DETECTING DOMESTIC VIOLENCE
Showcasing a Knowledge Browser based on Formal Concept Analysis and Emergent Self Organizing Maps

Paul Elzinga
Department of Business Information, Police Amsterdam-Amstelland, The Netherlands

Jonas Poelmans, Stijn Viaene, Guido Dedene
Faculty of Business and Economics, K.U. Leuven, Belgium
Vlerick Leuven Gent Management School, Belgium

Keywords: Formal Concept Analysis (FCA), Emergent SOM, Domestic violence, Knowledge discovery in databases, Text mining, Exploratory data analysis.

Abstract: Over 90% of the case data from police inquiries is stored as unstructured text in police databases. We use the combination of Formal Concept Analysis and Emergent Self Organizing Maps for exploring a dataset of unstructured police reports out of the Amsterdam-Amstelland police region in the Netherlands. In this paper, we specifically aim at making the reader familiar with how we used these two tools for browsing the dataset and how we discovered useful patterns for labelling cases as domestic or as non-domestic violence.

1 INTRODUCTION

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 for applying text mining in criminal analysis have been taken (Ananyan, 2002; Chen, 2004). In the case we are studying, we aim at automating the detection of domestic violence using the unstructured text contained in police reports. We use real-life data recorded during 2007 by the Amsterdam-Amstelland police in The Netherlands.

In 1997, the ministry of Justice of the Netherlands made its first inquiry into the nature and scope of domestic violence. 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 placed the topic high on the political agenda.

Pursuing an effective policy against offenders is one of the top priorities of the regional police Amsterdam-Amstelland. In order to do this, being able to swiftly recognize cases of domestic violence and label reports accordingly is of the utmost importance. 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 assign the label domestic violence to the report. That, however, has proven problematic. In the past, intensive audits of the police databases related to filed reports have shown that many reports are not well classified.

Therefore, a case triage system has been put in place to automatically filter out suspicious cases for in-depth manual inspection and classification. Still, to date more than 80% of these suspicious cases are wrongly selected for in-depth inspection by the system (i.e. false positives). Given that it takes at least five minutes to read and classify a case, it is clear that building an efficient and automated case labelling system would result in major savings.

In 2007, the database of the Amsterdam-Amstelland police contained more than 7000 cases with a statement made by the victim of a violent incident to the police. Because it is physically impossible for any individual to process this sheer amount of information, applying text mining technology seems a natural approach. Text mining has been defined as “the discovery by computer of

Elzinga P., Poelmans J., Viaene S. and Dedene G. (2009).
DETECTING DOMESTIC VIOLENCE - Showcasing a Knowledge Browser based on Formal Concept Analysis and Emergent Self Organizing Maps.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 11-18
DOI: 10.5220/0001844400110018

new, previously unknown, information by automatically extracting information from different written resources" (Fan, 2006).

For our data exploration we chose to use a combination of two visually appealing discovery techniques, known as Formal Concept Analysis (FCA) (Stumme, 2002) and Emergent Self Organizing Maps (ESOM) (Ultsch, 2005). Formal Concept Analysis (FCA) arose twenty-five years ago as a mathematical theory (Stumme, 2002). FCA was for the first time used as an exploratory data analysis and knowledge enrichment technique for analysing domestic violence cases in the Netherlands (Poelmans, 2008). In this setup, FCA is used as a concept generation engine, distilling formal concepts from the unstructured documents. We complement the knowledge discovery based on FCA with the ESOM. Emergent Self Organizing Maps are a special class of topographic maps, which are particularly suited for high-dimensional data visualization.

In this paper we aim at making the reader familiar with how we used these tools for browsing through the data in search of new knowledge for classifying new cases. The result of the research is a case labelling system that automatically and correctly assigns the domestic violence or non-domestic violence label to a large portion of the incoming cases.

The rest of this paper is structured as follows. In section 2, the dataset is discussed. In section 3, the use of FCA and ESOM for knowledge discovery is presented and applied to the data at hand. In section 4, the ensuing detection process is summarized. Section 5 concludes the paper.

2 THE DATASET

According to the Dutch police authorities and the department of Justice, 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 (Van Dijk, 1997).

The dataset we report on in this paper consists of a selection of 4814 police reports describing a whole range of violent incidents from the year 2007. The domestic violence cases for that period are a subset of this dataset. This selection came about by, among other things, filtering from a larger set 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 a police officer is sent to an incident and later on writes a report in which he/she mentions his/her findings, while the victim has not made an official statement to the police. The follow-up reports referring to previous cases were also removed from the initial set of reports. Ultimately, this gave rise to a set of 4814 reports that were used as input for our investigation. From these reports, 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 were extracted. Of the 4814 reports, 1657 were classified by police officers as domestic violence; the others were not.

3 HUMAN-CENTERED KNOWLEDGE DISCOVERY

In the literature, the need for exploratory data analysis has often been described (Marchionini, 2006). When beginning the analysis of a new dataset of which very little is known a priori, the first step is to explore the data. Data mining should be primarily concerned with making it easy, convenient and practical to explore very large databases for organizations with a lot of users but without requiring years of training into data analysis (Fayyad, 2002). Unfortunately, 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 (Brachman, 1996). A human-centered approach is proposed. A significant part of the art of data mining is the user's intuition with respect to the tools (Smyth, 2002). We argue that the combined use of FCA and ESOM fulfils this need. The visual representations of both tools provide a clear guide to the user for exploring the data.

Additionally, we aim at developing a classifier for automatically classifying cases as domestic or as non-domestic violence. Comprehensibility of the performed classification is a key requirement, requiring that the user understands the motivations behind the model's prediction (Martens, 2004). In the domain of police investigations, the lack of comprehensibility is a major issue and causes a

reluctance to use a classifier or even complete rejection of the model. This is for a large part due to the very high cost of classifying a case incorrectly as domestic or as non-domestic violence. Clarity and explainability of the performed classification are major constraints. Comprehensibility measures the "mental fit" of the classification model (Kodratoff, 1994).

The knowledge discovery process using the combination of FCA and ESOM is displayed in Figure 1.

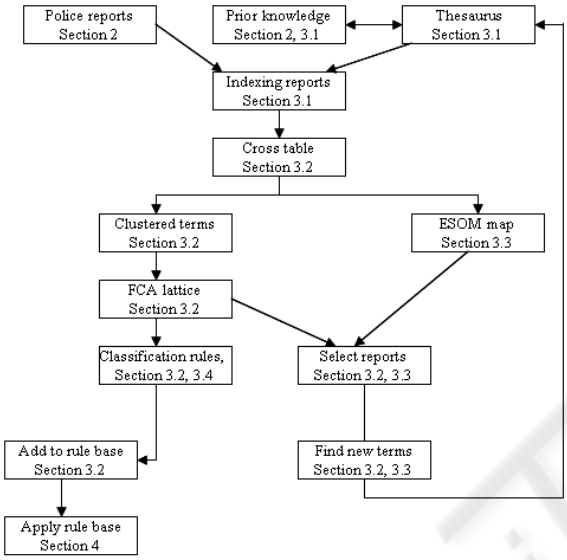

Figure 1: Knowledge acquisition process.

The knowledge acquisition process is explained and showcased in the next subsections.

3.1 Data Preparation

The initial phase of the knowledge acquisition process consists of translating the investigation area 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 3.2) 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", etc. Since according to the literature domestic violence is a phenomenon that typically occurs inside the house, we also added terms such as "bathroom", "living room", etc. We made an explicit distinction with public locations such as "under the bridge", "on the street", etc. The initial thesaurus contained 123 elements. An excerpt of this thesaurus is shown in Figure 2.

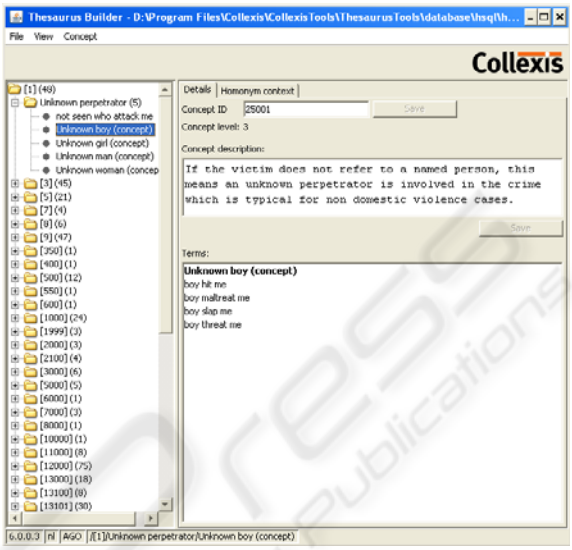

Figure 2: Thesaurus.

The reports were indexed using this thesaurus. For each report the found thesaurus elements were stored in a collection. This collection would be used as an input for both the FCA and the ESOM processing.

The thesaurus was refined after each iteration of re-indexing the reports and visualizing and analysing the data with the FCA lattice and ESOM maps. This process is showcased in section 3.2 and 3.3.

3.2 Exploratory Analysis with FCA

The tool we used to visualize and analyze the data with FCA is the tool Conexp (Yevtushenko, 2000).

The starting point of the FCA 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 Figure 3. 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 this term.

Prior to the analysis with FCA, the terms in the thesaurus have to be clustered in term clusters. The initial clustering, performed on the basis of expert prior knowledge, resulted in four term clusters: "acts

of violence", "public locations", "private locations" and "persons".

The formal context based on the domestic violence definition is displayed in Figure 3.

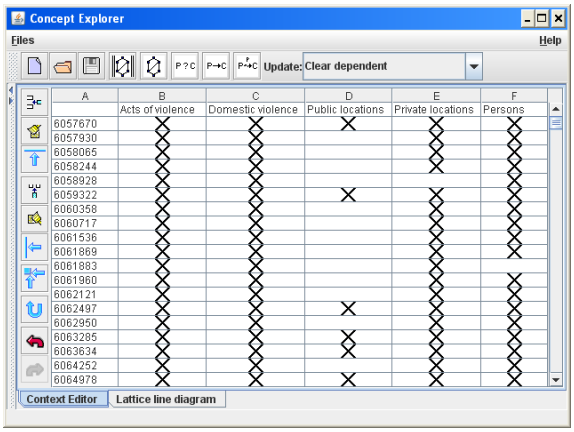

Figure 3: Initial cross table.

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 exercise.

Given a formal context, FCA 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. The line diagram corresponding to the cross table from Figure 3 is represented in Figure 4.

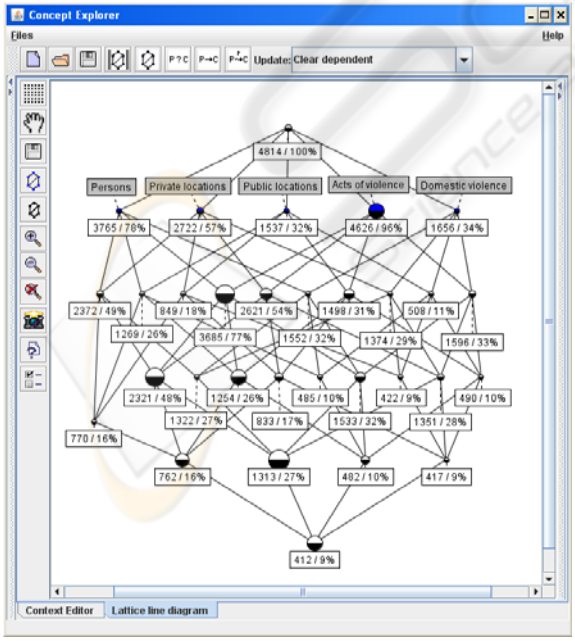

Figure 4: Initial lattice.

When the lattice has been built, it can be analyzed using the Conexp tool. The tool provides a user-friendly interface to navigating the lattice. The user can highlight specific aspects of the lattice, select and/or deselect attributes, display the case numbers that belong to each of the concepts, etc.

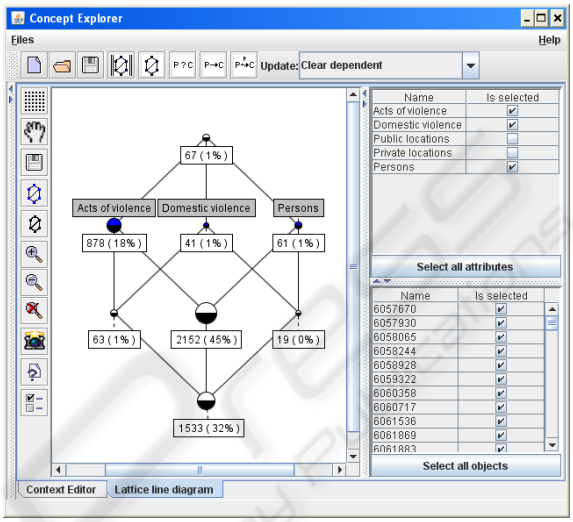

Figure 5: Lattice with own object count.

For example, the lattice displayed in Figure 5 was obtained by deselecting the attributes "private locations" and "public locations," and by showing the number of own objects for each concept.

Another example is the lattice displayed in Figure 6 that was obtained by making the case numbers belonging to an interesting concept visible.

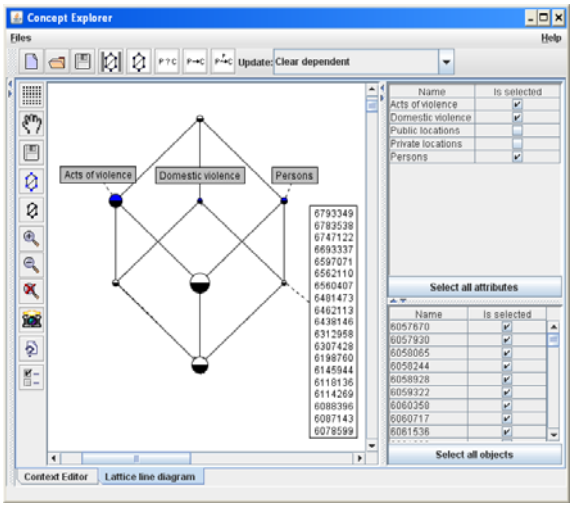

Figure 6: Lattice with case numbers.

Some interesting findings that emerged from navigating the lattice are displayed in Table 1.

Table 1: Key statistics from initial lattice navigation.

	Non-domestic violence	Domestic violence
No “acts of violence”	67	41
No “acts of violence” and “persons”	61	19
Only “acts of violence”	877	64

As can be seen from Table 1, a total of 60 (i.e. 41 and 19) domestic violence cases did not contain a term from the “acts of violence” term cluster. Of these 60 cases 19 contained a term from one of the clusters containing terms referring to a person in the domestic sphere of the victim. After in-depth manual inspection of these 19 cases, it turned out that they contained other violence terms, such as “abduction”, “strangle” and “deprivation of liberty”, which were lacking in the initial thesaurus. The remaining 41 cases, on the other hand, turned out to be wrongly classified as domestic violence.

Interestingly, some 28% (i.e. 877) of the non-domestic violence reports only contain terms from the “acts of violence” cluster, while there are only 64 domestic violence reports in the dataset that share that characteristic. Manual inspection, again, revealed that more than two thirds of these reports were wrongly classified as domestic violence. 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”. These term clusters are used as classification rules. If a case contains a term from the “burglary cases” cluster, for example, we found that it can be correctly classified as non-domestic violence.

After several iterations of exploring and refining the thesaurus, 44% of the cases could automatically be classified using such classification rules. To complement the use of FCA, the ESOM was introduced to function as a catalyst during the knowledge acquisition process, amongst others for finding new terms and term clusters. The ESOM tool is described in detail in the next section.

3.3 Exploratory Analysis with ESOM

The ESOM performs a non-linear mapping of the high-dimensional data space to a two-dimensional one which enables the exploration of the data (Ultsch, 2003). Emergence is the ability of a system to produce a phenomenon on a new, higher level (Ultsch, 1999). 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. In the traditional SOM, the number of nodes is too small to show emergence.

The ESOM can be used to detect clusters and maintains the neighbourhood relationships that are present in the input space. It also provides the user with an idea of the complexity of the dataset, the distribution of the dataset and the amount of overlap between the different classes. Only a minimal amount of expert knowledge is required for the user to be able to use it effectively for exploratory data analysis. The maps can be created and used for data analysis by means of the publicly available Databionics ESOM Tool (Ultsch, 2005). With this tool the user can construct ESOMs with either flat or unbounded (i.e., toroidal) topologies.

The ESOM is used as a catalyst in the knowledge discovery process, amongst others for finding new terms and term clusters. ESOM provides a highly interactive user interface that moves exploratory search process beyond predictable fact retrieval. Because of its highly interactive user interface, human participation and effective use of their expert prior knowledge in the search process is promoted.

Indexing the 4814 reports from 2007 with the initial thesaurus from section 3.1 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 presented in Figure 7. 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.

Some red squares in Figure 7 are located in the middle of a large group of green squares and vice versa. The ESOM tool allows the user to select neurons on the map (yellow area in the map of Figure 7). It displays the cases that had this neuron as a best match in the lower pane of Figure 7. After an in-depth manual inspection of the police reports corresponding to these 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. The reports also contained multiple confusing

situations that upon disclosure to us were used to refine the domestic violence definition.

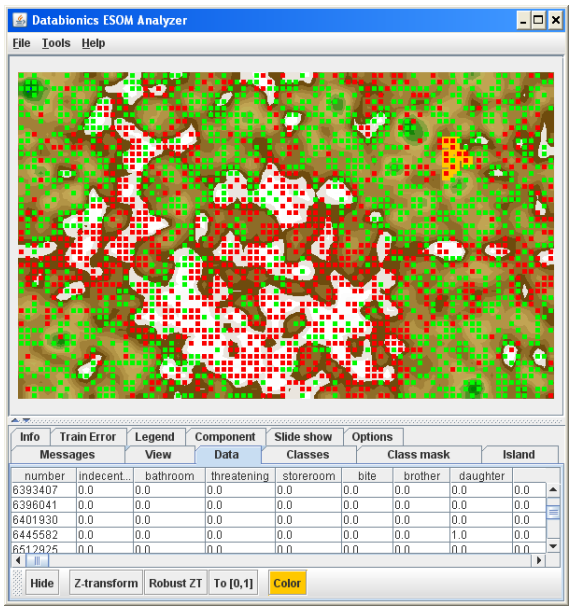

Figure 7: Toroidal ESOM trained on the thesaurus from section 3.1.

Every time new and important features were discovered in this way, they were used to enrich the thesaurus, and triggered a new iteration of the FCA analysis, followed 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 a 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”.

4 DOMESTIC VIOLENCE DETECTION SYSTEM

Based on the knowledge discovery laid out in the previous sections, 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.

4.1 Classification Rules

After a successive number of knowledge acquisition iterations using FCA and ESOM, we were able to enrich the initial thesaurus from 123 to 829 thesaurus elements. These 829 thesaurus elements include 50 term clusters and 779 domain-specific terms.

With FCA 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 can be classified automatically as either domestic or non-domestic violence. We also applied these rules to a validation set containing unstructured police reports from the year 2006 and a similar result was obtained. In the near future, these rules can be used to automatically classify the majority of incoming cases, while in the past all cases had to be dealt with manually

Ten of the domestic violence and five of the non-domestic violence classification rules are displayed in Table 2.

Table 2: 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	Honour related violence
6	Incest
7	(Court) injunction
8	Fear of domestic sphere
9	Committed by domestic sphere
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

These classification rules were also used to detect filed reports that were assigned a faulty case label.

In the next section we discuss the application that was developed to apply the discovered knowledge to classify police reports.

4.2 Labelling Process

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). A standard thesaurus is provided, but the user can modify or switch this thesaurus at will. The next step consists of indexing the set of police reports using the selected thesaurus and generating a cross table. The text mining tool Collexis is used to index the police reports and translate the found attributes into Prolog-facts.

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. Otherwise, it is verified whether one of the non-domestic violence rules is satisfied. If this is the case, the case is classified as non-domestic violence. Otherwise, the case is left unclassified.

The rule base was developed in Prolog, in order to make the knowledge base more flexible and expandable for other areas such as discrimination and weapons. The result of applying the classification rules to a set of police reports is displayed in Figure 8.

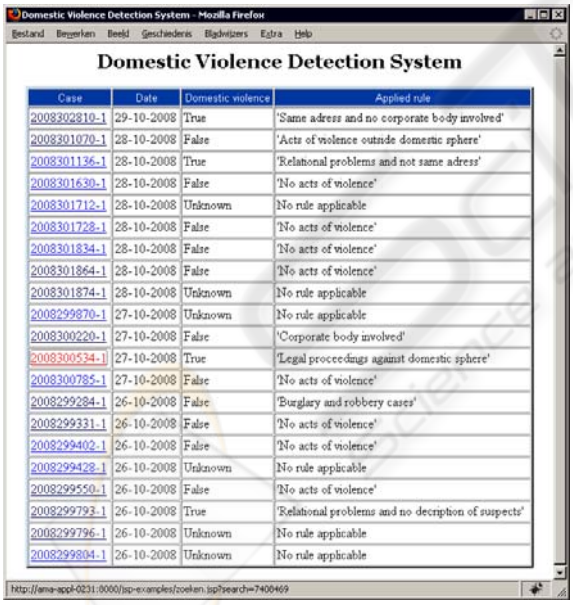

Figure 8: Domestic violence detection system.

For each report, the assigned label and the applied rule are shown. Moreover, for each case, there is a hyperlink on which the user can click to open the corresponding report in a popup window. As shown in Figure 9, the terms from the thesaurus, that were found in the report, are highlighted.

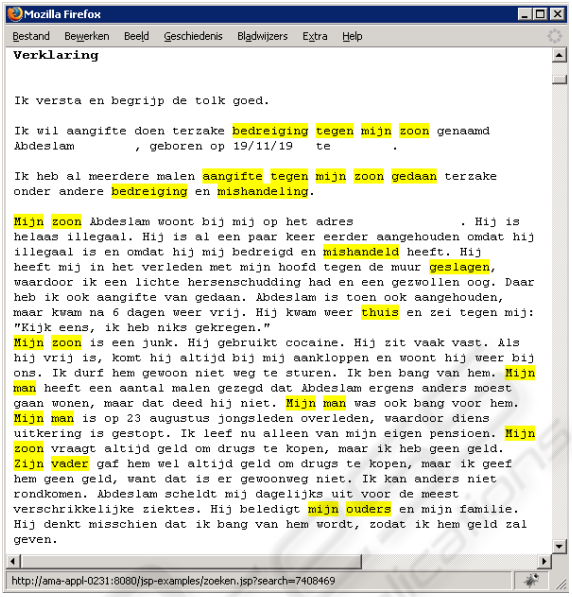

Figure 9: Police report with highlighted thesaurus terms.

The user can quickly verify whether the classification is correct. The application will also highlight the found attributes if no classification can be made (in 25% of the cases). The highlighting of terms makes it easier for the user to label a case.

5 CONCLUSIONS

This paper showed how the Amsterdam-Amstelland police is using text mining on police reports to better identify cases of domestic violence. At the heart of the knowledge discovery process are the lattices produced with FCA and the maps produced with ESOM. The knowledge browsing capabilities of these tools provide for a powerful framework for exploring unstructured data. In our search for improved detection of domestic violence cases, we discovered 37 classification rules, i.e. 22 domestic violence and 15 non-domestic violence rules. With these rules 75% of cases can be labelled automatically and correctly, whereas in the past, all cases had to be dealt with manually. These rules are the motor of a detection system that aids analysts in their assessment of incoming cases.

ACKNOWLEDGEMENTS

The authors would like to thank the police of Amsterdam-Amstelland for providing them with the necessary degrees of freedom to conduct and publish

this research. In particular, we are most grateful to Deputy Police Chief Reinder Doeleman and Police Chief Hans Schönfeld for their continued support. Jonas Poelmans is Aspirant of the "Fonds voor Wetenschappelijk Onderzoek – Vlaanderen" (FWO) or Research Foundation – Flanders.

REFERENCES

Brachman, R., Anand, T., 1996. *The process of knowledge discovery in databases: a human-centered approach.* In advances in knowledge discovery and data mining, ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. AAAI/MIT Press.

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M., 2004. *Crime data mining: a general frame-work and some examples*. IEEE Computer, April 2004.ijk, T van, 1997. *Huiselijk geweld, aard, omvang en hulp-verlening*. Ministerie van Justitie, Dienst Preventie, Jeugd-bescherming en Reclassering.

Fan, W., Wallace, L., Rich, S., Thang, T., 2006. *Tapping the power of text mining*. Communications of the ACM, Vol. 49, no. 9.

Fayyad, U., Uthurusamy, R., 2002. *Evolving data mining into solutions for insights.* Communications of the ACM, Vol. 45, no. 8.

Kodratoff, Y., 1994. *The comprehensibility manifesto*. KDD nuggets (94:9).

Marchionini, G., 2006. *Exploratory search: from finding to understanding*. Communications of the ACM, Vol. 49, no. 4.

Martens, D., Baesens, B., Van Gestel, T., Vanthienen, J., 2007. *Comprehensible credit scoring models using rule extraction from support vector machines.* European journal of operational research, 183(3), 1466-1476.

Poelmans, J, Elzinga P, Viaene S., Dedene G, 2008. *An exploration into the power of Formal Concept Analysis for domestic violence analysis.* 8th Industrial Conference, ICDM 2008 Leipzig, Germany, Juli 2008 Proceedings, LNAI 5077 Springer Heidelberg, pp. 404-416.

Pednault, E.P.D. 2000, *Representation is everything*. Communications of the ACM, Vol. 43, no. 8.

Smyth, P., Pregibon, D., Faloutsos, C., 2002. *Data-driven evolution of data mining algorithms.* Communications of the ACM, Vol. 45, no. 8.

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.

Ultsch, A., 1999. *Data mining and knowledge discovery with Emergent SOFMS for multivariate Time Series.* Kohonen Maps, pp. 33-46.

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

Ultsch, A., Moerchen, F., 2005. *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.

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