang voor mijn neefje; bang voor mijn nicht; bang voor mijn nichtje; bang voor mijn ex; bang voor mijn ex man; bang voor mijn ex-man; bang voor mijn exman; bang voor mijn ex vrouw; bang voor mijn ex-vrouw; bang voor mijn exvrouw; bang voor mijn ex partner; bang voor mijn ex-partner; bang voor mijn expartner; bang voor mijn ouders; bang voor mijn ex vriendin;…
geweld door mijn	door mijn zoon slaan; door mijn zoon geslagen; door mijn zoon sloeg; door mijn zoon slaat; door mijn zoon schoppen; door mijn zoon geschopt; door mijn zoon schopte; door mijn zoon bedreigen; door mijn zoon bedreigd; door mijn zoon bedreiging; door mijn zoon stompen; door mijn zoon gestompt; door mijn zoon stompte; door mijn zoon knijpen; door mijn zoon geknepen; door mijn zoon steken; door mijn zoon stak; door mijn zoon steekt; door mijn zoon gestoken; door mijn zoon vluchten; door mijn zoon gevlucht; door mijn zoon terroriseren; door mijn zoon terrorisering; door mijn zoon geterroriseerd….
Nieuwe vriend van ex	door de ex van; door de ex-vriend van; door de ex vriend van; door de exvriend van; door de ex-vriendin van; door de ex vriendin van; door de exvriendin van; ex-vriend van mijn vriendin; ex vriend van mijn vriendin; exvriend van mijn vriendin; ex-vriendin van mijn vriend; ex vriendin van mijn vriend; exvriendin van mijn vriend; nieuwe vriend van mijn ex-vriendin

Problemen met	problemen met mijn zoon; problemen met mijn zoontje; problemen met mijn vader; problemen met mijn moeder; problemen met mijn moeke; problemen met mijn broer; problemen met mijn broertje; problemen met mijn zus; problemen met mijn zusje; problemen met mijn neef; problemen met mijn neefje; problemen met mijn nicht; problemen met mijn nichtje; problemen met mijn ex; problemen met mijn ex man; problemen met mijn ex-man; problemen met mijn exman; problemen met mijn ex vrouw; problemen met mijn ex-vrouw; problemen met mijn exvrouw; problemen met mijn ex partner; problemen met mijn ex-partner; problemen met mijn expartner; problemen met mijn ouders; problemen met mijn ex vriendin; problemen met mijn ex-vriendin; problemen met mijn exvriendin; problemen met mijn papa; problemen met mijn pappa;…

APPENDIX C

Human trafficking thesaurus

Appendix C of this thesis contains a small excerpt of the thesaurus used during the human trafficking case study. The first column of the following table contains the name of the term clusters, the second column contains the associated search terms separated by a semicolon.

Term cluster	**Search terms**
Geweldsvormen	slaan; geslagen; sloeg; slaat; schoppen; geschopt; schopte; bedreigen; bedreigd; bedreiging; stompen; gestompt; stompte; knijpen; geknepen; steken; stak; steekt; gestoken; vluchten; gevlucht; terroriseren; terrorisering; geterroriseerd; geterroriseert; stalken; stalking; stalker; gestalkt; krassen; krabben; bekrast; gekrabt; bijten; gebeten; duwen; geduwd; worstelen; worsteling; geworsteld; verkrachten; verkrachting; verkracht; verkrachtte; verkrachte; ….
Aantasten lichamelijke integriteit	afstaan organen; onvrijwillig werken in de prostitutie; dwang prostitutie; dwang sexuele handelingen; gedwongen worden tot prostitutie; dwingen tot prostitutie; gedwongen prostitutie; gedwongen zijn als prostituee te werken; prostituees dwingen voor hem werken; prostituees dwingen werken; prostituees dwingen te werken; beweegt zich te prostitueren; bewegen tot prostitueren; gedwongen prostitutie; gedwongen tot prostitutie; tegen mijn zin prostituee; borstvergroting; dwingt ze te werken zonder condoom; dwingen zonder condoom; haar te laten werken in de prostitutie; …
Inperken basisvrijheid	geen medische hulp toestaan; onthouden medische hulp; geen bewegingsvrijheid; niet beschikken over eigen verdiensten; hoge afdracht verdiensten; ik verdiende daar dus niets; ik verdienen niets; zeer stevig de vrouw aan de hand vast; haar rechterhand niet losgelaten; kreeg zij zelf niet de mogelijkheid om te antwoorden; had een onderdanige houding; met gebogen hoofd; het meisje huilde; Zij mocht nooit alleen naar buiten; de deur van buitenaf afgesloten; zij niet weg kon; ondergedoken gezeten; intimideren om de verklaring in te trekken; naar nederland gebracht; zien er niet blij uit; zien er moe uit; man deed het woord voor de vrouw; …
Meervoudige afhankelijkheid	vals paspoort; illegaal verblijven; angst voor uitzetting; geen eigen woonruimte; geen eigen woning; geen vast adres; overnachten werkplek; onbekend met werkadres;

	schulden bij exploitant; schulden souteneur; schulden pooier; overnamebedrag betaald; geen werkkamer meer toegewezen krijgen; geen werkkamer toegewezen krijgen; worden begeleid; worden weggebracht; onder begeleiding; vrouwen gaan met verschillende mannen mee; aantal vrouwen aan de bar zitten; paspoort hierna nooit meer gezien; paspoort nooit meer gezien; …
Onbekende adressen	Adres wist hij uiteraard niet; Verhullen adres; verhullen daadwerkelijke verblijfsplaats; verklaarde geen adres in nederland te hebben; maar het exacte adres wist ze niet; niet wetende welke straat; zijn eigen adres wist hij ook niet; Hij wist zijn adres niet; niet het adres welke in Xpol staat; naam wist ze niet van het hotel; ze wist geen adres; geen adres weten; weigerden hun adres op te geven; adres weigeren op te geven; absoluut geen adres opgeven; …
Geen politie	nerveuze indruk; zenuwachtige indruk; keken zenuwachtig; kijken zenuwachtig; reed snel weg; snel weg rijden; geen pooier; zich niet bezig te houden met vrouwen; geen sex voor geld; niet betaald voor sex; erg zenuwachtig; durfde ons niet aan te kijken; vrouw praat niet; heeft geen woord gezegd; nam een zwijgende positie in; ze wou niets zeggen; heeft geen woord gesproken; zeer sombere blik in haar ogen; erg nerveus; vrouw achterin was erg stil; vrouw achhterin was erg stil; passagier heeft geen woord gezegd; angst voor de politie; raakten wat geirriteerd; …
ID-bewijzen	haar paspoort niet bij haar had; haar paspoort niet zelf bij haar droeg; niet over haar eigen identiteitsbewijs beschikte; hij had de beschikking over het paspoort; een geldig Hongaars document overhandigen; om haar paspoort te tonen en haar tas vragen; hadden geen identificatiemiddel bij zich; maar hij had wel haar paspoort; kon geen identiteitsbewijs tonen; Paspoort is door gehaald; haar id kaart had de oudere man in zijn portemonnee; had de ID van onder zich; …
Illegale prostitutie locaties	veel condooms; condooms in grote hoeveelheden; veel comdooms; hele rits condooms; keuken was omgebouwd tot woonkamer; twee grote twee-persoonsbedden; matras in de kofferbak; rondrijdt; volstonden met bedden; in de auto lagen condooms
Lichamelijke kenmerken	tatoeage; getatoueerd; tattoo; tatto; piercing; neuspiercing; navelpiercing; tribal-tattoo; tongpiercing; tribal op haar onderrug
Beïnvloedbaarheid slachtoffers	een moeilijk verleden; niet de slimste; zwak begaafd; verstandelijk gehandcapt; vatbaar is voor dit soort dingen

Werken onder slechte omstandigheden	slecht betaald; laag loon; werken onder gevaarlijke omstandigheden; gevaarlijk werk; lange werkdagen; lange werkweken; buitenproportioneel lang werken; chantage familie; bedreiging familie; smokkel alleenstaande vrouwen; slaafse houding; gebouwen met camera; schuilplaatsen; fake-inrichting; nooit meer zal zien; Mocht dat niet betaald worden; Jij gaat dood; ik moest altijd werken; altijd moeten werken; dubbele diensten draaien; haar familie iets aan zou kunnen doen; hij wist waar mijn ouders woonden; geld van haar afpakt; geld van haar afpakken; alle dagen per week te werken; …
Autohandel	autohandel; handel in autos; handel in auto; auto handel; handel in autoonderdelen; handel in auto onderdelen; auto naar het buitenland uit te voeren; autos te kopen en in Bulgarije de autos door te verkopen; handel in tweedehandsautos; handelen in auto's; handelen in auto; handelen in autos

APPENDIX D

Simulating the Trueblue Domestic Violence rule

To compare the distilled rules from chapter 3 with the rules of the in-triage system Trueblue, we used Cordiet to simulate the Trueblue knowledge rule with respect to suspicious domestic violence cases missing a project code label over 2010. Figure D.1 shows a screenshot of the Trueblue domestic violence rule named "10) Projectcode huiselijk geweld ontbreekt" (Project label domestic violence is missing).

Naam	**10) Projectcode Huiselijk Geweld ontbreekt**
Omschrijving	De projectcode Huiselijk Geweld is niet ingevuld bij het voorval.
Gemaild wordt	De eerste rapporteur van het hoofdincident.
Correctiemethode	Voeg de projectcode voor HG toe. U kunt kiezen voor de volgende codes: HG1 (Huiselijk Geweld), HG1.1 (HG gericht op (ex)partner), HG1.11 (HG gericht op partner (man)), HG1.12 (HG gericht op partner (vrouw)), HG1.13 (HG gericht op ex-partner (man)), HG1.14 (HG gericht op ex-partner (vrouw)), HG1.2 (HG gericht op kinderen (-18), HG1.3 (HG gericht op ouderen (+55), HG1.4 (HG gericht op ouders), HG1.6 (HG gericht op huisvrienden), HG1.7 (HG gericht op overige familieleden). **HG1.0 staat voor HG niet van toepassing (geen HG dus).**
Vertraging	Foutmelding wordt 1 dag na kennisname aangeboden aan de kwaliteitsbewaker. De kwaliteitsbewaker beoordeelt of er sprake is van huiselijk geweld.
Voorwaarden	Er is een persoon met de rol AANGEVER, MELDER, BETROKKEN of SLACHTOFFER. En er is een persoon met de rol VERDACHTE. 1) Deze personen hebben hetzelfde woonadres (gelijke woonplaats, straatnaam, huisnummer en toevoeging). 2) Of in de formulieren die bij het incident aanwezig zijn komt één van de volgende termen voor: relatie, ex vriend, ex vriendin, ex man, ex vrouw, huiselijk, stalk, samengewoond of samen gewoond of één van de volgende termen binnen een zin: <kind en bedreig>, <kind en bang>, <zoon en bang>, <zoon en bedreig>, <dochter en bang> en <dochter en bedreig>.
Maatschappelijke klassen	B71, C40, E13, E16, E33, E391, E40, F530, F531, F532, F540, F550, F551 en F552.

Figure D.1 screenshot of the Trueblue domestic violence rule

Trueblue uses two selection rules. First it restricts the dataset based on the classes of incident and activity reports ("Maatschappelijke klassen)" and second the conditions within the restricted dataset ("voorwaarden"). The restricted report classes are:

Table D.1 Overview of the incident and activity reports used by Trueblue

Class	**Description**
B71	Theft with violence (relational sphere)
C40	Destruction remaining objects
E13	Domestic quarrel (without consequences)
E16	Quarrel (without consequences)
E33	Nuisance by confused or overstrained person
E391	Nuisance by stalker
E40	Handling remaining mentions
F530	Threatening
F531	Remaining crimes against deprivation of personal liberty
F532	Hostage/kidnap
F540	Manslaughter/murder

F550	Simple assault
F551	Aggravated assault
F552	Remaining assault

We used an export application with input parameters report classes and period. The export application produces for each case a separate XML file which can be imported into Cordiet and indexed by Lucene. The export with the selected incident and activity reports over 2010 results in a dataset of 37,294 XML files. The XML files are imported in Cordiet and made available for analysis. The dataset also includes the already labelled domestic violence cases to visualize the detected domestic violence cases by Trueblue.

To simulate the knowledge rule, we build a thesaurus which uses the same conditions as formulated in Figure A.1.

Table D.2 the simulated Trueblue thesaurus

	Rule	**Description**
1	“Slachtoffer”	Person with the role of victim
2	“Verdachte”	Person with the role of suspect
3	“Zelfde adres”	Victim and suspect living at the same address
4	“HG-codes’	Cases labelled as domestic violence
5	“Ex leden’	one of the searchterms “relationship”,”ex boyfriend”, “ex girlfriend”,.”ex husband”, “ex wife”, “domestic”, “stalk”, “lived together”
6	“Familie en angst”	combination of two searchteams within one sentence: <child, threat> , <child, fear>, <son, thread> , <sun, fear>, <daughter, threat>, <daughter, fear>

Trueblue uses a COR construction between the first three elements of the thesaurus from table D.2 and the last three elements. If the condition of suspect and victim living at the same address is met, the case is detected by Trueblue as suspicious domestic violence. If this condition is not met, rules 5 and 6 are applied. Applying the simulated thesaurus on the dataset of 37,294 cases will result in the FCA lattice in Figure D.2.

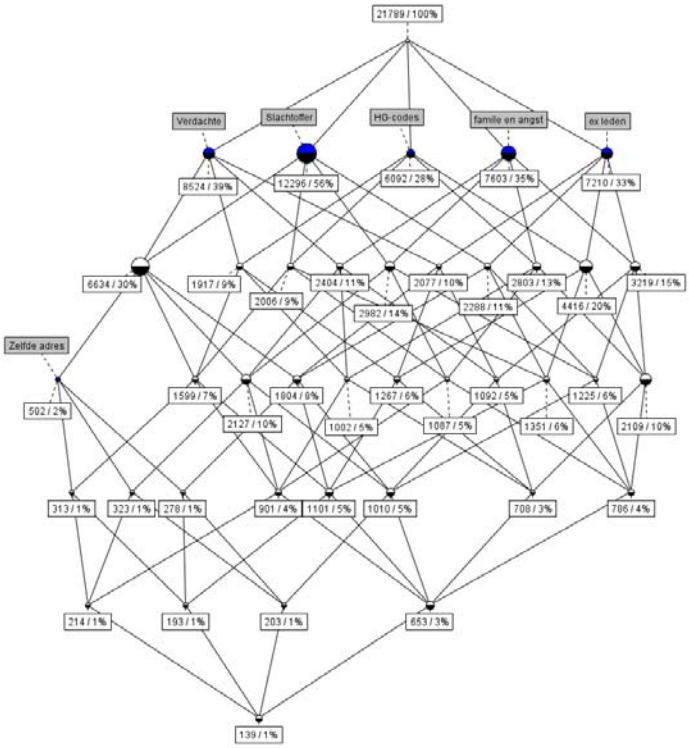


Fig D.2 FCA lattice with the Trueblue rule of domestic violence

There are 21,789 cases which met at least one of the 6 rules and 139 cases which met all rules. The first rule with victim and suspect living at the same address can be inferred by the node “zelfde adres” with 502 cases. Going down one node we retrieve 313 cases labelled as domestic violence which results in 189 cases suspicious cases according to Trueblue. The same we can infer from “familie en angst”: 7,603 – 2,803 = 3,800 cases and “ex leden”: 7,210 – 4,416 = 2,794 cases. The lattice shows that 3,216 cases have both search terms from “family en angst” and “ex-leden” and no domestic violence label. To show the exact number of cases which met the conditions and are not labelled as domestic violence, we use the lattice option of “show own objects” and we get the lattice in Figure D.3.

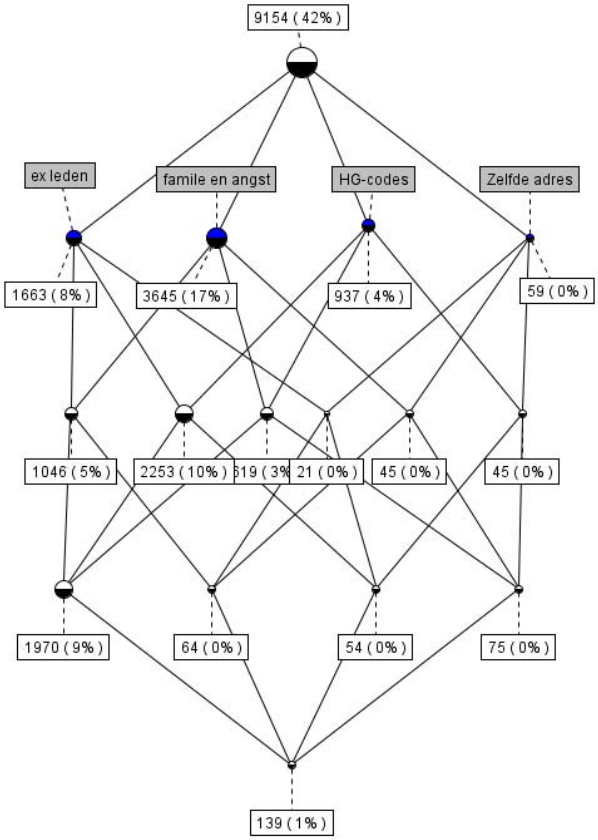


Fig D.3 Lattice showing the detected cases by Trueblue

From the lattice in Figure D.3 we can infer the results of the Trueblue domestic violence rule by summing the objects from the second row from the nodes “ex leden”, “familie en angst” and “zelfde adres”. This results in a total of 1663+3645+59 = 5367 suspicious domestic violence cases by Trueblue.

APPENDIX E

The rule based application

The rule based application is a java application with three methods which can be invoked with commandline options.

The first option is detecting the cases and uses two parameters, a set of classes of incident and activity reports and a period. The result of the detection is stored in a table of a relational database.
The second option reads the table of the detected cases and uses two options, the period and the name of the HTML file.
The third option reads the table of the detected cases and uses two options, the period and the name of the CSV file.

Detecting cases

The application uses a string with the codes of the classes of the incident and activity reports. Table E.1 show the selected classes of the dataset of domestic violence.

Table E.1 clustered BVH report classes

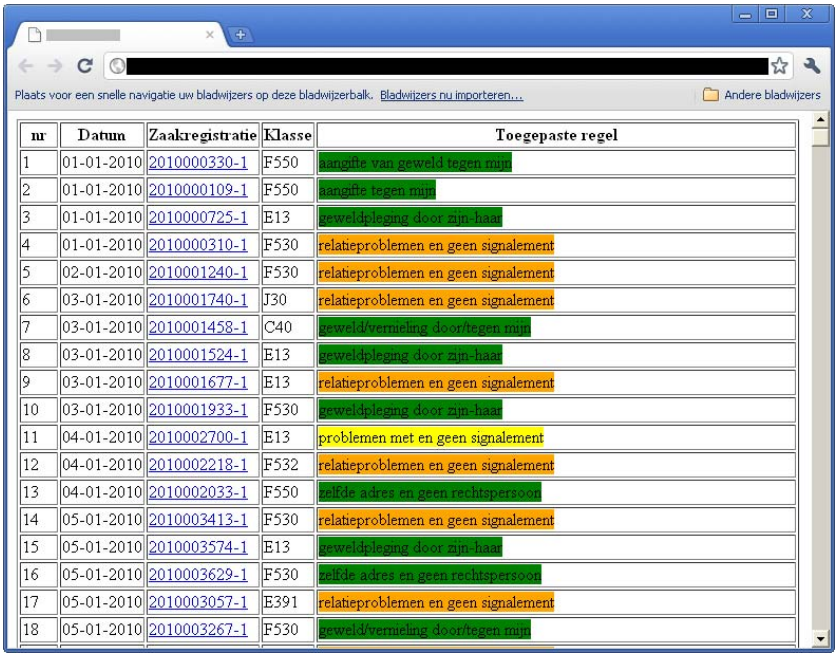
Group	**Class**	**Description**
Violence and vandalism		
	B71	Theft with violence (relational sphere)
	C40	Destruction remaining objects
	F530	Threatening
	F531	Remaining crimes against deprivation of personal liberty
	F532	Hostage/kidnap
	F540	Manslaughter/murder
	F550	Simple assault
	F551	Aggravated assault
	F552	Remaining assault
Quarrel and stalking		
	E13	Domestic quarrel (without consequences)
	E16	Quarrel (without consequences)
	E33	Nuisance by confused or overstrained person
	E391	Nuisance by stalker
General reports		
	E40	Remaining message
	J10	Attention message
	J30	General message

The result is stored in a table from the Trueblue RDBMS and is used for generating the HTML and CSV files. The stored attributes are

- recordID of the BVH system
- date and time observation of the BVH case
- the class code of the incident or activity report
- the project code
- applied rule
- the color of the rule
- detected concepts
- the URI to the web based application

Generating HTML files

The application uses a period and a filename as parameter and generates a HTML file from the table. An example of the HTML file is shown in Figure E.1



nr	Datum	Zaakregistratie	Klasse	Toegepaste regel
1	01-01-2010	2010000330-1	F550	aangifte van geweld tegen mijn
2	01-01-2010	2010000109-1	F550	aangifte tegen mijn
3	01-01-2010	2010000725-1	E13	geweldpleging door zijn-haar
4	01-01-2010	2010000310-1	F530	relatieproblemen en geen signalement
5	02-01-2010	2010001240-1	F530	relatieproblemen en geen signalement
6	03-01-2010	2010001740-1	J30	relatieproblemen en geen signalement
7	03-01-2010	2010001458-1	C40	geweld/vernieling door/tegen mijn
8	03-01-2010	2010001524-1	E13	geweldpleging door zijn-haar
9	03-01-2010	2010001677-1	E13	relatieproblemen en geen signalement
10	03-01-2010	2010001933-1	F530	geweldpleging door zijn-haar
11	04-01-2010	2010002700-1	E13	problemen met en geen signalement
12	04-01-2010	2010002218-1	F532	relatieproblemen en geen signalement
13	04-01-2010	2010002033-1	F550	zelfde adres en geen rechtspersoon
14	05-01-2010	2010003413-1	F530	relatieproblemen en geen signalement
15	05-01-2010	2010003574-1	E13	geweldpleging door zijn-haar
16	05-01-2010	2010003629-1	F530	zelfde adres en geen rechtspersoon
17	05-01-2010	2010003057-1	E391	relatieproblemen en geen signalement
18	05-01-2010	2010003267-1	F530	geweld/vernieling door/tegen mijn

Fig E.1 an example of the HTML generated file

The user can open the HTML file and select the link in the column named “zaakregistratie” or casenumber. The highlighter functionality of Cordiet is applied on the case and all searchterms of the report are highlighted. The use can examine the document more accurate than when he or she has to read the document without additional information about the applied rule

Figure E.2 shows an example of a highlighted document with the applied rule of “aangifte tegen mijn”.

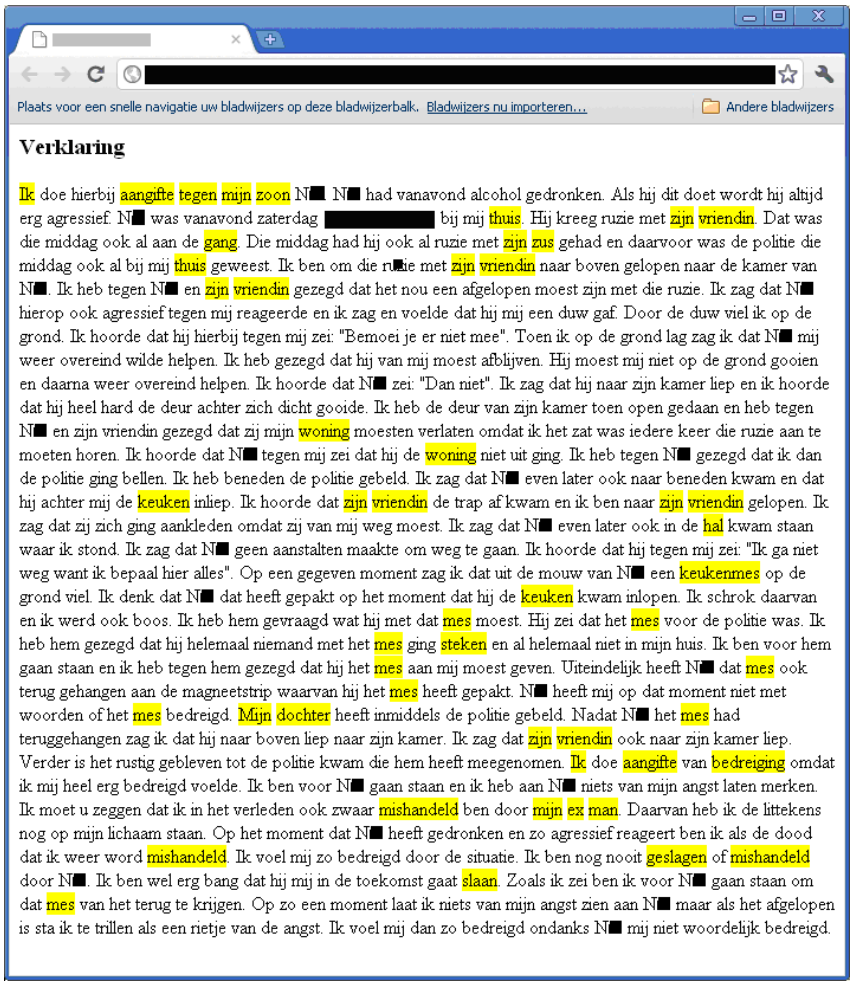
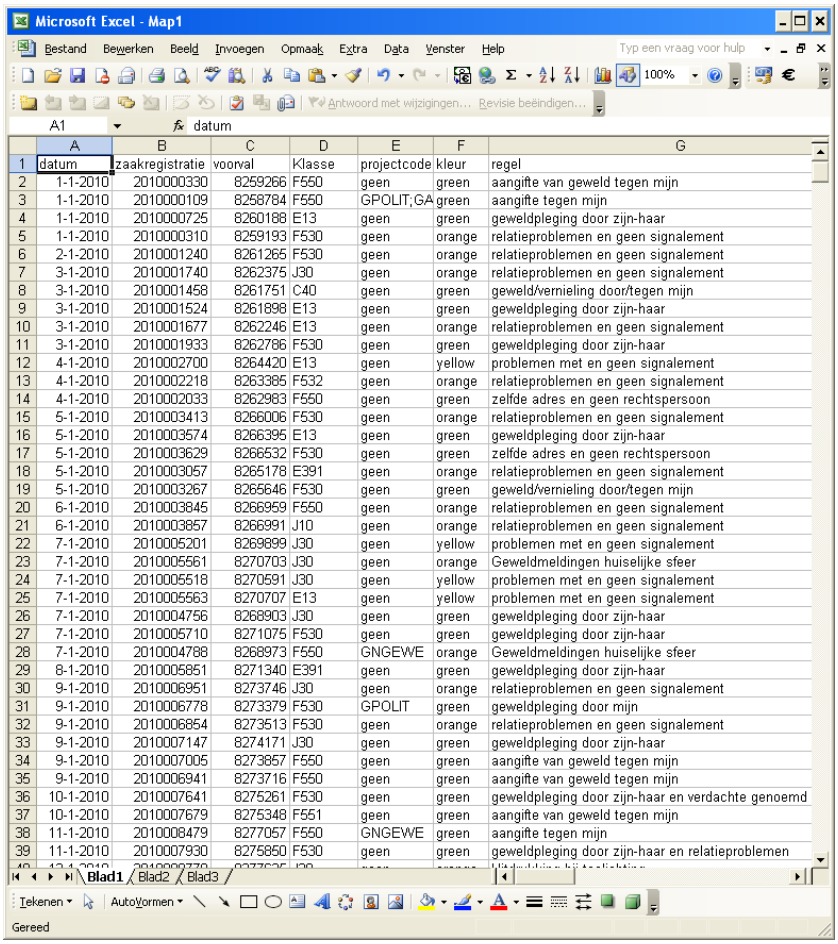


Fig E.2 An example of a highlighted report.

The example shows a statement of a mother who is threatened by her son. The beginning of the first sentence has validated the rule “aangifte tegen mijn” or legal proceedings against.

Generating CSV files

CSV files are generated for the member of the quality team to generate management statistics. An example is show below in Figure E.3.



Microsoft Excel - Map1

Bestand Bewerken Beeld Invoegen Opmaak Extra Data Venster Help

A1 datum

	A	B	C	D	E	F	G
1	datum	zaakregistratie	voorval	Klasse	projectcode	kleur	regel
2	1-1-2010	2010000330	8259266	F550	geen	green	aangifte van geweld tegen mijn
3	1-1-2010	2010000109	8258784	F550	GPOLIT;GA	green	aangifte tegen mijn
4	1-1-2010	2010000725	8260188	E13	geen	green	geweldpleging door zijn-haar
5	1-1-2010	2010000310	8259193	F530	geen	orange	relatieproblemen en geen signalement
6	2-1-2010	2010001240	8261265	F530	geen	orange	relatieproblemen en geen signalement
7	3-1-2010	2010001740	8262375	J30	geen	orange	relatieproblemen en geen signalement
8	3-1-2010	2010001458	8261751	C40	geen	green	geweld/vernieling door/tegen mijn
9	3-1-2010	2010001524	8261898	E13	geen	green	geweldpleging door zijn-haar
10	3-1-2010	2010001677	8262246	E13	geen	orange	relatieproblemen en geen signalement
11	3-1-2010	2010001933	8262786	F530	geen	green	geweldpleging door zijn-haar
12	4-1-2010	2010002700	8264420	E13	geen	yellow	problemen met en geen signalement
13	4-1-2010	2010002218	8263385	F532	geen	orange	relatieproblemen en geen signalement
14	4-1-2010	2010002033	8262983	F550	geen	green	zelfde adres en geen rechtspersoon
15	5-1-2010	2010003413	8266006	F530	geen	orange	relatieproblemen en geen signalement
16	5-1-2010	2010003574	8266395	E13	geen	green	geweldpleging door zijn-haar
17	5-1-2010	2010003629	8266532	F530	geen	green	zelfde adres en geen rechtspersoon
18	5-1-2010	2010003057	8265178	E391	geen	orange	relatieproblemen en geen signalement
19	5-1-2010	2010003267	8265646	F530	geen	green	geweld/vernieling door/tegen mijn
20	6-1-2010	2010003845	8266959	F550	geen	orange	relatieproblemen en geen signalement
21	6-1-2010	2010003857	8266991	J10	geen	orange	relatieproblemen en geen signalement
22	7-1-2010	2010005201	8269899	J30	geen	yellow	problemen met en geen signalement
23	7-1-2010	2010005561	8270703	J30	geen	orange	Geweldmeldingen huiselijke sfeer
24	7-1-2010	2010005518	8270591	J30	geen	yellow	problemen met en geen signalement
25	7-1-2010	2010005563	8270707	E13	geen	yellow	problemen met en geen signalement
26	7-1-2010	2010004756	8268903	J30	geen	green	geweldpleging door zijn-haar
27	7-1-2010	2010005710	8271075	F530	geen	green	geweldpleging door zijn-haar
28	7-1-2010	2010004788	8268973	F550	GNGEWE	orange	Geweldmeldingen huiselijke sfeer
29	8-1-2010	2010005851	8271340	E391	geen	green	geweldpleging door zijn-haar
30	9-1-2010	2010006951	8273746	J30	geen	orange	relatieproblemen en geen signalement
31	9-1-2010	2010006778	8273379	F530	GPOLIT	green	geweldpleging door mijn
32	9-1-2010	2010006854	8273513	F530	geen	orange	relatieproblemen en geen signalement
33	9-1-2010	2010007147	8274171	J30	geen	green	geweldpleging door zijn-haar
34	9-1-2010	2010007005	8273857	F550	geen	green	aangifte van geweld tegen mijn
35	9-1-2010	2010006941	8273716	F550	geen	green	aangifte van geweld tegen mijn
36	10-1-2010	2010007641	8275261	F530	geen	green	geweldpleging door zijn-haar en verdachte genoemd
37	10-1-2010	2010007679	8275348	F551	geen	green	aangifte van geweld tegen mijn
38	11-1-2010	2010008479	8277057	F550	GNGEWE	green	aangifte tegen mijn
39	11-1-2010	2010007930	8275850	F530	geen	green	geweldpleging door zijn-haar en relatieproblemen

Blad1 Blad2 Blad3

Tekenen AutoVormen

Gereed

Fig E.3 An example of the CSV generated file

The rule base

The application used a rule base which has been developed during the domestic violence case from chapter 3 of this thesis. The rule base is written in Prolog and the Prolog engine from tuProlog is used to evaluate the found concepts in the reports. The concepts are added to the rule base as facts. An excerpt of the rule base is shown in Figure E.4.

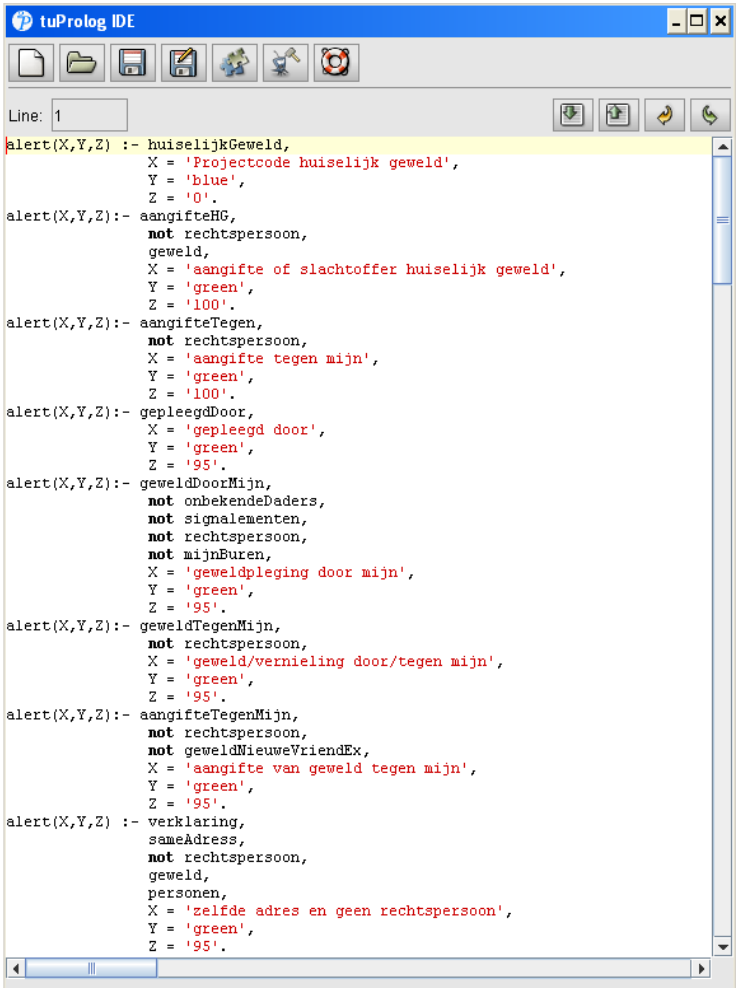


Fig E.4 An excerpt of the Domestic Violence rule base

The first rule of the rule base excludes the already labelled cases. The second and remaining rules are in order of probability being a domestic violence case. All rules in Figure E.4 have the predicate “not rechtspersoon” in the clause. This means that no cooperation may be involved in the report. If cooperation is involved, the case has a far lower probability of being a domestic violence case.
The variable’s “X”, “Y” and “Z” are returned to the rule based application whenever a set of concepts of a report satisfies the rule and stored in the Trueblue table.

APPENDIX F

Topicmap with FCA literature ontology examples

Appendix F of this thesis contains several screenshots of the web application Omnigator which can visualize the topicmap in various ways.

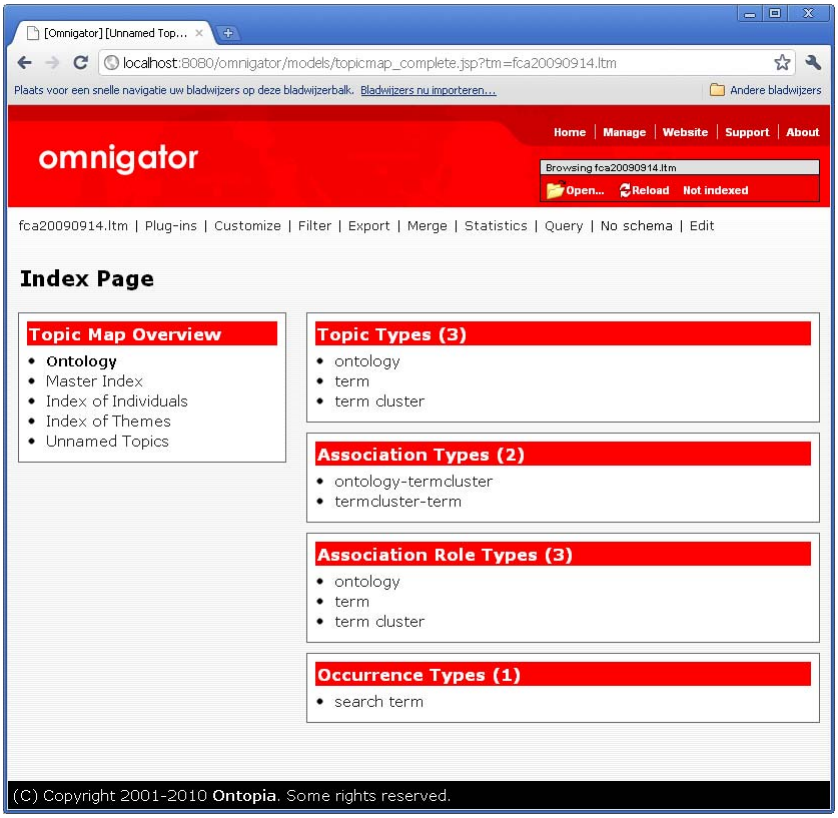


Fig F.1 the index page of the topicmap

The topicmap consists of topic types, association types, association role types and occurrence types. Because the topicmap concerns one ontology, there is also one instance of the topic type ontology, “FCA”. We select the topic type term cluster and Figure F.2 will show all instances of this topic.

Fig F.2 instances of the termclusters of the FCA ontology

The next selection we will make is showing all instances of knowledge discovery. This will result in Figure F.3

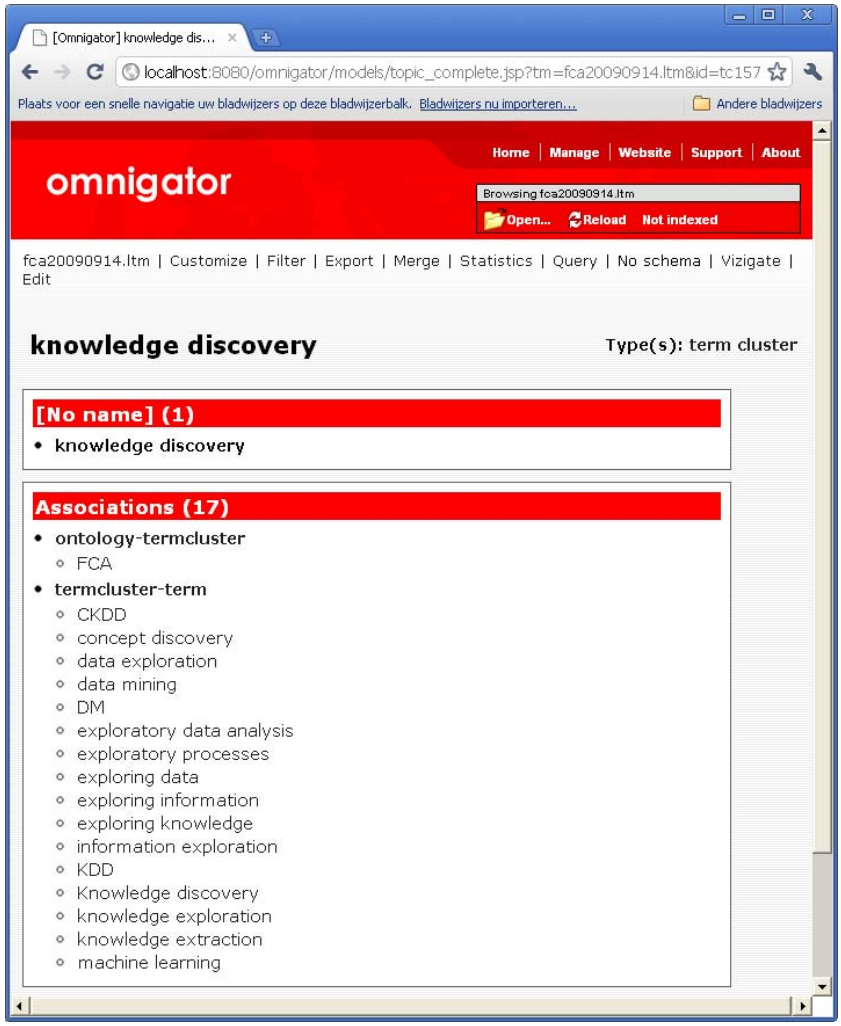


Fig F.3 instances of the termcluster “knowledge discovery”.

Figure F.3 shows two association relations, “ontology-termcluster”, which is the ontology to which “knowledge discovery” is related, and the association “termcluster-term” which shows all terms associated with “knowledge discovery”. The last selection we make is the association relation with “data mining”. This will result in Figure F.4.

Fig F.4 instance of the search terms of “data mining”

Figure F.4 shows the lowest level of the topicmap, the search terms which are used for generating the FCA input files. The term “data mining” used two search terms, “data mining” and “datamining”. The FCA literature ontology is applied on scientific papers where spelling errors are less frequent then the reports produced by police officers. Until now, the BVH system lacks an adequate spelling checker, so it is necessary to anticipate to all possible error which can be made of one word.

APPENDIX H

Human trafficking and Loverboy indicators

Human trafficking indicators

1. Dependency of the exploiter: Typically in human trafficking the housing, clothing and transportation of the woman are arranged through the exploiter, the woman will often have debts towards the exploiter and will be forced to earn the money back:
   - The woman did not arrange the travel, visa, etc. herself
   - The woman has a fake or counterfeit passport
   - The woman resides/works illegally in the Netherlands.
   - The woman fears for maltreatment and being set out of the country.
   - The woman sleeps over at the workplace
   - The woman has no proper living address in the Netherlands.
   - The woman does not know properly what her working address is.
   - The woman is socially being isolated by the exploiter.
   - The woman is in debt with a third party such as the exploiter.
   - The exploiter of the woman paid a take-over price.

2. Deprivation of liberty: Often the victim is not allowed to have contact with the outside world. They will typically not have their passports with them which are carried by the pimps. Also suspicious is when the victim cannot freely dispose of the money she earns.
   - The victim does not receive necessary medical treatment
   - The victim is not allowed to move around freely
   - The victim does not carry her own identity papers
   - The victim cannot freely dispose of her own money she earns.
   - The victim has to give an unreasonably large sum of her income to someone else

3. Being forced to work under bad circumstances:
   - The victim receives an unusually low wage compared to the market.
   - The victim works under dangerous circumstances
   - The victim works exceptionally long
   - The victim has to work under all circumstances and unreasonably long
   - The family of the victim is threatened and blackmailed
   - Indications of smuggling of single women
   - The combination of a non-European nationality, a marriage or stay with a partner and shortly after working in prostitution.
   - Relationships with persons with relevant antecedents and locations associated with human trafficking.
   - The woman is forced to earn a minimum amount of money each day
   - The woman has a slavish attitude towards exploiter

- The woman lives and/or works in buildings with internal cameras, hiding places, fake decoration, bodyguards, etc.

Violation of bodily integrity of the victim:
- Giving away organs
- Involuntarily employed in prostitution
- Threatened or confronted with violence
- Carrying traces of bodily maltreatment
- Certain things that may indicate the dependence of the exploiter such as tattoos or voodoo material.

4. Being forced to perform sexual deeds
   - Non-incidental pattern of abuse by suspect(s):
   - Working at different places from time to time
   - Tips of reliable third parties

Loverboy indicators

1. Preparatory activities to recruit girls: actual recruitment and arranging residence and shelter locations for the girls. Sometimes a girl is both a prostitute as well as a recruiter of other girls. Sometimes loverboys recruit girls for each other. During the first meeting, they estimate how vulnerable a girl is to attention and flattery. Their sensitivity to attention, presents, etc. made her fall in love with the pimp. They are not critical anymore and don't wonder where the money comes from and what the pimps intentions are.

2. Forcing her into prostitution: Pimps use a number of techniques to force the girls into prostitution:
   - Deception: They promise the girls they can keep the money or the money will be used for vacation or a house.
   - Deflowering and forcible rape: In particular Islamic girls, deflowering and the threat of being brought back home increase their anxiety to say no to the pimp's demands, because it can result in her abandonment by her family.
   - Blackmailing: If the girls don’t want to work in prostitution, the pimps threaten to bring her back to her parents.
   - Physical violence and threats: This is seen as the most effective technique to force the girl into prostitution

3. Keeping the girl in prostitution:
   - Emotional dependence: Feelings of love, nobody else to support her, the pimp is the father of her child, etc.\
   - Deception: in combination with the naivety and emotional dependence of girls.
   - Fear: the fear to be maltreated and the fear that her parents will be informed. In Islamic culture, virginity of the girl is a matter of family

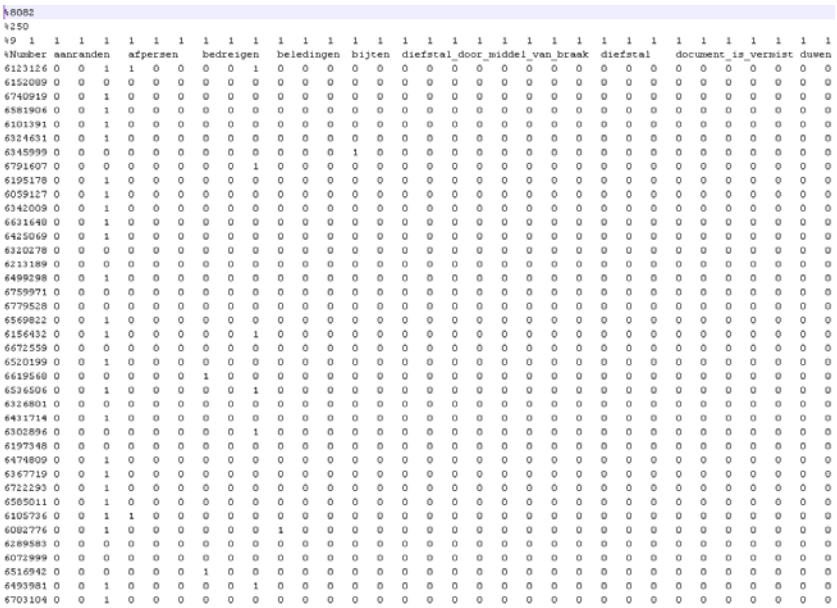
honor. If this girl is no longer virgin and the family finds out, she might not be welcome anymore.
- Social isolation: She becomes isolated from the outside world and only meets people from the prostitution circuit.
- Pride: by hiding the fact that they have to give away all the prostitution money, by acting as if they have a better life than many others, the girls justify for themselves the abuse they suffer and apparently have something they can be proud of.
- Police as an enemy: In particular under aged girls start seeing the police as their enemy.
- Competition and intermittent reinforcement: The pimp introduces competition between the girls and the girl who earns most will not be punished, but gets all attention and compliments from the pimp.

4. The pimp will also try to protect his organization:
   - Internal protection measurements: He will make sure that the girls are constantly under surveillance and with the threat of physical violence he completely dominates her life.
   - External protection: The pimp will threaten, bribe, interrogate, etc. the girls who have been in contact with the police. He may also force her to place a tattoo, to change her working address, etc. The tattoos are used in the prostitution world to trace girls who run away and are a powerful psychological instrument to make her consent to exploitation, When a girl runs away, the pimp may threaten to maltreat her or her family.

APPENDIX I

Excerpts of ESOM input files

Input file with ESOM cross table



```
%8082
%250
%9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
%Number aanranden afpersen bedreigen beledingen bijten diefstal_door_middel_van_braak diefstal document_is_vermist duwen
6123126 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6152089 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6740919 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6581906 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6101391 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6324631 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6345999 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6791607 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6195178 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6059127 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6342009 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6631640 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6425069 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6320278 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6213189 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6499298 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6759971 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6779528 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6569822 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6156432 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6672559 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6520199 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6619560 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6536506 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6326801 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6431714 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6302896 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6197348 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6474809 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6367719 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6722293 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6585011 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6105736 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6082776 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6289583 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6072999 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6516942 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6493981 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6703104 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

The first row is the number of the objects, 8082. The second row is the number of attributes, 250 (in fact 249 plus one for the name of the object). The third row is the data type of each of the attributes and the fourth row is the names of the attributes, the name of the object and the terms. White spaces are replaced by an underscore. The 8082 remaining rows are the objects with the attribute values of 0 or 1 whether a document contains the search terms associated with the term or not.

Input file with ESOM classification table

```
%8082
%1  labeled_domestic_violence   250 0   0
%2  no_labeled_domestic_violence    0   250 0
6123126 2
6152089 2
6740919 2
6581906 1
6101391 2
6324631 2
6345999 2
6791607 2
6195178 2
6059127 2
6342009 2
6631648 2
6425069 2
6320278 2
6213189 1
6499298 1
6759971 2
6779528 2
6569822 2
```

The first row is the number of objects, this should be equal to the number of objects of the cross table. The second and third rows are the classification parameters. Each object gets the value “1” and the color red (RGB 250 0 0) if the corresponding document is labelled as domestic violence and gets the value “2” and the color green (RGB 0 0 250) when it is not labelled as domestic violence. It is possible to use more than one classifier, each with separate color. The color is selected from the rule base; in this case only two rules are used.

BIBLIOGRAPHY

1 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

2 AIVD (2007), The radical dawa in transition, the rise of neoradicalism in the Netherlands. https://www.aivd.nl/aspx/download.aspx?file=/ contents/pages/90126/theradicaldawaintransition.pdf

3 Ananyan, S. (2002) Crime Pattern Analysis Through Text Mining. Proceedings of the Tenth Americas Conference on Information Systems, New York, August 2004.

4 Becker, K., Stumme, G., Wille, R., Wille, U. , Zickwolff, M. (2000) Conceptual information systems discussed through an IT-security tool. In: R. Dieng, O. Corby (eds.) Knowledge engineering and knowledge management. Methods, models and tools. Proc. EKAW. LNAI 1937, Springer, pp. 352-365

5 Beke, B.M.W.A., Bottenberg, M. (2003) De vele gezichten van huiselijk geweld. In opdracht van Programma Bureau Veilig / Gemeente Rotterdam. Uitgeverij SWP Amsterdam.

6 Black, C.M. (1999) Domestic violence: Findings from a new British Crime Survey self-completion questionnaire. London: Home Office Research Study.

7 Borg, I., Groenen, J.F. (2005) Modern multidimensional scaling: theory and applications. Springer series in statistics.

8 Bovenkerk, F., Van San, M., Boone, M., Van Solinge, T.B., Korf, D.J. (2004) “Loverboys” of modern pooierschap in Amsterdam. Willem Pompe Instituut voor Strafwetenschappen, Utrecht, December.

9 Brachman, R.J., Selfridge, P.G., Terveen, L.G., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., Mc Guinnes, D.L. and Resnick, L.A. (1993) Integrated support for data archaeology. International Journal of Intelligent and Cooperative Information Systems, 2: pp. 159-185.

10 Brachman, R., Anand, T. (1996) The process of knowledge discovery in databases: a human-centered ap-proach. In advances in knowledge discovery and data mining, U. Fayyad et al. (Eds.) AAAI/MIT Press.

11 Bullens, R.A.R., Horn, J.E. van (2000) Daad uit ‘liefde’: Gedwongen prostitutie van jonge meisjes, Tijdschrift: Justitiële verkenningen 'Jeugd en seksueel misbruik', jrg. 26, nr. 6, pag. 25-41

12 Carpineto, C, Romano, G. (1996). A lattice conceptual clustering system and its application to browsing retrieval Machine Learning, 24, 2, 1-28.

13 Carpineto, C., Romano, G. (2004) Concept data analysis: Theory and applications. John Wiley & Sons.

14 Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M., (2004) Crime data mining: a general framework and some examples. IEEE Computer, April.

15 Christopher, A. (1965) A city is not a tree. Architectural Forum, Vol 122, No 1, April, pp. 58-62 (Part I) and Vol 122, No 2, May, pp. 58-62 (Part II).

16 Cole, R., Eklund, P. (2001), Browsing Semi-structured Web Texts Using Formal Concept Analysis. H. Delugach, et al. (Eds.), Conceptual Structures: Broadening the Base, LNAI 2120, Berlin, Springer, 319-332.

17 Collier, P.M., Edwards, J.S. and Shaw, D. (2004) Communicating knowledge about police performance. International Journal of Productivity & Performance Management. Vol. 53, No. 5, pp. 458-467.

18 Carpineto, C., Romano, G. (2005) Using concept lattices for text retrieval and mining. In Formal Concept Analysis-State of the Art, Proc. of the first International Conference on Formal Concept Analysis, Berlin, Springer.

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

20 Correira, J.H., Stumme, G., Wille, R., Wille. U. (2003) Conceptual knowledge discovery - a human-centered approach. Applied artificial intelligence. 17: 281-302.

21 Cover, T., Thomas, J. (1991) Elements of information theory. New York: Wiley.

22 Dettmeijer – Vermeulen, C.E., Boot- Mattthijssen, M., Van Dijk, E.M.H., De Jonge van Ellemeet, H., Smit, H. (2008) Mensenhandel. Aanvullende kwantitatieve gegevens, Zesde rapportage van de Nationaal Rapporteur.

23 Ding, C., Peng, H.C. (2003) Minimum Redundancy feature Selection from MicroArray Gene Expression Data. Proc. Second IEEE Computational Systems Bioinformatics Conf., pp. 523-528, Aug .

24 Domingo, S., Eklund, P. (2005) Evaluation of concept lattices in a web-based mail browser. F. Dau et al. (Eds.): ICCS 2005, LNAI 3596, pp. 281–294.Springer.

25 Eidenberger, H. (2004) Visual Data Mining. Proc. SPIE Optics East Conf., Philadelphia 26-28 October. Vol. 5601, 121-132.

26 Eklund, P., Ducrou, J., Brawn, P. (2004) Concept Lattices for Information Visualization: Can Novices Read Line-Diagrams? P. Eklund (Ed.): ICFCA, LNAI 2961, 57-73. Springer.

27 Elzinga, P. (2006) Textmining by fingerprints. Onderzoeksrapport huiselijk geweld zaken. IGP project Activiteit 0504

28 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.

29 Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., Morsing, S. (2010) Terrorist treat 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.

30 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.

31 Fayyad, U., Uthurusamy, R. (2002) Evolving data mining into solutions for insights. Communications of the ACM, Vol. 45, no. 8.

32 Farley, M. Barkan, H. (1998) Prostitution, violence, and post-traumatic stress disorder. Women & Health 27(3):37-49

33 Ganter, B., Wille, R. (1999) Formal Concept Analysis. Mathematical foundations. Springer.

34 Ganter, B. and Wille, R. 1999a. Contextual Attribute Logic. In Proceedings of the 7th international Conference on Conceptual Structures: Standards and Practices (July 12 - 15, 1999). W. M. Tepfenhart and W. R. Cyre, Eds. Lecture Notes In Computer Science, vol. 1640. Springer-Verlag, London, 377-388.

35 Gill, P. (2000) Rounding up the usual suspects? Developments in Contemporary Law Enforcement Intelligence. Aldershot: Ashgate.

36 Godin, R., Gescei, J., Pichet, C. (1989), Design of browsing interface for information retrieval. N.J.Belkin et al. (Eds.), Proc. SIGIR ’89, 32-39.

37 Godin, R., Missaoui, R., April, A. (1993) Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. Int. J. Man-Machine Studies 38, 747-767.

38 Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325–328.

39 Gruber, T. (2009) Ontology. Encyclopedia of Database Systems. L. Liu et al. (Eds.), Springer.

40 Hatchuel, A. (1996) Les theories de la conception, Paris

41 Hatchuel, A., Weil, B. (1999) Pour une théorie unifiéee de la conception. Axiomatiques et processus collectifs, 1-27 CGS Ecole des Mines/GIS cognition – CNRS, Paris.

42 Hatchuel, A., Weil, B. (2002) La théorie C-K: fondements et usages d'une théorie unifiée de la conception. Proceedings of Colloque sciences de la conception, Lyon, 15-16 mars.

43 Hatchuel, A., Weil, B. (2003) A new approach of innovative design: an introduction to C – K theory. Proceedings of ICED'03, august 2003, Stockholm, Sweden, pp. 14.

44 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.

45 Hereth, J., Stumme, G., Wille, U., Wille, R. (2000) Conceptual knowledge discovery and data analysis. In: B. Ganter, G. Mineau (eds.) Conceptual structures: logical, linguistic and computational structures. Proc. ICCS. LNAI 1867, Springer, pp. 421-437.

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

47 Hollywood, J., Strom, K., Pope, M. (2009) Can Data Mining Turn Up Terrorists? OR/MS Today – February.

48 Hsu, C.W., Lin, C.J. (2002) A comparison of methods for Multi-Class Support Vector Machines. IEEE Trans. Neural Networks, vol. 13 pp. 415-425.

49 Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York.

50 IALEIA (2004) Law Enforcement Analytic Standards. Richmond, VA: Global Justice Information Sharing Initiative.

51 Keim, D.A. (2002) Information visualization and visual data mining. IEEE transactions on visualisation and computer graphics. Vol. 8, No. 1.

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

53 Kohonen, T. (1982) Self-Organized formation of topologically correct feature map, Biological Cybernetics, Vol. 43, pp. 59-69.

54 Kruskal, J. B., and Wish, M. (1978). Multidimensional Scaling, Sage University Paper series on Quantitative Application in the Social Sciences. Beverly Hills and London: Sage Publications.

55 Lakhal, L., Stumme, G. (2005) Efficient Mining of Association Rules Based on Formal Concept Analysis. B. Ganter et al. (Eds.): Formal Concept Analysis, LNAI 3626, 180-195. Springer.

56 Magers, J. S. (2004) Compstat: A new paradigm for policing or a repudiation of community policing? Journal of Contemporary Criminal Justice. 20 (1): 70-79

57 Manning, C.D., Raghavan, P., Schütze, H. (2008) Introduction to Information Retrieval. Cambridge University Press.

58 Marchionini, G. (2006) Exploratory search: from finding to understanding. Communications of the ACM, Vol. 49, no. 4

59 Matlab Arsenal (2009) http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenalDoc/MATLABArsenal/NeuralNet.html

60 Mili, H., Ah-Ki, E., Godin, R., Mcheick, H. (1997). Another nail to the coffin of faceted controlled-vocabulary component classification and retrieval. VCM SIGSOFT Software Engineering Notes, 22, 3, 89-98.

61 Ministerie van Justitie (2009) http://www.justitie.nl/onderwerpen/criminalitiet/mensenhandel/, retrieved on 22-08-2009.

62 NCTB (2008), Salafism in the Netherlands.http://english.nctb.nl/Images/Salafisme%20UK_tcm92-132297.pdf?cp=92&cs=25496

63 O'Neill, RA. (1999) International trafficking to the United States: a contemporary manifestation of slavery and organized crime- and intelligence monograph. Washington DC; Exceptional Intelligence Analyst Program.

64 Office on Violence against Women (2007) About Domestic Violence (http:www.usdoj.gov/ovw/domviolence.htm). Retrieved on 2007-10-22

65 Pednault, E.P.D. (2000) Representation is everything. Communications of the ACM, Vol. 43, no. 8.

66 Peng, H. Long, F. Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency. Max-Relevance, and Min-Redundancy. IEEE Transactions on pattern analysis and machine intelligence, Vol. 27, no. 8.

67 Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2008). An exploration into the power of formal concept analysis for domestic violence analysis. In: Lecture Notes in Computer Science, 5077. Industrial Conference on Data Mining ICDM. Leipzig (Germany), 16-18 July 2008 (pp. 404-416). Springer.

68 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: LNAI, Vol. 5633(XI), (Perner, P. (Eds.)). Industrial conference on data mining ICDM 2009. Leipzig (Germany), 20-22 July 2009 (pp. 402 p.).

69 Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., Dedene, G. (2009a). How emergent self organizing maps can help counter domestic violence. World Congress on Computer Science and Information Engineering (CSIE). Los Angeles (USA), 31 March - 2 April 2009.

70 Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Van Hulle, M. (2009b). Analyzing domestic violence with topographic maps: a comparative study. 7th International Workshop on Self-Organizing Maps (WSOM). St. Augustine, Florida (USA), 8-10 June 2009.

71 Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., Dedene, G. (2009c). Gaining insight in domestic violence with emergent self organizing maps. Expert systems with applications, 36(9), 11864-11874

72 Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2009d). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. 9th Industrial Conf. on Data Mining, LNCS, 5633, 247-260, Leipzig, Germany, July 20-22. Springer.

73 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, 167-191. Wiley and Sons, Ltd. Doi 10.1002/isaf.319.

74 Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010b) Formally Analyzing the Concepts of Domestic Violence, Expert Systems with Applications 38, 3116-3130. Elsevier Ltd. doi 10.1016/j.eswa. 2010.08.103 . [SCI 2009 = 2.908]

75 Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010c). A method based on Temporal Concept Analysis for detecting and profiling human trafficking suspects. Proc. IASTED International Conference on Artificial Intelligence. Insbruck, Austria, 15-17 february.

76 Poelmans, J., Dedene, G., Verheyden, G., Van der Mussele, H., Viaene, S., Peters, E. (2010d). 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

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

78 Politie Amsterdam-Amstelland (2009) http://www.politie-amsterdam-amstelland.nl/get.cfm?id=86, retrieved on 22-06-2009.

79 Priss, U. (1997) A Graphical Interface for Document Retrieval Based on Formal Concept Analysis. E. Santos (Ed.): Proc. of the 8th Midwest Artificial Intelligence and Cognitive Science Conf.. AAAI Technical Report CF-97-01, 66-70.

80 Priss, U. (2000), Lattice-based information Retrieval. Knowledge Organization, 27, 3, 132-142.

81 Priss, U., Old, L.J. (2005) Conceptual Exploration of Semantic Mirrors. B. Ganter et al. (Eds.): ICFCA, LNAI 3403, 21-32. Springer.

82 Priss, U. (2006) Formal Concept Analysis in Information Science. C. Blaise (Ed.): Annual Review of Information Science and Technology, ASIST, Vol. 40.

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

84 Ratcliffe, J. (2008) Intelligence-Led Policing. Collumpton, UK Willan Publishing.

85 Ritter, H. (1999) Non-Euclidean Self-Organizing Maps, pp. 97–109. Elsevier, Amsterdam.

86 Scheich, P., Skorsky, M., Vogt, F., Wachter, C., Wille, R. (1993) Conceptual data systems. In: O. Opitz, B. Lausen, R. Klar (eds.) Information and classification. Springer Berlin Heidelberg, pp. 72-84.

87 Skorsky, M. (1997). Graphische Darstellung eines Thesaurus. Deutscher Dokumentartag, Regensburg.

88 Smyth, P., Pregibon, D., Faloutsos, C. (2002) Data-driven evolution of data mining algorithms. Communications of the ACM, Vol. 45, no. 8.

89 Stumme, G., Wille, R., Wille, U. (1998) Conceptual knowledge discovery in databases using Formal Concept Analysis Methods. PKDD, 450-458.

90 Stumme, G., Wille, R. (eds.) (2000) Begriffliche Wissenverarbeitung – methoden und anwendungen. Springer Heidelberg.

91 Stumme, G. (2002) Efficient Data Mining Based on Formal Concept Analysis. A. Hameurlain et al. (Eds.): DEXA. LNCS, vol. 2453. Springer.

92 Stumme, G., Bestride, Y., Taouil, R., Lakhal, L. (2002a) Computing Iceberg Concept Lattices with TITANIC. Data and knowledge engineering 42 (2), 189-222.

93 Stumme, G. (2002b), Formal Concept Analysis on its Way from Mathematics to Computer Science. Proc. 10th Intl. Conf. on Conceptual Structures (ICCS 2002). LNCS, Springer, Heidelberg.

94 Stumme, G. (2003) Off to new shores: conceptual knowledge discovery and processing. Int. J. Human-Computer Studies 59, 287-325. Elsevier.

95 Thomas, J., Cook, K. (2005) Illuminating the path: research and development agenda for visual analytics. National Visualization and Analytics Ctr. IEEE.

96 Tilley, N. (2003) Community policing, problem-oriented policing and intelligence-led policing. Newburn, T. (Ed.). Handbook of policing. Collumpton: Willan Publishing, pp. 311-339.

97 Tilley, T. (2004) Tool support for FCA. P. Eklund (Ed.): ICFCA, LNAI 2961, 104-111. Springer.

98 Tilley, T., Eklund, P. (2007) Citation analysis using Formal Concept Analysis: A case study in software engineering. 18th int. conf. on database and expert systems applications (DEXA).

99 Tyldum, G., Brunowskis, A. (2005) Describing the unobserved: Methodological challenges in empirical studies on human trafficking. Data and research on human trafficking: A global survey. International Organization on Migration (IOM).

100 Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp. 305-308.

101 Ultsch, A. (1999) Data mining and knowledge discovery with Emergent SOFMS for multivariate Time Series.

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

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

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

105 Ultsch, A., Moerchen, F. (2005b) ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46.

106 United Nations General Assembly (2001) Protocol to prevent, suppress and punish trafficking in persons, especially women and children, supplementing the UN convention against Transnational Organized Crime (UN doc A/ 45/49, vol. 1). Palermo, UN.

107 United Nations, Economic and social council (2004) Economic causes of trafficking in women in the Unece region. Regional Preparatory Meeting for the 10-year review of implementation of the Beijing Platform for Action, 14-15 December.

108 Valtchev, P., Missaoui, R., Godin, R. (2004) Formal Concept Analysis for Knowledge Discovery and Data Mining: The New Challenges. P. Eklund (Ed.): ICFCA, LNAI 2961, 352-371. Springer.

109 Van Dijk, T. (1997) Huiselijk geweld, aard, omvang en hulpverlening (Ministerie van Justitie, Dienst Preventie, Jeugd-bescherming en Reclassering, oktober).

110 Van der Veer, R.C.P., Roos, H.T., Van der Zanden, A. (2009), Datamining for intelligence led policing, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris June 28th – July 1st.

111 Vapnik, V. (1995) The Nature of Statistical Learning Theory. New York: Springer.

112 Viaene S., De Hertogh S., Lutin L., Maandag A., den Hengst S., Doeleman R. (2009). Intelligence-led policing at the Amsterdam-Amstelland police department: operationalized business intelligence with an enterprise ambition. Intelligent systems in accounting, finance and management. 16 (4) : 279 -292

113 Vincent, J.P., Jouriles, E.N. (2000) Domestic violence. Guidelines for research-informed practice. Jessica Kingsley Publishers Londen and Philadelphia

114 Waits, K. (1985). The criminal Justice System's response to Battering: Understanding the problem, forging the solutions. Washington Law Review 60: pp. 267-330.

115 Walsh, W. F. (2001) Compstat: an analysis of an emerging police managerial paradigm. Policing: An International Journal of Police Strategies and Management, 24(3):347-362.

116 Watts, C., Timmerman, C. (2002) Violence against women: global scope and magnitude. The Lancet 359 (9313): pp.1232-1237. RMID 1155557

117 Webb, A.R. (1999) Statistical pattern recognition. Arnold.

118 Wehrens, R., and Buydens, L. M. (2007). Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software, 21 (5), 1–19.

119 Weisburd, D. and Eck, J. (2004) What can the police do to reduce crime, disorder and fear? Annals of the American Academy of Political and Social Science, 593 (1): 43-65.

120 Weka (2009) http://www.cs.waikato.ac.nz/ml/weka/

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

122 Wille, R. (1997) Introduction to formal concept analysis. G. Xesrlni (Ed.): Modelli e model Uzzazione. Models and modeling. Consiglio Xazionale delle Ricerche, Instituto di Studi sulli Ricerca e Documentazione Scientifica. Roma, 39-51.

123 Wille, R. (2002), Why can concept lattices support knowledge discovery in databases?, Journal of Experimental & Theoretical Artificial Intelligence, 14: 2, pp. 81-92.

124 Wille, R. (2005), Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. . B. Ganter et al. (Eds.): Formal Concept Analysis, LNAI 3626, 1-33. Springer

125 Wille, R. (2006) Methods of conceptual knowledge processing. R. Missouri et al. (Eds.): ICFCA, LNAI 3874, 1-29. Springer.

126 Wolff, K. E. (1994). A first course in formal concept analysis – how to understand line diagrams. F. Faulbaum (Ed.), SoftStat93, Advances in statistical software, Vol. 4, 429-438. Gustav Fischer.

127 Wolff, K.E. (2002) Transitions in conceptual time systems. In: D.M. Dubois (Ed.): Int. J. of Computing Anticipatory Systems, vol. 11, CHAOS, 398-412.

128 Wolff, K.E., Yameogo, W. (2003) Time dimension, Objects and life tracks – A conceptual analysis. A. De Moor et al. (Eds.) Conceptual structures for knowledge creation and communication. LNAI 2746, 188-200, Springer.

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

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