UNIVERSITY OF GOTHENBURG

Exploring the Potential for using Artificial Intelligence Techniques in Police Report Analysis:
A Design Research Approach

Bachelor of Science Thesis in the Programme Software Engineering and Management

AMADEUS HEIN
FREDRIK BENGTSSON

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
Göteborg, Sweden, May 2011

Exploring the Potential for using Artificial Intelligence Techniques in Police Report Analysis:
A Design Research Approach

Amadeus Hein
Fredrik Bengtsson

Examiner: Helena Holmström Olsson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden May 2011

BACHELOR THESIS

Exploring the Potential for using Artificial Intelligence Techniques in Police Report Analysis:

A Design Research Approach

Abstract

Data volume and complexity is increasing as technology advances. Therefore, manual data analysis suffers from increased time spent on analysis and an increased risk of errors. Previous research suggest artificial intelligence as a potential aid to these issues, and to explore these challenges this paper takes a design research approach. This study collaborates with the Swedish police to illustrate what steps can be taken to ease data analysis and develops a software prototype to aid human data analysts in their work.

Keywords: data analysis, volume, complexity, data mining, artificial intelligence, neural networks, criminology.

Amadeus Hein

Fredrik Bengtsson

University of Gothenburg

1 Introduction

Storage of data is no longer a problem, due to technical advancements in computing power and bandwidth, the problem is instead how we should use all the data we collect (Shoan & Woolf 2008). One drawback from this is that data analysis requires more time to complete than before (Chen et al. 2004; Liang & Austin 2005) as the data volume grows exponentially over time (Rajagopalan & Isken 2001). In addition to the growing data volume, another challenge is to correctly analyze complex data, as it is often difficult to interpret (Nath 2006). Data mining is one field, linked with artificial intelligence (AI) (Williams 1983), that strives to address these challenges (Liang & Austin 2005), saving time for the analyst (Charles 1998; Chen et al. 2004). However, mining complex data is difficult and often requires a skilled data miner and an analyst with good domain knowledge in the area of analysis (Nath 2006) to ensure a low rate of human errors (Chen et al. 2004; Charles 1998).

This highlights an opportunity in programmatically reducing the data volume to only include data that is relevant to the analysis. Completion of such procedures also becomes less time consuming when the raw data has been pre-processed. Additionally, software may be used to replicate repetitive existing human analysis steps in order to provide human analysts with higher quality

data to work with, e.g. showing data trends, narrowing the frame of analysis, and making decisions easier through suggesting likely beneficial answers. Examples of similar software already exist today. The Coplink project (Chet et al. 2004) uses data mining techniques to map criminal networks and assist the police by providing an overview of a criminal's connection to other people, while ReCAP (Charles 1998) analyzed crimes and showed at which times of day they occurred to guide police efforts. Research extends beyond the public safety and the police, also, including for medical data analysis and business forecasting (Liang & Austin 2005). An example of such use in the medicine field is presented by Mantzaris & Anastassopoulos (2008), where they look to help clinicians identify individuals with increased risk for osteoporosis that need to undergo further testing and treatment.

The intention of this paper is to explore the potential for volume and complexity of large data sets to be negotiated and leveraged for additional benefits through software implementation using AI techniques. Specifically, this is done by investigating the research question *how can artificial intelligence techniques be used to assist in identifying data trends that are likely to be of relevance for further investigation by human agents?* To address this, the study uses the increasingly recognized design research approach (cf. Hevner et al. 2004; Chatterjee & Hevner 2010) which is rooted in a desire to address practice and research interests through design artifacts. This strategy also allows us to iteratively assess and revise our design to better fit with the specific needs of our industrial partners. Thus, learning-by-doing (Jeffries et al. 1981) and reflection-in-action (Schön 1983) are important guiding principles for our emerging design and the contributions made by this study.

Since this paper follows a design research strategy, one research outcome is a software prototype. The prototype developed in this study uses a set of AI techniques to assist human agents in data analysis. In the background section of this paper we illustrate a matrix model that explains the relations between complexity and volume in a data set, and shows which considerable paths exist in the process of making data more easy to analyze. This paper starts by explaining why volume and complexity play a vital role in data analysis and illustrates it in the Complexity Volume matrix model (CVmatrix). In the subsequent section the method is introduced and justified, and the research setting explained fully. The paper concludes by covering the research outcomes, which includes both results and discussions of the research outcomes.

2 Theoretical Background

For the purposes of this paper, two attributes for data analysis are of particular interest: volume and complexity. Volume is the raw amount of data being analyzed, while complexity is a definition of how difficult the data is to interpret. The following two subsections define these two attributes, and establish why they both play a vital role in performing data analysis. In the final subsection, a matrix model based on the definitions of volume and complexity is built to provide the theoretical framework of this paper.

Data Volume

In the past decades the world has entered into the information age, resulting in exponential increases in computing power and communication bandwidth (Shoan & Woolf 2008). In turn, this has resulted in a rise of the amount of data that can be accessed (Rajagopalan & Isken 2001), and also increasing the time it takes to perform data analysis. Organizations involved in data analysis, such as intelligence-gathering

and medical data analysis, have to face the challenge of accurately and efficiently analyzing the growing data volume (Chen et al. 2004; Liang & Austin 2005). One reason for the challenges is that the relevant data is often hidden in a larger set of irrelevant data, making it difficult to find. For example, imagine analyzing network traffic with its frequent and busy online transactions, where only a small portion of the data is related to the analysis (Chen et al. 2004).

Data mining is one field that strives to address the challenges of the growing data volume. Data mining is a procedure that efficiently extracts information from large data sets, linked (Liang & Austin 2005) with artificial intelligence (Williams 1983). AI techniques are capable of analyzing vast amounts of information, with a low error rate, and can process this in seconds, saving time for the analyst (Charles 1998; Chen et al. 2004).

Data Complexity

Complex data may be differently expressed and structured in different data sets, changed periodically, generally diffuse, and/or span long periods of time (Tanasescu, Boussaid & Bentayeb 2005; Chen et al. 2004). For example, data that is collected for reasons other than analysis limit the analyst's ability to extract meaningful output (McCue 2006). The definition of complex data implies that complexity have different meanings depending on the data itself.

The challenge deriving from complexity in data is to correctly analyze it, given that complex data is more difficult to interpret (Nath 2006). Humans are more likely to see complex patterns in data sets than through the use of AI. However, the more complex the data is, the more time the analysis will take and the risk of suffering from human error increases (Chen et al. 2004; Charles 1998).

Mining complex data is a demanding process, and often requires not only a skilled data miner, but also an analyst with good domain knowledge in the area of analysis (Nath 2006). Before the actual data mining can begin the data has to be prepared. This includes collecting, cleaning and transforming the data to fit the purpose of the usage (Rajagopalan & Isken 2001). The challenge of mining complex data becomes evident when approximately 80% of the data mining process is spent on preparing the data for the actual mining (Helberg 2002; McCue 2006).

Illustrating the relationship between complexity and volume

Based on the definitions made of the two data attributes, volume and complexity, it is possible to represent the relationship between them and their data sets in a Complexity Volume matrix (CVmatrix (step one), figure 1).

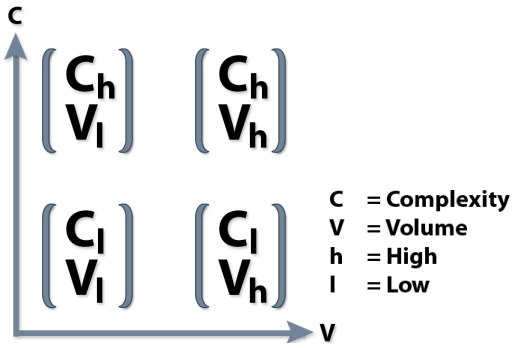

***Figure 1:** CVmatrix (step one)*

The CVmatrix consist of four states (represented as vectors) of volume and complexity that a data set may exist in. The suitability for one particular agent (human or AI) to analyze a data set depends on what state it is in. A human is good at analyzing complex data in smaller volumes, while an AI can handle large volumes of less complex data more efficiently (Charles 1998). This relationship is represented in figure 2. When the data set is both complex and contains a large volume of data the risk of errors are

highest (Chen et al. 2004). Thus, the ideal state for either actor is when the volume and complexity of the data set is low. Finding a method for reducing either or both attributes would thus make the analysis of the data set more efficient and more correct. Returning to our research objective, it appears that AI has potential to handle large data volumes to assist human agents, while humans are better at reducing complexity. Step two of the CVmatrix (figure 3) illustrates the possible paths for simplification of a data set.

Figure 2: Agent relationships

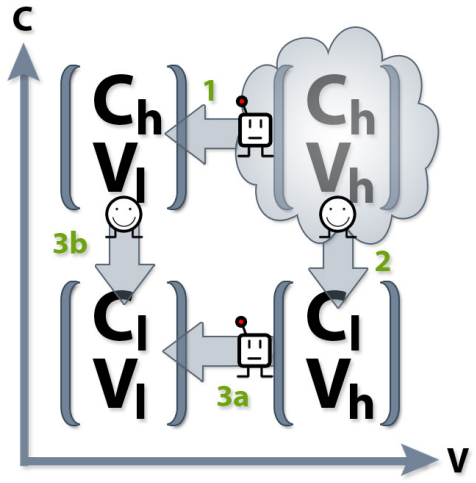

Figure 3: CVmatrix (step two)

Each path (1, 2, 3a and 3b) in the CVmatrix is assigned an agent that has the highest potential in reducing the specific attribute. The illustration visualizes a human reducing the data complexity (2 and 3b), and an AI reducing the data volume (1 and 3a). Traversing path 1, an AI is assigned to reduce the volume of a complex and large data set. The objective is to reduce the risk of errors and increase the efficiency of a human agent. From a complex, but small, data set a human can either choose to start the analysis or attempt to reduce the complexity to move to the ideal state of the data set (path 3b). Following the opposite way (path 2 and 3a) a human agent is assigned to reduce the complexity in a large and complex data set in order to assist an AI and make it easier for the AI to reduce the volume. However, this way is not optimal as the more volume, the more time it takes for a human and the risk of errors increases (Chen et al. 2004).

Arriving at the ideal state is a two-step process in which the human and AI collaborates. By arriving at the ideal state fewer errors are made, and results in more efficient analysis. Naturally, it is possible for the data set to initially be in any of the states, and from there one can apply the specific path to make the data set easier to analyze. For example, a data set might be low on complexity by default but have high volume. Thus it is preferable to take step 3a to reach the ideal state. However, this scenario does not satisfy the purpose of the study, instead the focus will be set on moving from the state where most errors are made (indicated by the cloud in figure 3), because if it is possible to move from this state, the possibility of moving from other states increases. We want to show all the steps we have identified through related literature, but focus on exploring an approach to traverse path 1.

The CVmatrix shows which agent should be assigned to reduce a specific attribute in a data set. Relating to our research objective, the CVmatrix illustrates the assisting roles an artificial intelligence may have in aiding manual data analysis, by manipulating the data set, specifically in trying to move the attributes of the data set towards a point where the risk of errors is low. While the CVmatrix is a contribution in itself, this paper will focus on exploring how AI can reduce volume in a complex data set, visualized in path 1 of the CVmatrix.

3 Method

Research Setting

Settings dealing with complex data sets in large volumes are intelligence units of the police. Specifically, this study collaborates with the intelligence unit of Gothenburg, Sweden. Intelligence units deal with growing volume and complexity of data, which is typical for organizations involved in crime analysis (Chen et al. 2004). Nowhere are the data volume issues more evident than in the amount of police reports that were added each day in 2009. An average of 607 police reports were filed each day, totaling in 221 708 for the whole year, in the Västra Götaland region of Sweden alone[1]. The Swedish police have a clear directive to work against 'crimes of quantity', such as burglary, physical abuse, vandalism and more. Lately, the Gothenburg police have acknowledged a particular issue with 'roaming burglary'. Roaming burglary describes crime in which the criminals rapidly travel around the country striking around their path. One of the fundamental difficulties with such crimes is that the criminals are constantly in motion, and the police detect it too late. The criminals have most likely traveled to another location and this impedes the preliminary investigation, and some inquiries might even have been abandoned altogether. For example, statistics in Sweden show that only 4% of burglaries were resolved in 2009[1]. This is a low number, which media is also reporting as a problem from a societal perspective.

Many burglaries are part of organized crime. One example of organized crime within burglary is criminal organizations that get special orders on specific items to burgle. Often, organized crime moves around the country and eventually the offenders leave to sell the goods in another country. This makes it difficult for the police to track down and apprehend the offenders before they are gone. Part of the problem is that the intelligence analysis takes days, and sometimes weeks, before patterns in burglaries may be observed. This is mostly caused by the volume and complexity of the data, often spanning several years back, being analyzed manually to draw conclusions and make analysis reports.

In response to this, the intelligence unit in Gothenburg is interested in exploring the development of a software prototype (Sundhage & Lindgren, 2010). The prototype (Sherlock) developed during the study is intended as an analysis tool using AI techniques to analyze and identify patterns in police reports. Relating back to the research objective, Sherlock follows path 1 in the CVmatrix (figure 3, theoretical background section), by reducing the volume in a highly complex and large data set. Sherlock, and its development, is further discussed and elaborated on in the next section (4).

Research Design

This paper takes a qualitative strategy (Creswell 2003), using interpretation of data along with an iterative design research (Hevner et al. 2004) strategy to the research and design of Sherlock. The analysis process investigated is used by an organization in which the data analysts serves as this study's unit of analysis (Sundhage & Lindgren 2010), explained in the research setting section.

Design research is characterized by its intention of improving the state-of-the-world, by for example improving the efficiency of an organization (Hevner et al. 2004). As our knowledge within the criminology domain is low, it is valuable to use the iterative nature of design research (Kuechler & Vaishnavi 2007). Thus, learning-by-doing (Jeffries et al. 1981), and reflection-in-action (Schön 1983) principles are allowed that enable iterative assessment and revision of our design as the study matures. This strate-

[1] *Statistics taken from BRÅ, Brottsförebyggande Rådet (www.bra.se/)*

gy makes it possible to develop a prototype for our industrial partner, giving not only an academic contribution but an industrial contribution as well.

This study combines design research and qualitative strategies. Design research and qualitative strategies are suitable strategies for research in novel areas, due to the iterative and exploratory nature of both strategies where knowledge may be gradually and collaboratively developed through the design process (Kuechler & Vaishnavi 2007; Creswell 2003).

Qualitative research is characterized by its interpretive and iterative nature, taking place in practice with participant involvement encouraged through open-ended interviews (Creswell 2003; Wolcott 1994). As our knowledge in the criminology area is low, it is possible, by using qualitative research, to adapt the data collection methods while new knowledge is gained. Interaction in the industrial setting gives us the benefit of frequent face-to-face meetings and interviews at the convenience of the industrial partners. Open-ended interviews, with some general questions prepared, is used to gather data on people's reflections, experience, local knowledge and practices (Myers & Newman 2007). This affords understanding of the industrial challenge and the domain of analysis.

Our research objective focuses on exploring the role an AI can have in assisting human agents, it is an area which is difficult to find in related literature or case studies. The qualitative strategy allows us to learn as we advance, while the study matures, and to learn from the process itself.

Data Collection

In order to facilitate an understanding of the industrial challenge and to gain domain knowledge, the interviews are open-ended as it's important that the interviewees are able to guide the discussions. The interviews are semi-structured, and follow a theme so that the discussions are relevant to the study. For example, one theme was on how to establish a *modus operandi* (the method of a crime), which data to look at and how to analyze it. We conducted six one hour interviews, involving a total of four intelligence analysts, two present at each interview. In addition we had a two hour interview with a system administrator to gain technical understanding of the industrial setting. The interviews are used in this paper throughout the design iteration, and are presented and discussed in section 4 as they become relevant. Each interview is transcribed to make the analysis of the collected data easier. In addition to the interviews we kept field notes on observations made during our time at the intelligence department, as we did the implementation of Sherlock on site.

Myers and Newman (2007) note that qualitative interviews come with a series of problems and pitfalls. Thus we prepare an interviewing strategy to deal with them. The interviews are conducted in the interviewee's office together with another colleague in order to make the interviewee feel more confident and comfortable in the interview environment. To deal with time pressure we make sure to have follow-up interviews in which the interviewees are able to add or deduct to the previous interview. By involving multiple analysts we make sure to gather a shared view of their analysis processes, and since they share the same status within the organization there is no risk of elite bias. Since we and the interviewees stem from different domains (software engineering and criminology respectively) the interview questions will not involve any technical software terms, and any criminology terms brought up will be thoroughly explained.

We use a literature review to facilitate theory development and collect data in areas that are found through both the design process and interviews that we need to elaborate on (Webster & Watson 2002). We focus the literature search within top-rated journals, as it is more likely major contributions are made there (Webster & Watson, 2002), and then search the reference list in interesting articles for more literature.

4 Discussion of research process and outcomes

This section combines the results and the discussion of this study. The research approach taken is influenced by an iterative design research strategy. Thus, to make this easy to follow, we organize the following subsections into each iteration of the design process the study explores. We include the diagram (figure 4) of the design research cycle, adopted from Kuechler and Vaishnavi (2007), where it is illustrated which steps are taken, and what flows the iterations can take.

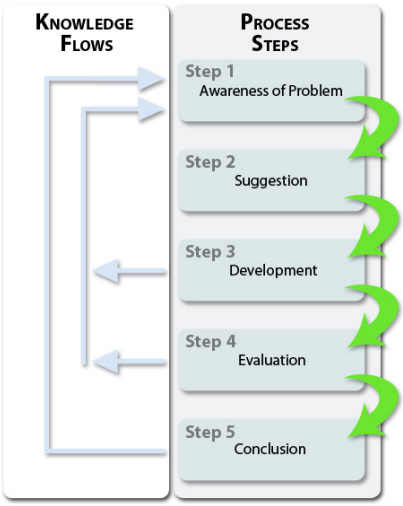

***Figure 4:** Design Cycle*

As illustrated in the diagram, the design research process is based on iterations, and our research process consists of three iterations. In the following subsections, named after each iteration, the research outcomes are explained and discussed. The last subsection, reflections on the iteration outcomes, summarizes the discussion and gives an overview of the paper's contributions.

Iteration 1: Preperation

In the first iteration we interviewed the intelligence analysts at the intelligence unit of the Gothenburg police. These interviews were used to gain an understanding of the industrial challenge and to learn what data is used in crime analysis. We discovered that the *modus operandi* (a criminal's method of operating) of a crime has a major impact on identifying crime series. For a burglary, the *modus operandi* reveals the entry method and crime scene behavior of the offender. The literature review was focused on exploring related literature within the fields of criminology to understand how the analysis process works in theory. We also investigated artificial neural networks as it became evident in previous research that this technique could be used for crime analysis (Charles 1998; Chau et al. 2002; Chen et al. 2004).

After investigating artificial intelligence techniques, we established that neural networks have potential to assist the intelligence analysts, as this technique offers complex pattern recognition (Heaton 2008). Thus, we decided that the foundation of Sherlock should be based on neural networks. We focused the initial development on preparing the data for neural network analysis, as the preparation process plays an extensive role of the whole analysis procedure (Helberg 2002).

Throughout the development step we designed and implemented a way of normalizing all crime data to be suitable as input for the neural network. We created a protocol that addressed each of the relevant crime attributes which made it possible to

								Gender				
$Crime_1$	1	1	2	1	1	4	5	2	630529	4	4	0
$Crime_2$	1	1	2	1	1	4	2	2	0	0	0	0
$Crime_3$	2	4	4	5	1	2	1	1	0	0	2	0
$Crime_4$	2	4	4	5	1	2	1	1	0	0	1	0

***Table 1:** Normalized Crimes*

normalize the report data. The normalizing process' primarily objective is to translate the relevant data from police reports to be suitable as input for the neural network. Because the details of the protocol are confidential, we will only give one example of it. Each attribute can be assigned a predefined amount of values. In turn, each of these values has a unique meaning, for example: The attribute 'Gender' can be assigned value '0', '1' and '2' which is translated to 'unknown', 'man' and 'woman'. Looking at table 1, it is illustrated that the two first crimes were made by women, and the last two were made by men. After normalizing the data the intention was to compare two normalized crimes using them as input for the neural network.

Iteration 2: Development

Through the second iteration, interviews were held to understand how to look at the data the analysts presented in the first iteration. This includes the analysts' explanation of how to establish a *modus operandi*, what to look for, and what impact it has compared to other crime factors. We discovered that extracting meaningful data from the police reports is a complex and highly demanding task. As emphasized by Chau et al. (2002), Nath (2006) and Bache et al. (2007), the open-entry field (also called free text or narrative) of the police report contains much valuable information, but is difficult to interpret because the data may be differently expressed and structured depending on who writes it. As already mentioned in the theoretical background section, this is one definition of complex data (Tanasescu, Boussaid & Bentayeb 2005). We decided that interpretation of the open-entry field is important for the purposes of the software prototype as it may provide additional data for the analysis and further increase the accuracy of the analysis. Thus, during the second iteration we broadened our knowledge of data mining to discover an efficient way of retrieving the relevant data from a complex data set.

We decided to rely on a neural network architecture called Self-Organizing Map (SOM), which is commonly used for classifying patterns. This architecture is based on unsupervised training, meaning that it trains itself in order to indicate what kind of pattern was introduced (Heaton 2008). Thus we set out to explore its potential of analyzing police reports to find crime patterns. Subsequently, we realized that neural networks could not fulfill the entire intention of the prototype. The data from the police reports had to be pre-processed and filtered (Goodwill & Alison 2006) by Sherlock. The neural network should then analyze the pre-processed and filtered data to suggest to the intelligence analysts whether they should look into the police reports more thoroughly or not. Thus, we decided that Sherlock required a way to extract data from the open-entry fields of the police reports, which was vital for the network to produce relevant output.

As mentioned earlier, a major challenge in using AI for analysis of police reports is to interpret the open-entry field. An open-entry field is where the police officer freely writes the details of a crime. Unfortunately, as discovered through the interviews and literature (Chau et al. 2002; Nath 2006; Bache et al. 2007), the most relevant data is hidden within the open-entry field and is difficult to find as the report can be differently structured and expressed. The relevant data may in the worst case be implicitly expressed, making the interpretation more challenging to perform with software (Nonaka & Takeuchi 1995). For example, if the open-entry field contains "The plaintiff was asleep when the burglary took place", a human would assume that the crime was

performed during the night. However, these tacit meanings are difficult for software based systems to understand (Nonaka & Takeuchi 1995). Interpreting tacit meanings is not within the scope of this study and would suit better in its own research report. We decided that for the purposes of this paper enough information about a crime is gained without interpreting tacit information, and still illustrates the potential of Sherlock.

To address the challenge of interpreting open-entry fields, we implemented an open-entry interpreter, based on a lexical-lookup (Chau et al. 2002) extraction approach, adapted to establish a *modus operandi*. The lexical-lookup approach matches and extracts keywords in open-entry fields

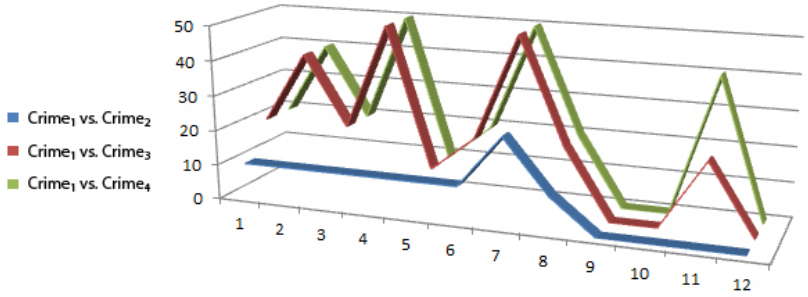

***Figure 5:** Merged Crimes (Comparing to Crime$_1$)*

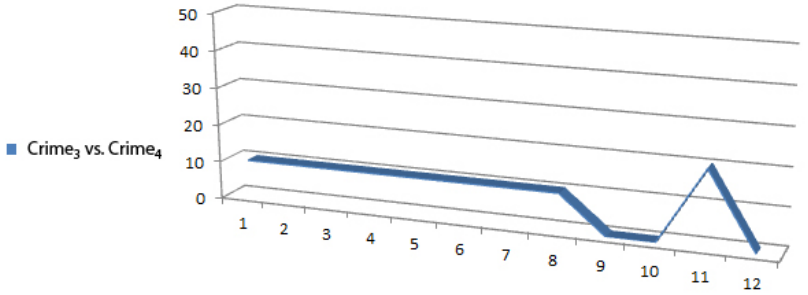

***Figure 6:** Merged Crimes (Comparing to Crime$_3$)*

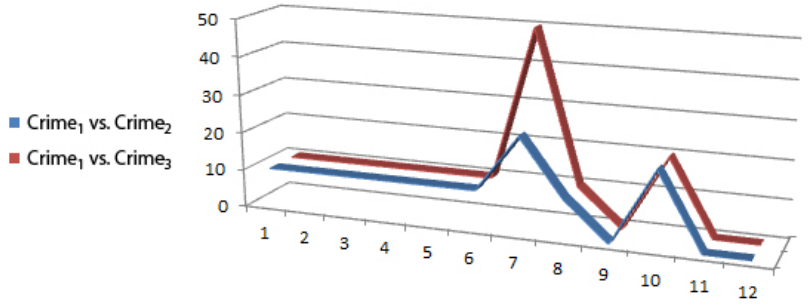

***Figure 7:** The Merged Crime Problem*

(Chau et al. 2002), these keywords were defined in collaboration with the intelligence analysts. We implemented the open-entry interpreter to softly match keywords, using an algorithm called Q-gram (Younghoon et al. 2010), to match bending of words and misspelled words. In order to establish a *modus operandi*, the algorithm initially searches for a keyword, for example 'door' or 'window', then it tries to find additional keywords in the vicinity of that word, for example, 'brake' or 'drill'. The more matches the interpreter finds in the vicinity of a keyword, the more that group "scores", and the group with the highest score is determined as the winner, and thus becomes the crime's *modus operandi*.

The neural network input structure, developed in iteration 1, consisted of two normalized crimes that the network compared. This resulted in an unlimited amount of different patterns, as if you change the positions of two crimes a new pattern would be generated even though that is not intended. We realized that there was a need of modifying this input structure in order for the neural network to find crime patterns. The proposal was to create an algorithm that merged two normalized crimes and generated a pattern that would always look similar when two relatively homogeneous crimes were compared. Additionally, when two relatively heterogeneous crimes were compared, the generated pattern should be differentiating from the one representing the similar crimes. This would make it easier for the neural network to recognize whether the crimes are related or not. Figure 5 shows the different patterns that can be generated by comparing each crime ('Crime$_2$', 'Crime$_3$' and 'Crime$_4$'), from table 1, to the first crime ('Crime$_1$').

Illustrated in figure 5 is when comparing the two first crimes ('Crime$_1$' and 'Crime$_2$') the pattern has a flat overall structure, with a major amount of '10's (the chosen ideal

value). This indicates that the chance of the two crimes belonging to the same series is high. Comparing the first crime ('$Crime_1$') with the rest of the crimes ('$Crime_3$' and '$Crime_4$') generates patterns that differentiate, meaning that they are most likely not related to the first crime ('$Crime_1$'). However, notice that the last two crimes ('$Crime_3$' and '$Crime_4$') have a relatively identical structure. Comparing them with each other would generate the pattern shown in figure 6.

The pattern illustrated in figure 6 indicates that the two crimes ('$Crime_3$' and '$Crime_4$') are likely to belong to the same crime series. However, even if the patterns generated by the merge algorithm can look flat, it does not immediately imply that the two crimes belong to the same series. The advantage of using neural networks for recognizing and classifying complex patterns is that the network can be taught over time. This means that the neural network's analysis ability should increase as it is used. Training a neural network involves gathering both training and evaluation data (Heaton 2008). Sherlock's neural network will be trained and evaluated through data retrieved by the intelligence analysts on previous crime series in order to decide what patterns should be regarded as a series. Throughout the development of Sherlock it was important that the intelligence analysts contributed with their perspective of how to analyze and classify crimes. Their role as human agents is to teach the neural network to "think" like they do when they use it in their daily work.

We evaluated the implementations of the SOM neural network and open-entry interpreter and discovered that the SOM network did not function as expected. The network had difficulty mapping crimes due to the input structure we required, and caused random output. This meant that we had to find a more suitable neural network model in the next iteration. We evaluated the open-entry interpreter by running it through 60 pre-analyzed police reports. The results showed that all modus operandi's were correctly structured. Sometimes the information extracted through it was sparse, however, when manually going through those reports it was evident that the report itself lacked all the information expected. When evaluating the merge algorithm it became evident that the patterns generated could contribute with deficient results, as they were constructed partly based on prioritization. This means that when comparing two patterns, some values in one pattern could be very high while in the other pattern relatively low, but still over the ideal value, thus implying that the latter pattern is more likely to be a crime series. This issue is illustrated in figure 7. All values that are not '10' imply that the merge algorithm found attributes in the police reports that were not matching. When the difference from '10' is relatively low, a neural network interprets the attributes of the two crimes as similar, even though in reality they are not. Thus, the neural network might interpret the two patterns differently, meaning that one pattern is identified as a crime series while the other one is not, even though both should be in the same crime series. However, this drawback is not critical for the objective of this study, because it is still possible to show the potential of using Sherlock, as further explained when evaluating in the next iteration.

Iteration 3: Improvement of the neural network model

We discovered during the second iteration that the output of the SOM was difficult to manipulate and interpret, which is also hinted towards in literature (Heaton, 2008). Therefore, we decided that it was important to investigate what other neural network alternatives existed that had potential for the purpose of Sherlock.

The investigation of another neural network model, the feedforward backpropagation network (FB), showed more potential for crime analysis. Feedforward is a method for recalling patterns, and backpropagation is a supervised training method that requires sample inputs and anticipated output (Heaton 2008). By using a supervised training method we can manipulate the neural network during the training session. This means that it is possible to introduce two crimes and tell the FB whether it should interpret it as a belonging to the same crime series or not. Thus, we are able to train it to recognize specific patterns.

The intelligence analysts presented us with a crime series they identified in 2009, containing a total of 58 linked crimes in the same crime series. We evaluated the FB network analysis capabilities by using a training set consisting of a subset (10 crimes) of the crime series presented by the intelligence analysts (58 crimes). The FB network was trained with this subset, and the hypothesis was that it should be able to recognize the rest of the 48 crimes in the series. Subsequently we let it analyze the police report database of burglaries in the Swedish region of Skaraborg 2009 (containing a total of 318 crimes, with the 48 linked crimes included), and the network found 41 crimes that were matching each other. This indicates that there still exists an error rate (17 of 58 crimes) that has to be tuned by further training, but showed that the network model is more suitable than the SOM. The human agent has an important role to play for the reliability of the neural network. As the intelligence analysts train the network with more crime series the network may learn to better determine whether the crimes belong to a crime series, making the error rate decrease. Thus, the reliability of the neural network analysis should improve.

Reflections on the iteration outcomes

From the evaluation of the FB network we have demonstrated that it is possible to reduce the volume of a complex data set through combining a set of AI techniques. Sherlock uses open entry interpretation, a crime merge algorithm and a FB neural network to make it possible to traverse path 1 of the CVmatrix (the circle in figure 8).

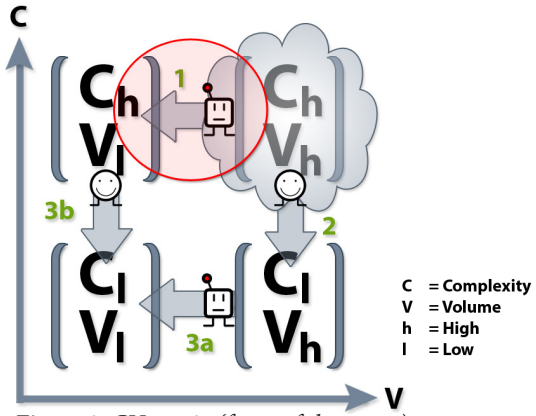

Figure 8: CVmatrix (focus of the paper)

While further tweaking of Sherlock is suggested, the objective of this study is accomplished. This paper set out to show that an AI can identify a data trend that needs further investigation by a human agent, and in doing so reduce the volume of data that needs to be manually analyzed.

Due to the police reports containing confidential data and the diffuse structure of information available it was difficult to collect proper amounts of training and evaluation sets for Sherlock. Heaton (2008) emphasizes that a lot of training data is ideal for neural networks. Thus, if the neural network was trained with more sets of crime series it would likely increase the accuracy of the results. Additionally, there exist features that may further support the neural network analysis. We suggest that exploring how to extract more of the relevant data would have positive effect on the assisting role of the neural network, as the higher quality of the input, the higher quality of the analysis.

While the crime merge algorithm produced patterns that were easy for a neural network to analyze, it had an issue with constructing the patterns by unintended prioritization. For example, if an attribute in one pattern differed by '20', and the other by '50', the neural network assumed that the closer to '10' the more likelihood of a pattern, which was not always the case. We suggest that a solution to this would be to sort all values that each crime attribute can have by priority such that all values that are close to the ideal value ('10') actually, in reality, are similar while values that are highly differing from the ideal value indicates a big difference. For example, prioritization could be made by assigning similar values to similar *modus operandi* values, making '3.1' and '3.4' represent 'break' and 'crush' respectively. Thus, the merge algorithm would generate similar and more flattened patterns.

The open-entry interpreter used by Sherlock is a technique that could be used separately from the neural network to reduce the volume of complex data sets. The interpreter would allow the human analysts to run more specific search queries, for example matching the *modus operandi* of a crime, and by doing so get more precise data to analyze. We suggest that open-entry interpretation is one area that should be researched further, perhaps covering tacit meanings.

Through the iterations we have implemented several AI techniques as a foundation of Sherlock. The final prototype's analysis behaviour can be illustrated as a set of phases that utilize different techniques, as shown in figure 9. To clarify how Sherlock functions a detailed description of each phase is provided in table 2.

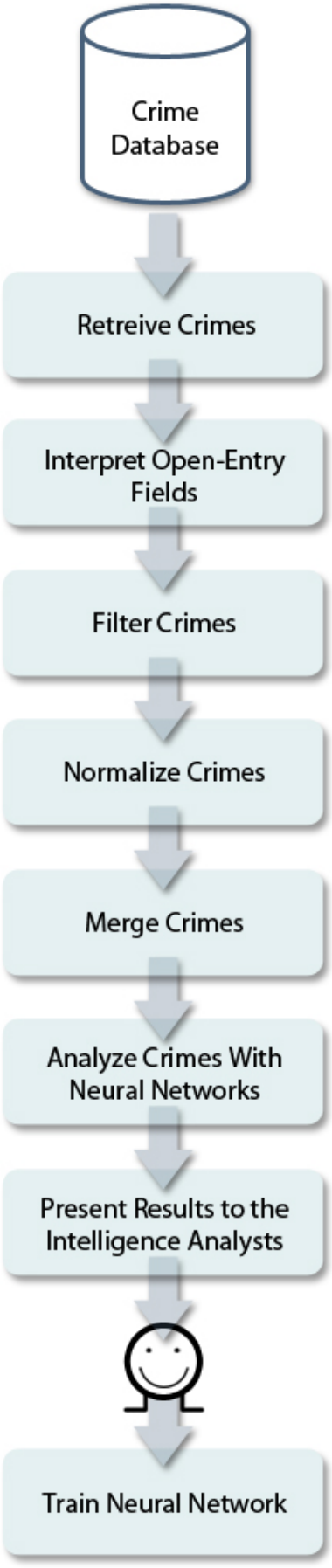

Figure 9: *Sherlock System Overview*

Element	Description
Crime Database	The crime database is the official database used by the police which includes all crime reports and will continuously be appended with new reports as they get reported by police officers. This database contains high volumes of complex data.
Retreive Crimes	In this phase all new crimes are collected that have been reported since the last analysis. This is done within a predefined frame, including only attributes that are likely to be relevant for the particular analysis. The predefined frame is structured based on the intelligence analysts' knowledge and experience in the crime analysis area, meaning that it includes only factors that may be of interest for crime series analysis.
Interpret Open-Entry Fields	Interpreting open-entry fields of police reports is important for the quality of the result of the analysis. By help of lexical lookup, and soft matching with the Q-gram algorithm it is possible to interpret the open-entry field and thus include more data from the police reports, making the analysis more specific and precise. This is elaborated in section 4 – iteration 2.
Filter Crimes	We have implemented a filter based on input from the intelligence analysts, and existing filter models for dividing a crime scene into several aspects such as geo-spatial and temporal aspects (Goodwill & Alison 2006). Each analysis is utilizing a specific filter depending on what type of crime is being analyzed e.g. this study focuses on dwelling burglaries along with additional related aspects as crime scene behavior. In this phase all crimes that do not match the specified filter are ignored while the rest are used in the next phase (Normalize Crimes).
Normalize Crimes	A neural network is limited to only numeric input, thus making it important to convert non-numeric data of the crime reports. This is done by following a predefined protocol, elaborated in section 4 – iteration 1, and shown in table 1, that was developed in cooperation with the intelligence analysts.
Merge Crimes	The neural network architecture used in this study requires a pattern as input. To address this an algorithm was developed with influences of the protocol and was used for performing comparisons between the different attributes of two normalized crimes, generating a pattern that shows how the two compared crimes differ from each other. This is further elaborated in section 4 – iteration 2, and shown in figure 5, 6 and 7.

Table 2: Sherlock Analysis Phase Description

Element	Description
Analyze Crimes with Neural Networks	In this phase the generated pattern is analyzed with the chosen neural network architecture, Feedforward Backpropagation Neural Network, to produce an output that may help the intelligence analysts in further investigation of crime series analysis. Additionally, how the patterns are interpreted by the neural network is dependent on the input given by the intelligence analysts during the training phase (see section 4 – Iteration 2 & 3).
Present Results to the Intelligence Analyst	The result will be presented to the intelligence analyst as a number representing the percentage of how likely the neural network believes it is that the compared crimes belong to the same crime series. Similar approaches exist within the medicine field today, where it is used for helping clinicians identify which people are at increased risk for osteoporosis and should therefore undergo further testing with bone densitometry (cf. Mantzaris et al. 2008).
Intelligence Analyst	The intelligence analysts choose if they are to look deeper into the crime reports or not based on the neural network output given and decides what outputs are accurate or not.
Train Neural Network	Based on the decisions made by the intelligence analyst the neural network will be trained to produce more accurate results. For example, if the neural network indicated that the crimes compared did not belong to the same crime series while the intelligence analysts did, then the neural network will be trained to, for future analysis, indicate that these crimes should be interpreted as belonging to the same series.

Table 2: Sherlock Analysis Phase Description

5 Conclusion

This study set out to explore the assisting role artificial intelligence (AI) may have in identifying data trends that are likely to be of relevance for additional investigation by human agents. In order to investigate our research objective we used a qualitative design research strategy, as it is a suitable strategy for research in novel areas. A setting which deals with large volumes of complex data is the police. Specifically, this study collaborates with the intelligence unit police department of Gothenburg to develop an analysis tool prototype using AI techniques.

To understand the challenge of reducing volume and complexity we created a matrix model (the CVmatrix, figure 8) based on related literature. The CVmatrix illustrates the roles human and AI agents may have in reducing volume and complexity. Subsequently we decided to solely focus on reducing the volume of a complex data set, following path 1 of the CVmatrix (figure 8), as that path can be seen as the most challenging. Following this path, we developed an analysis tool prototype (Sherlock), which showed that it was possible for an AI to reduce volume in a complex data set to assist human agents. Sherlock uses a series

of different techniques to assist a human agent. By using the implemented open-entry interpreter, it is possible for Sherlock to extract the relevant information from the open-entry field of police reports. Sherlock can then normalize the extracted information and apply the crime merge algorithm to generate a suitable input for its neural network. Sherlock uses a feedforward back-propagation neural network to analyze police reports and see what crimes may belong to each other.

While Sherlock showed a potential in finding crime patterns, our evaluation showed that there exists an error rate of approximately 29,3%. Thus, Sherlock's neural network requires further training and evaluation to gain more accurate results. For future research it would be interesting to delve deeper into interpreting tacit knowledge and extracting more information from police reports in order to gain more data for the neural network, which in turn should reduce the error rate.

Acknowledgements

The authors wish to acknowledge the patience and support from our university supervisor, **Carl Magnus Olsson**. We also like to thank the intelligence analysts, our industrial supervisors, **Carin Sundhage** and **Marianne Saether** and the system administrator **Julita Hein**. Their continuous participation made this study possible.

References

Bache, R, Crestani, F, Canter D, Youngs, D 2007 'Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes', *IAS '07 Proceedings of the Third International Symposium on Information Assurance and Security*, IEEE Computer Society Washington, DC, USA, pp. 399-404.

Charles, J 1998, 'AI and law enforcement', *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 1, pp. 77-80.

Chau, M, Xu, J, Chen, H 2002, 'Extracting meaningful entities from police narrative reports', *Proceedings of the 2002 annual national conference on digital government research*, Digital Government Society of North America, pp. 1-5

Chen, H, Chung, W, Xu, J.J, Wang, G, Qin, Y, Chau, M 2004, 'Crime data mining: a general framework and some examples', *Computer*, vol. 37, no. 4, pp. 50-56.

Creswell, J W 2003, *Research design: qualitative, quantitative and mixed methods approaches*, 2nd edn, Sage Publications Inc, London.

Goodwill, A, Alison, L 2006, 'The development of a filter model for prioritising suspects in burglary offences', *Psychology, Crime and Law*, vol. 12, no. 4, pp. 395-416.

Heaton, J 2008, *Introduction to neural networks for java*, 2nd edn. Heaton Research, Chesterfield.

Helberg, C 2002, *Data mining with confidence – SPSS* (2nd edn). SPSS Inc, Chicago, Illinois.

Hevner, A. R, Chatterjee, S 2010, *Design research in information systems: Theory and practice*, Springer, New York; London.

Hevner, A. R, March, S, Park, J, Ram, S 2004, 'Design science in information systems research', *MIS Quarterly*, vol. 28, no. 1, pp. 75-105.

Jeffries, R, Turner, A. A, Polson, P. G, Atwood, M. E 1981, 'The processes involved in designing software', in Anderson, J. R. (ed.), *Cognitive Skills And Their Acquisition*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp, 255-283.

Johnson-Laird, P. & Byrne, R 2000, A gentle introduction, Mental models Website, School of Psychology, Trinity College, Dublin (available at www.tcd.ie/Psychology/Ruth_Byrne/mental_models/).

Kuechler, W. and Vaishnavi, V. (2007). 'Design [science] research in IS: A work in progress' *Proceedings of 2nd International Conference on Design Science Research in Information Systems and Technology (DESRIST '07)*, Pasedena, CA, May 13-16, 2007.

Liang, B, Austin, J 2005, 'A neural network for mining large volumes of time series data', *Conference on industrial technology,* pp. 688-693, 14-17 December 2005.

Mantzaris, D. H, Anastassopoulos, G. C, Lymberopoulos, D. K 2008, 'Medical disease prediction using Artificial Neural Networks', *8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008 Athens 8-10 Oct. 2008,* IEEE, pp. 1-6.

McCue, C 2006, 'Data mining and predictive analytics in public safety and security', *IT Professional*, vol. 8, no. 4, pp. 12-18.

Myers, M, Newman, M 2007 "The qualitative interview in IS research: examining the craft", *Information and Organization*, vol. 17, no. 1, pp. 2-26.

Nath, S V, 2006 'Crime pattern detection using data mining', *Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligent Agent Technology*, Oracle Corporation, Florida, Atlantic University, pp. 41-44.

Nonaka, I, Takeuchi, H 1995, *The knowledge-creating company : how Japanese companies create the dynamics of innovation / Ikujiro Nonaka and Hirotaka Takeuchi*, Oxford University Press, New York.

Rajagopalan, B, Isken, M.W. 2001, "Exploiting data preparation to enhance mining and knowledge discovery," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol.31, no.4, pp.460-467.

Schön, D. A 1983, *The reflective practitioner: How professionals think in action*, Basic Books, New York.

Shaon, A, Woolf, A 2008, 'An OASIS based approach to effective long-term digital metadata curation', *Computer and Information Science*, vol. 1, no. 2, pp. 2-12.

Sundhage, C, Lindgren, A 2010, 'Bakgrund till thesis-arbete', *Polismyndigheten i Västra Götaland.*

Tanasescu, A, Boussaid, O, Bentayeb, F 2005, 'Preparing complex data for warehousing' *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, ,vol., no., pp. 30-33.

Webster, J, Watson, R 2002, 'Analyzing the past to prepare for the future: Writing a literature review', *MIS Quarterly*, vol. 26, no. 2, pp. 13-23.

Williams, C 1983 'A brief introduction to artificial intelligence', *OCEANS'83 Proceedings*, vol., no., pp. 94-99.

Wolcott, H.T. 1994, *Transforming qualitative data: description, analysis, and interpretation*, Sage, Thousand Oaks, CA.

Younghoon, K, Kyoung-Gu, W, Hyoungmin, P, Kyuseok, S 2010, 'Efficient processing of substring match queries with inverted q-gram indexes' *International Conference on Data Engineering (ICDE)*, vol., no., pp. 721-732, 01-06 March 2010.