Expert Systems with Applications 38 (2011) 3116–3130

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Formally analysing the concepts of domestic violence

Jonas Poelmans [a,*], Paul Elzinga [c], Stijn Viaene [a,b], Guido Dedene [a,d]

[a] K.U. Leuven, Faculty of Business and Economics, Naamsestraat 69, 3000 Leuven, Belgium
[b] Vlerick Leuven Gent Management School, Vlamingenstraat 83, 3000 Leuven, Belgium
[c] Amsterdam-Amstelland Police, James Wattstraat 84, 1000 CG Amsterdam, The Netherlands
[d] Universiteit van Amsterdam Business School, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Formal Concept Analysis (FCA)
Domestic violence
Knowledge discovery in databases
Text mining
Exploratory data analysis
Knowledge enrichment
Concept discovery

ABSTRACT

The types of police inquiries performed these days are incredibly diverse. Often data processing architectures are not suited to cope with this diversity since most of the case data is still stored as unstructured text. In this paper Formal Concept Analysis (FCA) is showcased for its exploratory data analysis capabilities in discovering domestic violence intelligence from a dataset of unstructured police reports filed with the Amsterdam-Amstelland police in the Netherlands. From this data analysis it is shown that FCA can be a powerful instrument to operationally improve policing practice. For one, it is shown that the definition of domestic violence employed by the police is not always as clear as it should be, making it hard to use it effectively for classification purposes. In addition, this paper presents newly discovered knowledge for automatically classifying certain cases as either domestic or non-domestic violence. Moreover, it provides practical advice for detecting incorrect classifications performed by police officers. A final aspect to be discussed is the problems encountered because of the sometimes unstructured way of working of police officers. The added value of this paper resides in both using FCA for exploratory data analysis, as well as with the application of FCA for the detection of domestic violence.

1. Introduction

Concept discovery is a relatively new approach for discovering knowledge from textual information (Poelmans, Elzinga, Viaene, & Dedene, 2010). At the core of the method is the visualization of the underlying concepts of the data by means of Formal Concept Analysis (FCA) lattices (Ganter & Wille, 1999; Wille, 1982) which are interpreted, analysed and discussed by domain experts. FCA arose twenty-five years ago as a mathematical theory (Stumme, 2002) and has over the years grown into a powerful framework for data analysis, data visualization (Eklund et al., 2004), information retrieval and text mining (Carpineto & Romano, 2005; Godin, Gescei, & Pichet, 1989; Priss, 2000). In this paper FCA is for the first time used as an exploratory data analysis and knowledge enrichment technique for police data. Compared to traditional black-box data mining techniques, this human-centred approach has the advantage of actively engaging expert knowledge in the discovery process.

The goal of Intelligence Led Policing (ILP) is to complement intuition led police actions with information coming from analyses on aggregated operational data, such as crime figures and criminal characteristics (Collier, 2006; Collier, Edwards, & Shaw, 2004; Viaene et al., 2009). While over 80% of all information available to police organizations resides in textual form, analysis has to date been primarily focused on the structured portion of the available data. Though text mining has been identified as a promising area in the formal framework for crime data mining by Chen et al. (2004), this work has hardly found its way into mainstream scientific literature. One of the notorious exceptions is the paper by Ananyan (2002) in which historical police reports were analysed to identify hidden patterns.

According to the Ministry of Justice of the Netherlands, 45% of the population once fell victim to non-incidental domestic violence and for 27% of the population, the incidents even occurred on a weekly or daily basis (van Dijk, 1997). These gloomy statistics brought this topic to the centre of the political agenda and made it to one of the pivotal projects of the Balkenende administration when it took office in 2003[1] and the Amsterdam-Amstelland police in the Netherlands (Politie Amsterdam-Amstelland (2008)). Sufficient insight into the nature of domestic violence, being able to swiftly recognise suspicious cases and label reports accordingly is of the utmost importance. However, in the past intensive audits of the police databases related to filed reports established that many

* Corresponding author.
E-mail addresses: Jonas.Poelmans@econ.kuleuven.be (J. Poelmans), Paul.Elzinga@amsterdam.politie.nl (P. Elzinga), Stijn.Viaene@econ.kuleuven.be (S. Viaene), Guido.Dedene@econ.kuleuven.be (G. Dedene).

[1] http://www.regering.nl/Het_kabinet/Eerdere_kabinetten/Kabinet_Balkenende_II/Regeerakkoord#internelink4.

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.103

reports tended to be wrongly labelled as domestic or as non-domestic violence cases.

In this paper we shall demonstrate the effectiveness of concept discovery methods for distilling new knowledge from the unstructured text in police reports. FCA amongst others helped us to improve the definition, the understanding by police officers and the management of the notion of domestic violence. Additionally, we aim at automating detection of domestic violence from the unstructured text in police reports. The very first steps taken in this direction are described in Poelmans, Elzinga, Viaene, and Dedene (2008) and Poelmans, Elzinga, Viaene, Van Hulle, and Dedene (2009) an independent research track pursued in parallel with the work presented in this paper based on Emergent Self Organizing Maps is described. Although the usage of FCA for browsing text collections has been suggested before by Cole and Eklund (2001) and Cole (2000), almost none of these papers have focused on how FCA can be used for knowledge enrichment and for discovering different types of knowledge in unstructured text. Neither has it been thoroughly discussed in the literature how FCA can be used to incrementally construct and refine a high-quality domain-specific thesaurus (which is a prerequisite for developing an effective information retrieval system). Moreover, only minor attention has been paid to the possibilities offered by FCA to incorporate prior knowledge in the knowledge discovery process. Finally, some of the aspects of this paper have already been discussed in the literature in a fragmented way (e.g. information retrieval, knowledge browsing), but an integrated approach has never been pursued.

FCA is particularly suited for exploratory data analysis because of its human-centredness. Representations that expose the underlying conceptual structure of the information promote the creation of new knowledge. What makes FCA an especially appealing technique for knowledge discovery in databases from a practitioner's 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. Concepts are the elementary units of human reasoning and this notion of concept is central to FCA (Arnauld & Nicole, 1985; Peirce, 1992). The underlying structure of the information is considered to be a concept system and FCA concept lattices are used to visualize the concepts and their interrelationships. These visual representations support human actors in their information discovery and knowledge creation exercise.

This paper is composed as follows. In Section 2 we describe the current situation of the domestic violence reporting procedure and previous attempts to improve the situation. In Section 3 we cover the essentials of FCA theory, introducing the pivotal FCA notions of concept and concept lattice and describing the process of FCA for knowledge discovery. Section 4 elaborates on the dataset used in our research, while Section 5 focuses on how this dataset was analysed and discusses the results of the application of FCA for exploratory analysis of domestic violence cases using this dataset. In Section 6 the results of the domain exploration are validated. Finally, Section 7 presents a number of concluding remarks.

2. Domestic violence discovery

According to the US Office on Violence against Women, domestic violence is a "pattern of abusive behavior in any relationship that is used by one partner to gain or maintain power and control over another intimate partner" (Office on Violence against Women, 2007). Domestic violence can take the form of physical violence, which includes biting, pushing, maltreating, stabbing or even killing the victim. Physical violence is often accompanied by mental or emotional abuse, which includes insults and verbal threats of physical violence towards the victim, the self or others, including children. Domestic violence occurs all over the world, in various cultures (Watts & Timmerman, 2002) and affects people throughout society, irrespective of economic status (Waits, 1985).

2.1. Current situation

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

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

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

A summary of the current domestic violence reporting procedure is displayed in Fig. 1. There are several problems associated with this process. First, recent audits have confirmed that many of the retrieved cases are wrongly selected for in-depth manual inspection. Going back to 2006, the system retrieved 1157 cases, 80% of which actually turned out to be non-domestic violence cases. For example, going back to 2007, the triage system retrieved 1091 of such cases in which the victim made a statement to the police. Second, because of a lack of manpower the data management quality team was not able to analyse each retrieved police report. Third, audits of the police databases revealed that not all domestic