A case of using Formal Concept Analysis in combination with Emergent Self Organizing Maps for detecting domestic violence

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

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

Abstract. In this paper, we propose a framework for iterative knowledge discovery from unstructured text using Formal Concept Analysis and Emergent Self Organizing Maps. We apply the framework to a real life case study using data from the Amsterdam-Amstelland police. The case zooms in on the problem of distilling concepts for domestic violence from the unstructured text in police reports. Our human-centered framework facilitates the exploration of the data and allows for an efficient incorporation of prior expert knowledge to steer the discovery process. This exploration resulted in the discovery of faulty case labellings, common classification errors made by police officers, confusing situations, missing values in police reports, etc. The framework was also used for iteratively expanding a domain-specific thesaurus. Furthermore, we showed how the presented method was used to develop a highly accurate and comprehensible classification model that automatically assigns a domestic or non-domestic violence label to police reports.

Keywords: Formal Concept Analysis, Emergent Self Organizing Map, text mining, actionable knowledge discovery, domestic violence.

1 Introduction

In this paper, we propose a framework for knowledge discovery from unstructured text based on the synergistic combination of two visually appealing discovery techniques known as, Formal Concept Analysis (FCA) [10, 11] and Emergent Self Organizing Maps (ESOM) [2, 5]. The framework 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.

FCA arose twenty-five years ago as a mathematical theory [10, 15] and has over the years grown into a powerful tool for data analysis, data visualization and information retrieval [12, 13]. We complement the knowledge discovery based on FCA with a special type of topographic map known as ESOM. The ESOM functions as a catalyst for the FCA based discovery. A key contribution of this paper is that we ground the knowledge discovery approach based on FCA and ESOM in the theoretical foundation of C-K theory. C-K theory is used to give a clear structure to the discovery process based on FCA and ESOM.

We apply the presented framework to a real life case study using data from the Amsterdam-Amstelland police. The case zooms in on the problem of distilling concepts for domestic violence from the unstructured text in police reports. These concepts are used to iteratively enrich the actionable knowledge available to police officers for recognizing cases of domestic violence. Domestic violence is one of the top priorities of the Amsterdam-Amstelland police force [19]. Unfortunately, in the past intensive audits of the police databases related to filed reports established that many police reports tended to be wrongly labelled as domestic or as non-domestic violence cases.

In this paper, we demonstrate that by applying the discovery framework to the unstructured text in police reports we can obtain essential knowledge for upgrading the definition and understanding of the domestic violence phenomenon, and for improving its management. In addition, we show how early detection of domestic violence could be automated.

The remainder of this paper is composed as follows. In section 2, we introduce the data exploration techniques of FCA and ESOM. In section 3, we discuss the dataset. In section 4, we elaborate on the knowledge discovery process, apply it to the data set at hand, and report results. Section 5 concludes the paper.

2. Exploration techniques

According to R.S. Brachman and T. Anand [17] much attention and effort has been focused on the development of data mining techniques, but only minor effort has been devoted to the development of tools that support the analyst in the overall discovery task. The authors 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. This can only be achieved by tools that offer highly interactive user interfaces that continuously engage human control over the information seeking process [16]. According to Brachman et al. [18] this is best embedded into a knowledge discovery support environment. FCA and ESOM are particularly suited for exploratory data analysis because of their human-centeredness. FCA and ESOM offer the user intuitive visual displays of different types of structures available in the dataset and guide the user in its exploration.

2.1 Formal Concept Analysis

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*). In our case, 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. 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).

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, [10, 11]. The extension consists of all objects belonging to the concept, while the intension comprises all attributes shared by those objects. 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 a report from the dataset used in this paper, for example. By collecting all reports of this context that share these attributes, we get to a set $O \subseteq M$ consisting of reports. This set O of objects is closely connected to set A consisting of attributes

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

In other words, 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. 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 and only if the extent of d is a subset of the extent of e (or equivalently, if and only if the intent of d is a superset of the intent of e), or

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

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 (labelled) line diagram.

In section 4 of this paper, a line diagram is displayed. We use the following conventions. The circles or nodes in the line diagram represent the formal concepts. The diagram 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 concept lattice can be distilled from the line diagram 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 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 Emergent Self Organizing Map

Emergent Self Organizing Maps (ESOM) [2] are a special class of topographic maps [7]. ESOM is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of its structure [6]. 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 [4]. ESOM is a more recent type of topographic map. According to Ultsch [5], "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 [3]. In the traditional SOM, the number of nodes is too small to show emergence.

The ESOM map is composed of a set of neurons *I*, arranged in a hex-grid map structure. A neuron $i \in I$ is a tuple (w_i, p_i) consisting of a weight vector $w_i \in W$ and a position $p_i \in P$ in the map. The input space $D \subset R^n$ is a metric subspace of R^n.

Consider a training set $E = \{x_1, ..., x_k\}$ with $x_1, ..., x_k \in R^n$ representing the input vectors for ESOM training. The training algorithm used is the online training algorithm in which the best match for an input vector is searched and the neighborhood of the map is updated immediately. When an input vector x_i is supplied to the training algorithm, the weight of a neuron $n_1 = (w_1, p_1)$ is modified as follows. Let $\eta \in [0,1]$, then

$$\Delta w_1 = r \times h \times (bm_i, n_1, r) \times (x_i - w_1)$$

where

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

represents the best-matching neuron of an input vector , being the neuron $n_b \in I$ having the smallest Euclidean distance to x_i , or

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

where $d(x_i, w_j)$ stands for the Euclidean distance between input vector x_i and weight vector w_j .

The neighborhood of a neuron

$$N_i = N(n_i) = \{n_j \in M \mid h_{ij}(r) \neq 0\}$$

is the set of neurons surrounding neuron n_i and determined by the neighborhood function *h*. The neighborhood defines a lattice of neurons in the map space *K,* while *r* is called the neighborhood radius.

The produced map 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 different classes of objects. Finally, only a minimal amount of expert knowledge is required for an analyst to start use it effectively for exploratory data analysis. An additional advantage of an ESOM is that it can be trained directly on the available dataset without first having to perform a feature selection procedure [5]. ESOM maps can be created and used for data analysis by means of the publicly available Databionics ESOM Tool [8].

3. 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 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 investigate an incident, afterwards writes a report in which he/she mentions his/her 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. Of those 4814 reports, 1657 were classified by police officers as domestic violence.

We also used a validation data set for our experiment. It 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. 1734 of these 4738 cases were classified as domestic violence by police officers.

4. Iterative knowledge discovery with FCA and ESOM

In this section, we elaborate on the applied process for knowledge discovery based on the synergistic combination of the visually appealing discovery techniques presented in section 2. In this setup, FCA is used as a concept generation engine, distilling formal concepts from the unstructured text documents described in section 3. We

complement this knowledge discovery with the capabilities of ESOM, which functions as a catalyst for the FCA based knowledge extraction. To frame our approach to knowledge discovery we make use of C-K theory.

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 [1]. The theory makes a formal distinction between concepts and knowledge: the K space consists of propositions with logical status (i.e. either true or false) for a designer, and the C space consists of propositions without logical status in the knowledge space. According to Hatchuel et al. [1], concepts are candidates to be transformed into propositions of K but are not themselves elements of K.

From the point of knowledge discovery in data, the K space could be viewed as composed of actionable information. It contains the existing knowledge used to operate and steer in the action environment. The C space, on the other hand, is 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. The transformations within and between the C and K spaces are accomplished by the application of four transformation operators: $C \rightarrow K$, $K \rightarrow C$, $C \rightarrow C$, and $K \rightarrow K$. These transformations form what is referred to as the design square, which fundamentally provides structure to the design process. At the basis of the knowledge discovery process are iterations through the design square. The knowledge discovery documented in this paper was driven by a data analyst and a domain expert, who collaborated intensely and continuously interacted with the visual exploration tools. They immersed themselves in a process of iterating through the transformation operators of the design square. This process is summarized in Figure 1.

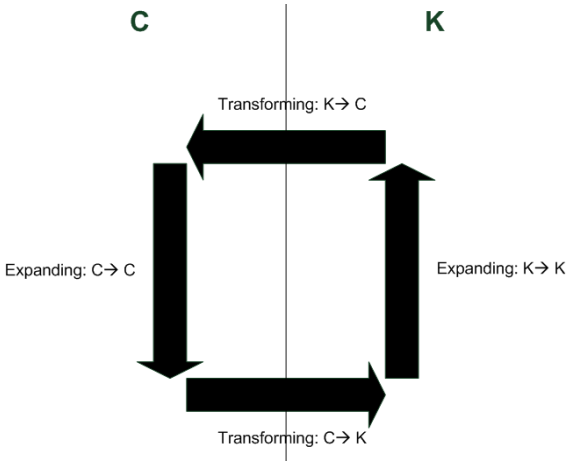

Fig. 1. Design square in action

In this case, the process in Figure 1 works as follows:

- The process starts with the data analyst constructing an initial FCA lattice and ESOM map using the police reports contained in the dataset, and the terms contained in a thesaurus (i.e. K $\rightarrow$ C). 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.
- The FCA lattice and the ESOM map provide a reduced search space to the domain expert, who then visually inspects and analyses the lattice and map (i.e. C $\rightarrow$ C). In other words, the FCA lattice and ESOM map are used in the capacity of information browser.
- Based on anomalies and counter-intuitive elements found by analyzing the lattice, or using ESOM to pinpoint outliers, clusters and areas of the map containing a mixtures of case types, police reports can be selected for in-depth manual inspection (i.e. C $\rightarrow$ K).
- These police reports are in turn used, for example, to discover new referential terms to improve the thesaurus, to enrich and validate prior domain knowledge, to discover new classification rules, and for operational validation (i.e. K $\rightarrow$ 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) for starting a new iteration of the design square.

In the remainder of this section we limit ourselves to illustrating how ESOM and FCA help to operationalise the four design cube operators that constitute the knowledge discovery process.

4.1 Transforming: K→C

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 [16]."

The lattice structure in Figure 2 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 a thesaurus. For example, to verify whether a criminal offence has occurred, reports were searched for terms such as "hit", "stab" and "kick". These terms were grouped into the term cluster "acts of violence". Another term cluster, for example, verified whether a person from the domestic sphere of the victim was involved in the crime. In this case, reports were 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".

Using the reference definition of domestic violence employed by the police was but one way to come up with term clusters to structure the lattice in Figure 2. Term clusters also emerged from in-depth scanning of certain reports highlighted during a knowledge iteration cycle. This is the way, for example, in which the term cluster

"relational problems" was created. Terms in that cluster made reference to a broken relationship. In-depth scanning of reports is also the way in which we found that many cases did not have a formally labelled suspect. Thus, we incorporated the attribute "no suspect" into the lattice. Reports that were assigned the label "domestic violence" had been classified as such by police officers. The remaining reports were classified as non-domestic violence.

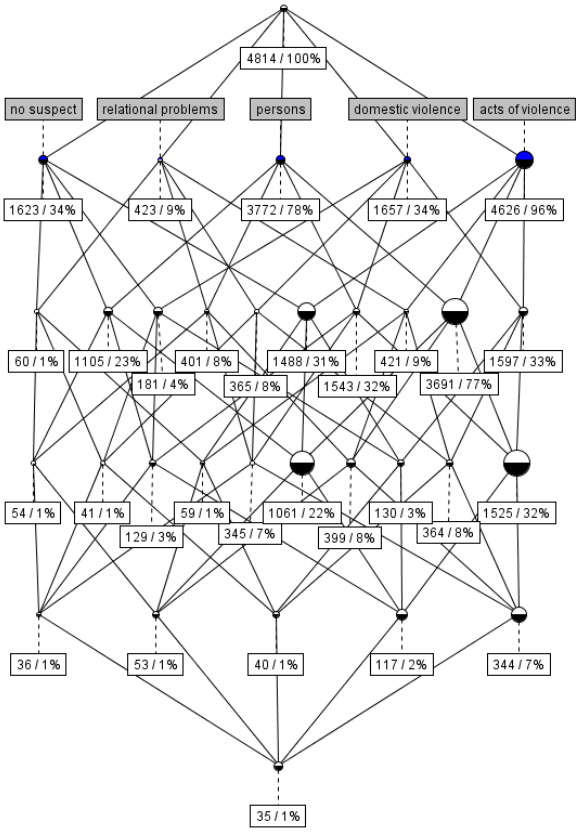

Fig. 2. Lattice based on the police reports from 2007

Some 123 elements were collected into an initial thesaurus. This gathering process was not only influenced by the above domestic violence definition, but also incorporated information from other prior knowledge sources such as expert advice. Indexing the set of 4814 reports from 2007 with this thesaurus resulted in a data set (i.e. reports are objects and thesaurus elements are object attributes) for training a toroidal ESOM map. The latter is presented in Figure 3. 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.

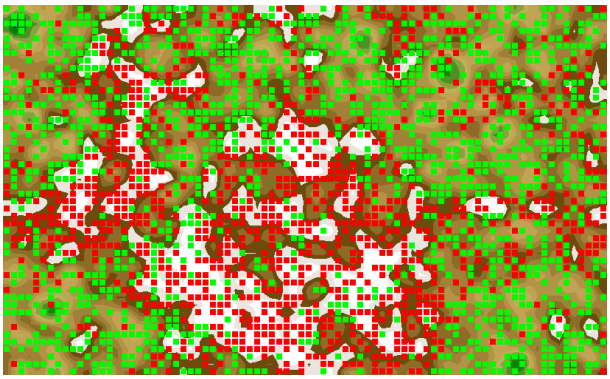

Fig. 3. ESOM map based on the police reports from 2007

4.2 Expanding: C→C

The notion of expansion plays a key role in C-K theory. An analysts ability to recognize an expansion opportunity depends on his sensitivity to these opportunities, his training, and the knowledge at his disposal. We argue here that FCA and ESOM help analysts recognize and exploit these opportunities. Basically, C space expansion is driven by the analyst spotting and investigating anomalies, outliers and concept gaps from these visual exploration tools. For example, the FCA lattice in Figure 2 allowed us to make some interesting observations associated with the data at hand. We use the numbers in Table 1 to illustrate.

Table 1. Interesting observations from the lattice in Figure 2

	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

From Table 1 it is clear that 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 cluster 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 labelled as domestic violence did not fit with the definition of domestic

violence that was used to start this discovery exercise in the first place. The associated reports were therefore selected for in-depth investigation.

Zooming in onto the term cluster "relational problems" we observed some more interesting facts. Apparently, only 58 non-domestic violence reports contained one or more terms from that cluster. We concluded that the presence of at least one of the terms from 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.

Visual inspection of the patterns laid down by the ESOM map in Figure 3 also allowed us to make interesting observations. For example, using colour coding made it easy to spot outlying observations, that is, red squares located in the middle of a large group of green squares, and vice versa. For inspection we made use of the ESOM tool's functionality to select neurons for displaying the cases that had these neurons as best match. We thought that these neurons were associated with cases that might have been wrongly classified by police officers.

4.3 Transforming: C→K

The C → K operator transforms concepts in C into logical questions in K. An answer to such a question is in our case found by manually inspecting 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 labellings, niche cases, and data quality problems.

For example, and with reference to Table 1, the 18 domestic violence cases that contained a term from the cluster "persons of domestic sphere" but no violence term were selected for manual investigation. In-depth analysis showed that these reports contained violence related terms, such as "abduction", "strangle" and "deprivation of liberty", that were originally lacking from the initial thesaurus. Another example, are the 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.

Table 1 also indicates that there were 58 police reports that were classified as non-domestic violence while containing a term from the "relational problems" cluster. Investigation revealed that a startling 95% of these cases had been wrongly labelled as non-domestic violence. Moreover, about 70% of these cases had in common that a third person made a statement to the police for someone else. Moreover, analysis of the remaining 30% led to the discovery of an important new concept that turned out to be lacking from the domain expert's initial understanding of domestic violence. Many of the reports included expressions such as "I was attacked by the ex boyfriend of my girlfriend" and "I was maltreated by the ex girlfriend of my boyfriend". This gave rise to a term cluster "attack by ex-person against new friend" that would be added to the thesaurus.

We also went after novel and potentially interesting classification attributes. The fact of not making mention of a suspect in a report is such an attribute. This can be inferred from the FCA lattice in Figure 2. It shows how some 34% of the reports (1623 cases) did not mention a suspect. However, going back to our original

definition of domestic violence, which talked about a perpetrator belonging to the domestic sphere of the victim, an offender thus had to be known for a valid labelling of domestic violence. However, the lattice in Figure 1 shows that 181 of these "no suspect" cases were actually labelled as domestic violence, which led us to further investigate these reports.

Some of the ESOM outliers also helped us to enrich the K space. Inspection of these outlier cases helped us to enrich the thesaurus. Some of the features that constituted this enrichment are mentioned in Table 2.

Table 2. Thesaurus features discovered by exploring ESOM outliers

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: 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.

Many of the reports also contained confusing situations that upon disclosure were used to refine the notion of domestic violence.

4.4 Expanding: K→K

This expansion of the K space constitutes validation or testing of the proposed expansion with the ultimate goal of producing actionable intelligence. K-validation of a concept comes down to a confrontation of the output from the C→K transformation with a selection of knowledge sources available to the K space (e.g. cross-checking with other databases, setting up field experiments, soliciting expert advice).

For example, analysis of the misclassified reports of which we made mention in the previous section showed that apparently, for some unknown reason, police officers regularly misclassified burglary, car theft, bicycle theft and street robbery cases as domestic violence. To consolidate the agreement on this kind of mistake a new term cluster was introduced that would not only influence subsequent iterating through the design square, but police reporting itself by addressing it during police training. The latter would also elaborate on what seemed to be confusing situations to police officers in terms of labelling cases correctly (e.g. when third persons make statements for someone else).

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 doubtful cases to the board members responsible for the domestic

violence policy. Police officers and policy makers confirmed that this type of situation was indeed to be seen as domestic violence, mainly because the perpetrator often aims at emotionally hurting the ex-partner. Consequently, the expectation was for the terms contained in this cluster to relatively frequently occur in cases labelled as domestic violence reports. However, this assumption turned out to be incorrect when scrutinizing the data. Police officers clearly had trouble allocating a correct label to these kind of cases, which anew gave rise to a need for training

The ”no suspect” cases were yet another example of the potential for knowledge expansion. Remember that some of the cases that were labelled as domestic violence by police officers did not make mention of a suspect, which was weird. Studying the way in which police officers registered victim reports helped us uncover some haphazard behaviour in the process. 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 people before casting the label of suspect. In the latter cases, the person that was mentioned would then be added to the list of persons who were said to be involved in or to have witnessed the crime. These lists tended to account for a rather diverse and extensive set of people. 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. This analysis made a strong case for such a policy.

In the process of digging up evidence and confronting the different stakeholders with that evidence we exposed a serious mismatch between the management’s conception of domestic violence and that of police officers. We found that management employed a much broader definition of domestic violence than most police officers.

4.5 Actionable intelligence

Several iterations through the design cube resulted in truly valuable upgrades of the K space from the point of improving action in the field, that is, on the street. Some of the most important achievements of our work are the following:

- We were able to upgrade the definition for domestic violence that would act as a principle guideline for labelling cases.
- Several types of niche cases were identified as valid exceptions to the general definition, and advice, grounded in evidence, was formulated for policy redesign.
- We ended up extracting a set of 22 domestic violence and 15 non-domestic violence classification rules. Using these rules, 75% of cases from the year 2007 could be labelled automatically and correctly as either domestic or non-domestic violence. Validation of these results by applying these rules to the police reports from the year 2006 allowed us to obtain a similar performance at 72%. These rules have been incorporated in an early case filter to identifying cases that warrant in-depth manual inspection. Before, all claims had to be manually checked.
- The set of identified classification rules did not just allow the police to classify newly incoming cases. The rules could also be usefully employed to reclassify

cases from the past to provide for 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.

5. Conclusion

In this paper, we proposed an approach to knowledge discovery from unstructured text using FCA and ESOM. The approach was framed and illustrated with reference to C-K theory (i.e. the design square) to provide for a deep understanding of the nature of the exploration; an exploration that essentially is human-centered. With this paper 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 progress with knowledge expansion by progressively looping through the design square in an effective way. We showcased the framework using a real life case study with data from the Amsterdam-Amstelland police. The case zoomed in on the problem of distilling concepts for domestic violence from the unstructured text in police reports. The data exploration for this case study resulted in several improvements related to the operational reporting on and the handling of domestic violence cases. This included the implementation of an effective early case filter to identify cases that truly warrant in-depth manual inspection.

Acknowledgments

The authors would like to thank the Amsterdam-Amstelland police 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 and relentless support. Jonas Poelmans is aspirant of the Research Foundation – Flanders.

References

[1] Hatchuel, A., Weil, B. (2003) A new approach of innovative design: an introduction to C – K theory. Proc. of ICED'03, august 2003, Stockholm, Sweden, pp. 14.

[2] 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.

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

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

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

[6] Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp. 305-308.

[7] Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York.
[8] http://databionic-esom.sourceforge.net/
[9] Ultsch, A. (1999) Data mining and knowledge discovery with Emergent SOFMS for multivariate Time Series. In Kohonen Maps, pp. 33-46.
[10] Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.
[11] Wille, R. (1982) Restructuring lattice theory: an approach based on hierarchies of concepts, I. Rival (ed.). Ordered sets. Reidel, Dordrecht-Boston, pp. 445-470.
[12] Priss, U. (2005) Formal Concept Analysis in Information Science, Cronin, Blaise (ed.), Annual Review of Information Science and Technology, ASIST, Vol. 40.
[13] Wille, R. (2002) Why can concept lattices support knowledge discovery in databases?, Journal of Experimental & Theoretical Artificial Intelligence, 14: 2, pp. 81-92.
[14] 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. 2nd European Symposium on PKDD '98, LNAI 1510, Springer, Heidelberg, pp. 450-458.
[15] 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.
[16] T. van Dijk (1997) Huiselijk geweld, aard, omvang en hulpverlening. Ministerie van Justitie, Dienst Preventie, Jeugd-bescherming en Reclassering.
[17] 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.
[18] 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.
[19] http://www.politie-amsterdam-amstelland.nl/get.cfm?id=86