An exploration into the power of Formal Concept Analysis for domestic violence analysis

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

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

Abstract. The types of police inquiries performed are very diverse in nature and the current data processing architecture is not sufficiently tailored to cope with this diversity. Many information concerning cases is still stored in databases as unstructured text. Formal Concept Analysis is showcased as an exploratory data analysis technique for discovering new knowledge from police reports. It turns out that it provides a powerful framework for exploring the dataset, resulting in essential knowledge for improving current practices. It is shown that the domestic violence definition employed by the police organisation of the Netherlands is not always as clear as it should be, making it hard to use it effectively for classification purposes. In addition, newly discovered knowledge for automatically classifying certain cases as either domestic or non-domestic violence is presented. Moreover, essential techniques for detecting incorrect classifications, performed by police officers, are provided. Finally, some problems encountered because of the sometimes unstructured way of working of police officers are discussed. Both using Formal Concept Analysis for exploratory data analysis and its application on this area are novel enough to make this paper into a valuable contribution to the literature.

Keywords: Formal Concept Analysis (FCA), domestic violence, knowledge discovery in databases, data mining

1 Introduction

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" [1]. 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 to the

victim, the self or others including children. Domestic violence occurs all over the world, in various cultures [2] and affects people across society, irrespective of economic status [3].

Domestic violence is one of the top priorities of the police organization of the region Amsterdam-Amstelland in The Netherlands. 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 have established that many reports tended to be wrongly classified as domestic or as non-domestic violence cases. One of the conclusions was that there was a need for an in-depth investigation of this problem area.

In this paper, it shall be demonstrated that from the unstructured text in police reports, essential knowledge regarding domestic violence can be obtained by using a technique known as Formal Concept Analysis (FCA) [8, 9]. FCA arose twenty-five years ago as a mathematical theory [14]. It has over the years grown into a powerful framework for data analysis, data visualization, [10, 15, 18], information retrieval and text mining [16, 17, 20]. However, FCA has never been used for exploratory data analysis, which is one of the core contributions of this paper. What makes FCA into an especially appealing knowledge discovery in databases technique from a practitioner point of view is the compactness of its information representation and the minimal need for users to tune (hyper-) parameters to distill a useful, actionable picture of the mining exercise.

The remainder of this paper is composed as follows. In section 2, we shall cover the essentials of FCA theory, introducing the pivotal FCA notions of concept and concept lattice. In section 3, the dataset used in our research will be elaborated on. Section 4 then showcases and discusses the results of the application of FCA for exploratory analysis of domestic violence cases using this data set. Finally, section 5 rounds up with conclusions.

2 FCA essentials

This section introduces the main ideas of Formal Concept Analysis in a very elementary way.

2.1 Concepts and lattices

The starting point of the analysis is a database table consisting of rows (i.e. objects), columns (i.e. attributes) and crosses (i.e. relationships between objects and attributes). The mathematical structure used to reference such a cross table is called a formal context. An example of a cross table is displayed in table 1. In the latter, reports of domestic violence (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 the term. The dataset in table 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 ac-

cording to a "subconcept-superconcept" relation. This results in a line diagram (a.k.a. lattice).

Table 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	

The notion of "concept" is central to FCA. 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 1. Take the attributes that describe report 5, for example. By collecting all reports of this context that share these attributes, we get to a set O consisting of reports 2, 3 and 5. This set O of objects is closely connected to set A consisting of the attributes "cursing" and "scratching." 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 O is called the extent, while A is called the intent 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. A concept d is called a subconcept of a concept e (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 "cursing," "scratching" and "stabbing" is a subconcept of a concept with intent "cursing" and "scratching." With reference to table 1, the extent of the latter is composed of reports 2 and 3, while the extent of the former is composed of reports 2, 3 and 5.

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 1, for example, represents the concept lattice of the formal context abstracted from table 1. 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 the node represent the objects used to name the concept. The information contained in the formal context of table 1 can be distilled from the line diagram in figure 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, report 5 is described by the attributes "cursing" and "scratching".

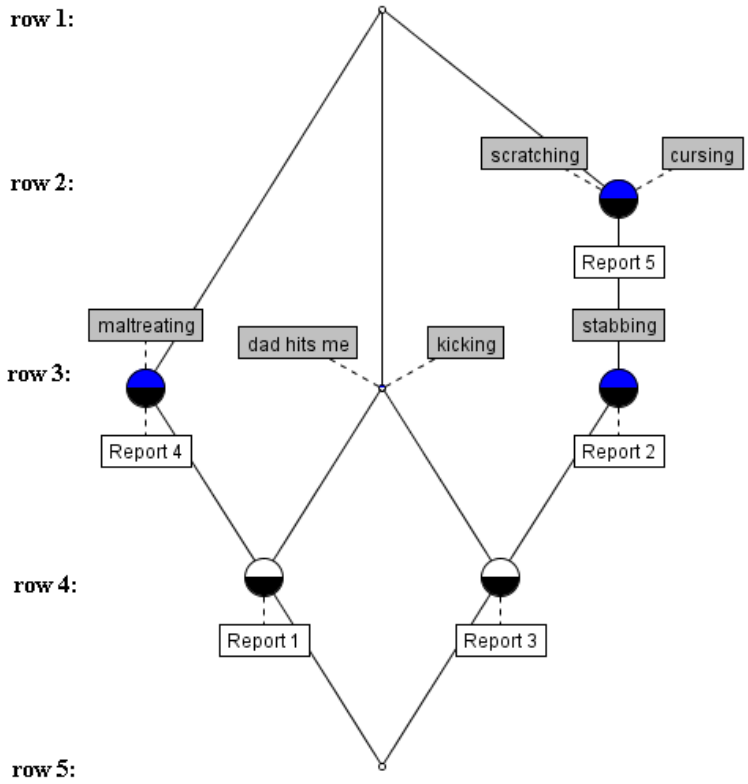

Fig. 1. Line diagram corresponding to the context from table 1

Retrieving the extension of a formal concept from a line diagram such as the one in figure 1 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 three 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 three 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. 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. 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.

2.2 FCA software

We used FCA as an unsupervised clustering technique [11, 13]. Police reports containing terms from the same term-clusters were grouped together in concepts. The aim was to make these concepts as pure as possible. When a concept contained domestic and non-domestic violence reports, we investigated these reports and searched them for new attributes that can be used to discriminate between the domestic and non-domestic violence reports from this concept. This process was repeated until a classification was obtained that minimizes the number of false negative cases (i.e. domestic violence cases that were not classified as such).

The tool Concept Explorer [7] was used to visualize the concepts and their relationships.

3 Data set

The domestic violence definition employed by the police organization of the Netherlands is as follows: "Domestic violence can be characterized as serious *acts of violence* committed by *someone of 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. Family friends are those persons who have a friendly relationship with the victim and who (regularly) meet the victim in his/her home", [6].

The XPol database – the database of the Amsterdam police organization – contains all documents relating to criminal offences. Documents related to certain types of crimes receive corresponding labels. 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 a domestic violence case, they can indicate this by assigning the project code "domestic violence" to the report. However, not all domestic violence cases are recognized as such by police officers and by consequence, many documents are wrongly lacking a "domestic violence" label. The in-place case triage system is used to filter out these reports for in-depth manual inspection and classification. For example, going back to the first quarter of 2006, the in-place triage system retrieved 367 of such cases.

The dataset used during the research consists of 4146 police reports describing all violent incidents from the first quarter of 2006. All domestic violence cases from that period are a subset of this dataset. The 367 cases selected by the in-place case triage system are also a subset of this dataset. Unfortunately, many of these 4146 police reports did not contain the reporting of a crime by a victim, which is necessary for establishing domestic violence. Therefore, we only retained the 2288 documents in which the victim reported a crime to a police officer. From these documents, we removed the follow-up reports referring to previous cases. This filtering process resulted in a set of 1794 reports. From these reports, we 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. These data were used to generate the 1794 html-documents that were used for our research. An example of such a report is displayed in figure 2.

Title of incident	Violent incident xxx
Reporting date	26-11-2007
Project code	Domestic violence against seniors (+55)
Crime location	Amsterdam Keizersgracht yyy
Suspect (male) Suspect (18-45yr)	zzz
Address	Amsterdam Keizersgracht yyy
Involved (male) Involved (18-45yr)	Neighbours
Address	Amsterdam Keizersgracht www
Victim (male) Victim (>45jr)	uuu
Address	Amsterdam Keizersgracht vvv

Reporting of the crime

Last night I was attacked by my only son. I was watching television in the living room when he suddenly attacked me with a knife. I felt on the floor. Then he tried to kick me. I tried to escape through the back door while I was yelling for help. I ran to the neighbours for help. They called the emergency services. Meanwhile my son ran away. My leg was bleeding etc.

Fig. 2. Example police report

We also have at our disposal a thesaurus – a collection of terms – that was obtained by performing frequency analyses on these police reports. The terms that occurred most often were retrieved and added to the initially empty thesaurus. This resulted in a set of 123 terms.

Our validation set consists of 9147 cases describing all violent incidents from the year 2005 where the victim made a statement to the police. After removing the follow-up reports, 7817 cases were retained. In 2005, the in-place case triage system retrieved 2668 documents that had to be manually classified by police-officers. 1526 of them were classified as domestic violence, while 1142 of them were classified as non-domestic violence. These documents are a subset of our validation set.

We intend to verify whether a report can be classified as domestic violence by checking that it contains one or more terms from each of the two components of the domestic violence definition. A case can be labelled as domestic violence if:

1. a criminal offence has occurred. This may range from verbal threats over pushing and kicking to even killing the victim. To verify whether a criminal offence has occurred, the report is searched from terms like “hit”,

"stab", "kick", etc. These terms are grouped into the term-cluster "acts of violence".

2. and a person of the domestic sphere of the victim is involved in the crime. 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" to the persons involved in the crime. Therefore, the report is searched for terms like "my dad", "my mom", "my son", etc. These terms are grouped into the term-cluster "family members". The report is also searched for terms like "my ex-boyfriend", "my ex-husband", "my ex-wife", etc. These terms are grouped under the term-cluster "ex-partners". Furthermore, the report is searched for terms like "my nephew", "her uncle", "my aunt", "my step-father", "his step-daughter", etc. These terms are grouped under the term-cluster "relatives". Then the report is searched for terms like "family friend", "co-occupant", etc. These terms are grouped under the term-cluster "family friends".

The reports having the attribute "domestic violence" were classified by police officers as domestic violence. The remaining reports were classified as non-domestic violence. This results in the following lattice:

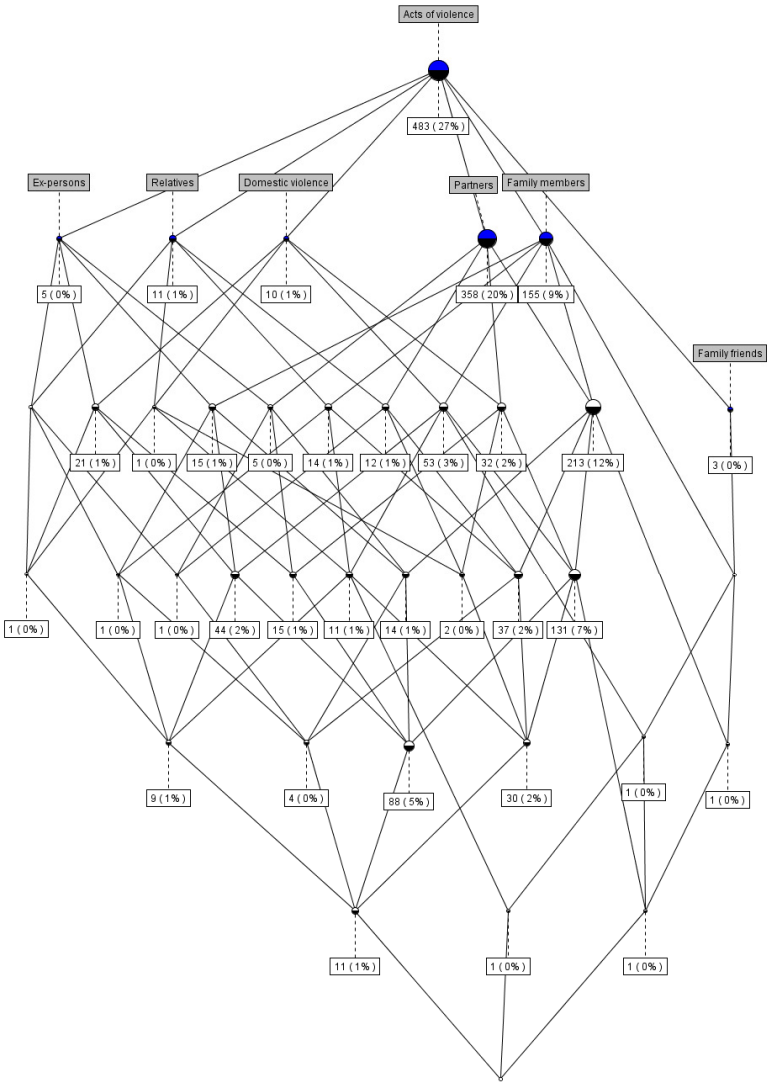

Fig. 3. Initial lattice based on the police reports from the first quarter of 2006

4 Results

By analysing the lattice displayed in figure 3, it became clear that a lattice containing only term-clusters based on the domestic violence definition does not sufficiently discriminate domestic from non-domestic violence reports. In other words, the definition cannot be used to automatically classify domestic violence cases as such, because many non-domestic violence also contain terms belonging to one or more of these clusters. However, the following interesting knowledge emerged out of this lattice.

Table 2. Results from lattice in figure 3

	Non-domestic violence	Domestic violence
Acts of violence	483	10

Some 36% of the non-domestic violence reports only contain terms from the "acts of violence" cluster, while there are only 10 such domestic violence reports in the dataset. After in-depth manual inspection, these reports turned out to be wrongly classified as domestic violence. Although the lattice can not be used to distinguish domestic violence reports from non-domestic violence reports, it can be used to detect cases that were wrongly classified as domestic violence. Therefore, this lattice was constructed for all domestic violence reports from the year 2005. 116 (5%) of the cases that were classified as domestic violence only contained one or more terms from the "acts of violence" cluster, while there was no person of the domestic sphere of the victim involved. In-depth manual inspection of these reports proves that they were almost all wrongly classified as domestic violence. Furthermore, 49 (2%) of the domestic violence cases did not describe a violent incident. They were also wrongly classified as domestic violence. 7% of the domestic violence cases were thus reclassified as non-domestic violence. By using this lattice, it was also possible to discover the most frequently occurring types of domestic violence cases from the year 2005.

Table 3. Most frequently occurring types of domestic violence in 2005

	% of all domestic violence cases of 2005
Acts of violence and family members and partners	27%
Acts of violence and family members and partners and ex-persons	15%
Acts of violence and family members	13%
Acts of violence and family members and ex-persons	12%
Acts of violence and family members and partners and relatives	6%
Acts of violence and partners	6%

The next step consisted of searching for additional attributes that can be used to distinguish a domestic violence report from non-domestic violence reports and vice versa. First of all, it became apparent that in a large number of the domestic violence cases (137 cases or 28%), the perpetrator and the victim lived at the same address at

the time the victim made his/her statement to the police. 128 of these cases were classified as domestic violence. When we studied the 9 non-domestic violence cases we found that the perpetrator and the victim always lived together in the same institution (e.g. a youth institution, a prison, an old folk's home). Of the 21 cases where the perpetrator and the victim lived in the same institution, only 12 were classified as domestic violence. This finding brought about a lively discussion amongst the police officers of the Amsterdam police force. More importantly, it exposed the mismatch between the management's conception of domestic violence and the classification as performed by the police officers. Going back to 2005, 138 cases described incidents between inhabitants of an institution. Police officers classified a substantial part of these as non-domestic violence, while only a limited number were classified as domestic violence. However, according to the board members responsible for the domestic violence policy, all these cases should have been classified as domestic violence. In other words, the definition employed by the management was much broader than the one employed by the police officers performing the classification task.

To classify the remaining cases, we explored the corresponding police reports, in search of new attributes. We found that 42% of these reports (749 cases) did not mention a suspect. However, according to the domestic violence definition (which says that the perpetrator must belong to the domestic sphere of the victim), the offender should be known. By consequence, we assumed that these reports described non-domestic violence cases. However, 30 of them turned out to be domestic violence cases. After in-depth inspection of these reports, we concluded that this was due to unstructured way of working of the police officers. Some officers immediately label a person as a suspect, when the victim mentions this person as a suspect. However, other officers first want to interrogate the suspect. In the latter case, this person is added to the list of persons who were involved in or witnessing the crime. This list of persons might include friends and family members of the victim, bystanders, etc. and can be very large. By consequence it is often very difficult to identify the suspect from this list in filed reports. We asked the proper authorities whether or not there exists a policy that regulates the labelling of persons as suspects. It turned out that such a regulation did not exist. Our research proves that such a regulation is necessary. We also found that a 37% of these reports that lack a suspect contain a description of the suspect (277 cases). These 277 reports were all classified as non-domestic violence.

After the consultation of the proper authorities with regard to this subject, it became clear that the best and the most feasible solution would be to introduce an additional field in police reports that can be used by police officers to record the person who was mentioned by the victim as the offender. This relatively small change makes it easier to identify the suspect(s) for a given case.

According to the literature, domestic violence is a phenomenon that mainly occurs inside the house [4, 5, 6, 21]. Therefore, an attribute called "private locations" was introduced. This term-cluster contained terms like "bathroom", "living room", "bedroom", etc. As was expected, 393 (86%) of the domestic violence cases from the dataset contained one or more terms from this term-cluster. However, 568 (43%) of the non-domestic violence cases also contained one or more terms from this term-cluster. An attribute called "public locations" was also introduced. It was expected that there would be almost no domestic violence case that took place on the street. Surprisingly, this turned out to be incorrect. In about one-fourth of the domestic violence cases

there had been an incident on a public location. While studying these police reports, we discovered that this was often the case when ex-partners were involved in the case. It thus became clear that it is 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 like “family members” or “ex-persons” for example did not yield the expected results either. While exploring the domestic violence reports, terms like “divorce”, “marriage problems”, “relational problems”, etc. regularly occurred. Therefore, a new term-cluster called “relational problems” was introduced.

In order to keep the lattice surveyable, we clustered the terms from the clusters “family members”, “relatives”, “partners”, “ex-partners” and “family friends”, together in the cluster “persons” and added the newly discovered attributes “same address”, “institution”, “no suspect”, “description of suspect” and “relational problems”. This resulted in the following lattice:

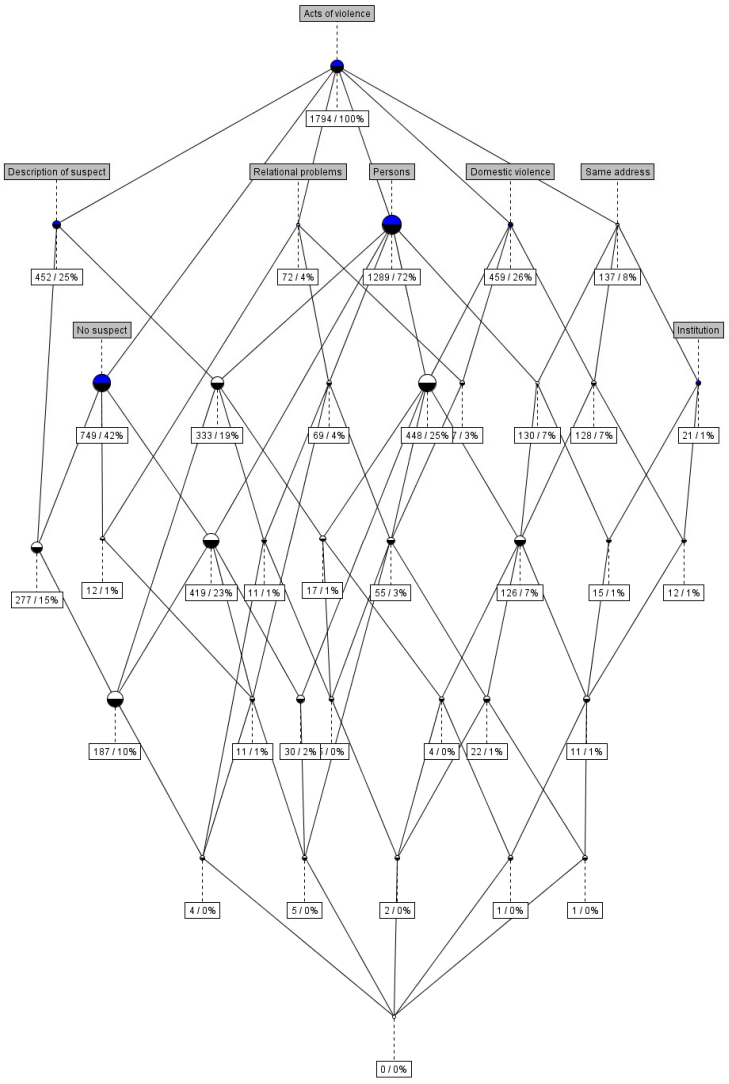

Fig. 4. Refined lattice based on the police reports from the first quarter of 2006

This lattice provides us with essential knowledge needed to discriminate domestic from non-domestic violence reports and vice versa. The most interesting findings are displayed in the following table:

Table 4. Results from lattice in figure 4

	Non-domestic violence	Domestic violence
Acts of violence	483	10
Acts of violence and same address	9	128
Acts of violence and no suspect and description of suspect	277	0
Acts of violence and no suspect	719	30

This lattice was also constructed for the validation set containing police reports from the year 2005. The most interesting findings are displayed in the following table.

Table 5. Discovered knowledge applied on dataset from 2005

	Non-domestic violence	Domestic violence
Acts of violence	2065	116
Acts of violence and same address	56	422
Acts of violence and no suspect and description of suspect	937	27
Acts of violence and no suspect	2579	186

The 56 cases where the perpetrator and the victim lived at the same address and that were not classified as domestic violence contained many cases where the perpetrator and the victim lived in the same institution. The remaining cases turned out to be wrongly classified after in-depth manual inspection. After in-depth manual inspection of the 116 police reports containing only one or more terms from the "acts of violence" cluster and that were classified as domestic violence, they turned out to be wrongly classified.

5 Conclusions

In this paper, the possibilities of using FCA as an exploratory data analysis technique for discovering new knowledge from police reports were explored. The construction of an initial lattice containing term-clusters created by a domain expert on the basis of the domestic violence definition and the incremental refinement of this lattice provided the user with a powerful framework for exploring the dataset.

It was shown that the domestic violence definition is often not applied properly by police officers, but incorrect classifications can be automatically corrected on the basis of this definition. In addition, some essential characteristics that discriminate domestic from non-domestic violence reports were discovered. However, the limitations of the FCA technique also became apparent. Concept lattices could only be used effectively with a maximum of 14 attributes. If more attributes were used, the lattices became too cluttered to be useful for data exploration. Issues for future research include improving the used thesaurus by amongst others taking negated sentences like "my dad didn't hit me" into account.

Acknowledgements

The authors would like to thank the police organization of the region Amsterdam-Amstelland and in particular Deputy Chief Organisation Information and Programme Manager Intelligence Led Policing Reinder Doeleman and Chief Information and Operations Hans Schönfeld for supporting this research.

References

[1] About Domestic Violence (http:www.usdoj.gov/ovw/domviolence.htm). Office on Violence against Women. Retrieved on 2007-10-22
[2] Watts, C., Timmerman, C.: Violence against women: global scope and magnitude. The Lancet 359 (9313): pp.1232-1237. RMID 1155557
[3] Waits, K. (1984-1985). The criminal Justice System's response to Battering: Understanding the problem, forging the solutions. Washington Law Review 60: pp. 267-330
[4] Vincent, J.P., Jouriles, E.N. (2000) Domestic violence. Guidelines for research-informed practice. Jessica Kingsley Publishers Londen and Philadelphia
[5] Catriona Minleer-Black (1999) Domestic violence: Findings from a new British Crime Survey self-completion questionnaire. London: Home Office Research Study.
[6] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
[7] Yevtushenko, S.A. (2000). System of data analysis "Concept Explorer". Proceedings of the 7th national conference or Artificial Intelligence. KII-2000. 127-134, Russia
[8] Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.
[9] Wille, R. (1982), Restructuring lattice theory: an approach based on hierarchies of concepts, I. Rival (ed.). Ordered sets. Reidel, Dordrecht-Boston, 445-470.
[10] Priss, U. (2005), Formal Concept Analysis in Information Science, Cronin, Blaise (ed.), Annual Review of Information Science and Technology, ASIST, Vol. 40.
[11] Wille, R. (2002), Why can concept lattices support knowledge discovery in databases?, Journal of Experimental & Theoretical Artificial Intelligence, 14: 2, 81-92.
[12] Stumme, G., Wille, R., Wille, U. (1998), Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods, In: J.M. Zytkow, M. Quafofou (eds.): Principles of Data Mining and Knowledge Discovery, Proc. 2 nd European Symposium on PKDD '98, LNAI 1510, Springer, Heidelberg, 1998, 450-458.
[13] Stumme, G. (2002) Efficient Data Mining Based on Formal Concept Analysis. Lecture Notes in Computer Science Vol. 2453, Springer, Heidelberg, 3-22
[14] Stumme, G. (2002), Formal Concept Analysis on its Way from Mathematics to Computer Science. Proc. 10th Intl. Conf. on Conceptual Structures (ICCS 2002). LNCS, Springer, Heidelberg 2002.
[15] Priss, U. (2000), Lattice-based information Retrieval. Knowledge Organization, 27, 3, 132-142.
[16] Godin, R., Gescei, J., Pichet, C. (1989), Design of browsing interface for information retrieval. In: N.J.Belkin, C.J. van Rijsbergen (Eds.), Proc. SIGIR '89, 32-39.
[17] 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.

[18] Cole, R. , Eklund, P. (2001), Browsing Semi-structured Web Texts Using Formal Concept Analysis. In H. Delugach, G., Stumme (Eds.), Conceptual Structures: Broadening the Base, LNAI 2120, Berlin, Springer, 319-332.
[19] Eklund, P., Ducrou, J., Brawn, P. (2004), Concept Lattices for Information Visualization: Can Novice Read Line Diagrams? In P. Eklund (Ed.), Concept lattices: Second International Conference on Formal Concept Analysis, LNCS 2961, Berlin, Springer, 14-27.
[20] Priss, U. (1997), A Graphical Interface for Document Retrieval Based on Formal Concept Analysis. In: E. Santos (Ed.), Proc. of the 8th Midwest Artificial Intelligence and Cognitive Science Conference. AAAI Technical Report CF-97-01, 66-70.
[21] 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.