Mining Modus-operandi Patterns of Swedish Serial Burglaries

Simin Cai
Abstract

**Mining Modus-operandi Patterns of Swedish Serial Burglaries**

*Simin Cai*

Around 22,000 burglaries are reported to the Swedish police in 2012. It is not only inefficient to analyze these records by human experts, lots of valuable information remains hidden due to weakness of human information processing. Data mining is a promising technique to uncover hidden, unknown and potentially valuable information from large amount of data.

The goal of this project is to analyze burglary records and find crime patterns from a burglary dataset by using data mining and machine learning techniques. In this paper from the perspective of data mining I redefine the crime patterns by International Association of Crime Analysts. Then a series of correspondent algorithms and techniques are introduced to mine these patterns. A prototype is implemented to analyze the provided dataset. Crime patterns are identified and visualized in an understandable and user friendly fashion.
## Contents

1. Introduction .................................................................................................................. 1

2. Background ..................................................................................................................... 2

2.1 Data Mining .................................................................................................................. 2

2.2 Association rule mining ............................................................................................... 4

2.3 Cluster analysis ............................................................................................................. 4

3. Related Work .................................................................................................................. 6

3.1 Data mining and intelligent crime investigation systems .............................................. 6

3.2 Crime patterns mining ................................................................................................. 6

4. Research questions .......................................................................................................... 10

4.1 Research motivations .................................................................................................... 10

4.2 Research questions ....................................................................................................... 11

5. Research Methods .......................................................................................................... 12

5.1 Methods for RQ1 .......................................................................................................... 12

5.2 Methods for RQ2 .......................................................................................................... 14

6. Data Understanding and Preprocessing ......................................................................... 18

6.1 Data Understanding ...................................................................................................... 18

6.2 Data preprocessing ........................................................................................................ 18

6.3 Data consistency ............................................................................................................ 20

6.4 Data usage .................................................................................................................... 20

7. Analysis Models ............................................................................................................. 21

7.1 Analysis model for Spree ............................................................................................. 21

7.2 Analysis model for Hot Product ..................................................................................... 21

7.3 Analysis model for Hot Spot ......................................................................................... 22

7.4 Analysis model for Hot Place ....................................................................................... 22

7.5 Analysis model for correlating frequent patterns ......................................................... 23

7.6 Analysis model for Series ............................................................................................ 24

8. Prototype design ............................................................................................................. 26

8.1 Tools ............................................................................................................................. 26
1. Introduction

During the congressional hearings regarding the intelligence failure in the 9/11 attack at the World Trade Center in New York City, FBI director Robert S. Mueller indicated two primary problems of law enforcement agencies: they have been emphasizing too much in dealing with crimes after commitment rather than preventing the crimes beforehand; they have been making too much effort to collect data in investigations, while too little in analyzing the information they get [1].

With dramatic increase of data volume and lack of experienced analysts, analyzing the gathered evidence becomes a big challenge of the police [2]. In years 2003/2004, for example, 5.9 million crimes were reported in England and Wales [3]. Moreover, research and practice have demonstrated that human experts alone are not enough for crime investigation and prevention due to the nature of human being in information processing and decision making [4]. With these concerns many software systems and tools have been developed to aid investigation and analysis. By exploiting data mining, statistics and visualization techniques large amount of data can be analyzed by computers to find patterns and other useful information.

While “hard” evidences such as DNA or fingerprints are most important and desirable for burglary investigations, they are not always available due to many reasons. But some “soft” evidences are unavoidably left by burglars, such as Modus Operandi (MO) [5], which is defined as a summary of criminal’s habits, techniques and peculiarities of behavior [6]. Criminology experts believe that multiple crimes are likely to be conducted by the same criminal or group if a high similarity of MO’s is discovered among these crimes [7]. Together with spatial and temporal information MO can be used to find interesting crime patterns using data mining techniques.

Among other types of crimes burglary takes up a large proportion of all recorded crimes. However, the clearance rate for a volume crime such as burglary is very low (13% in England and Wales, year 2003/2004) [3]. In Sweden, about 22,000 burglary cases were reported in 2011, with a clearance rate of 3-4% [8].

This project aims to aid burglary investigation and prevention in Sweden by finding interesting crime patterns from annual national burglary data with relevant data mining and data visualization techniques. The MO information as well as spatial and temporal data are collected and analyzed.

In this project the burglary data are investigated and visualized. The similarity between burglary entries is designed, and based on this we are able to find potential criminal clusters. Using association mining techniques interesting crime patterns are discovered and future possible crimes can be predicted.
2. Background

This thesis is part of a project regarding providing uniform storage and intelligent processing for burglary analysis in Sweden. The project is cooperation between Blekinge Institute of Technology (BTH) and the Swedish police. Police officers in Sweden use a check-box form to collect information about each burglary in a structured way. With a national coverage the dataset will increase by 22,000 reports per year. At the moment researchers from BTH have done or are doing other parts of this project, including transferring data from check-box forms into a database, and developing a web-based query system with Geographic Information System (GIS) techniques.

2.1 Data Mining

Data mining is the process of finding useful patterns from data. Seen as a key step of Knowledge Discovery in Databases (KDD), which refers to the overall discovering of useful knowledge from data, data mining is the application of Artificial Intelligence, Machine Learning and statistical algorithms to extract patterns from data under acceptable computational efficiency limitations [9].

Data mining is an interdisciplinary field closely connected to Artificial Intelligence, Machine Learning, Statistics, as well as Database systems and data visualization. Data mining involves six common classes of tasks: Classification, Clustering, Dependency modeling, Regression, Summarization, Change and Deviation Detection [9]. A similar categorization is given by Tan, which categorizes data mining methods into Classification, Association Rule Mining, Clustering and Anomaly Detection [10]. Despite variance in categorization, a specific data mining task is either to predict unknown values given known values (Prediction), or to find understandable patterns that describe the data (Description). Tan and his coworkers give a rough illustration of predictive and descriptive data mining methods [10] (Figure 1).

![Figure 1. Predictive and descriptive data mining methods](image)
A number of algorithms have been applied in real business world and contributed to considerable business value and more algorithms are continuously introduced. Some of the most popular data mining algorithms used in industry are listed in Table 1.

Table 1. Categorization of data mining techniques and algorithms

<table>
<thead>
<tr>
<th>Category</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Decision Tree algorithms, Rule-based algorithms, Bayesian Network, Backpropagation, KNN, etc.</td>
</tr>
<tr>
<td>Association Rule Mining</td>
<td>Apriori, FP-Growth, etc.</td>
</tr>
<tr>
<td>Clustering</td>
<td>K-Means, DBSCAN, Self-Organizing Feature Map (SOM), etc.</td>
</tr>
</tbody>
</table>

The discovered patterns are evaluated with a series of criteria. They are desired to achieve some degree of certainty, novelty, potential usefulness and understandability

Certainty is usually considered an objective criterion regarding to how well the discovered pattern is valid on new data. Classification models are often judged by the classifying capability on test set, which can be measured with confusion matrix, accuracy or other quantitative measurements. For clustering models, external indices such as Entropy can be used to measure cluster validity when a test set with class labels is supplied; otherwise internal indices, such as Sum of Squared Error (SSE) can be defined to measure the goodness of a clustering without respect to external information.

While quantitative measurements can be defined for certainty criterion, the evaluation for other criteria is mostly subjective. Novelty requires the discovered patterns are non-trivial, and potential usefulness means the patterns are beneficial to the tasks at hand or in the future. Patterns are supposed to be understandable to analyzers after proper visualization and/or post-processing.

By combining these four criteria an interestingness measure can be defined for evaluating the overall value of patterns. Using this measurement, a pattern can be defined as knowledge discovered from data if it exceeds some interestingness threshold. Due to the subjectivity of interestingness, the desired knowledge and the evaluation of patterns are highly user oriented and domain specific, and is determined by whatever functions and thresholds the user chooses.

Data mining techniques have been exploited in a wide variety of business fields. In Fayyad’s paper he concluded the main business areas including marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents. Since then, data mining has been applied in many more areas involving large scale data analysis, such as law enforcement, bioinformatics, and text mining.
2.2 Association rule mining

An association rule is an implication expression of the form $X \rightarrow Y$, where $X$ and $Y$ are itemsets and $X \cap Y = \emptyset$. It represents a co-occurrence pattern of $X$ and $Y$ in a large dataset. Two metrics are used to evaluate the effectiveness of an association rule, support and confidence:

$$\text{support}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

$$\text{confidence}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)},$$

where $N$ is the number of transactions in the transaction set $T$ and $\sigma(X) = |\{t_i | X_t, t_i \in T\}|$.

Given a transaction set $T$, association rule mining is to find all association rules having a support greater than or equal to $\text{minsup}$ and a confidence greater than or equal to $\text{minconf}$, where $\text{minsup}$ and $\text{minconf}$ are user-defined thresholds for support and confidence respectively.

A common strategy to mine association rules is to divide the task into two steps. Firstly, find all itemsets whose support is greater than or equal to $\text{minsup}$. These itemsets are named Frequent Itemsets. For the rule $X \rightarrow Y$, the union of $X$ and $Y$ is a frequent itemset. Secondly, generate high confidence rules from each frequent itemset. In some cases we only care about the frequent patterns, which are frequent co-occurrences of different attributes. Those patterns are in fact frequent itemsets.

Many algorithms have been introduced to mine association rules though; Apriori is the best-known and most popular algorithm in use.

2.3 Cluster analysis

Cluster analysis is to find groups of objects such that objects in one group are similar (related) to one another and different from (unrelated to) objects in other groups. Sometimes cluster analysis is called “unsupervised classification” in that it tries to find hidden structures in unlabeled data. The clustering step in cluster analysis doesn’t require training data since it is performed only based on the interior information of the analyzed data. Before the clustering step, however, a training dataset with labeled classes is often needed in defining similarity/distance measurement and selecting significant features.

The goal of cluster analysis is to minimize the intra-cluster distance and maximize the inter-cluster distance. In other words, it is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. Simple Matching Coefficient and Jaccard Coefficient are suggested as suitable similarity measurements for binary vectors. Given two data objects $p$ and $q$, which have only binary attributes, a Simple Matching Coefficient is defined as
\[
S_{MC}(p, q) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}}
\]  
(3)

A Jaccard Coefficient is defined as

\[
J(p, q) = \frac{M_{11}}{M_{11} + M_{10} + M_{01} + M_{00}}
\]  
(4)

Here,

- \(M_{01}\) = the number of attributes where \(p\) was 0 and \(q\) was 1,
- \(M_{10}\) = the number of attributes where \(p\) was 1 and \(q\) was 0,
- \(M_{00}\) = the number of attributes where \(p\) was 0 and \(q\) was 0,
- \(M_{11}\) = the number of attributes where \(p\) was 1 and \(q\) was 1.

Many algorithms have been introduced in cluster analysis. Selecting a clustering algorithm is usually problem oriented and heavily dependent on the shape and size of the dataset. Due to simplicity and efficiency K-Means, DBSCAN and Self Organizing Maps are widely used methods in cluster analysis applications [10].

The validity of clusters can be evaluated by supervised evaluation, unsupervised evaluation or a combination of them [10]. In supervised evaluation, external class labels are provided and the validity is measured as how good the clusters match the classes. In unsupervised evaluation no external information is provided. The validity is measured by the coherence within each cluster and separation between different clusters. A number of indices are used as validity measurement, such as entropy (supervised) and Sum of Squared Error (SSE, unsupervised).
3. Related Work

3.1 Data mining and intelligent crime investigation systems

Data mining involves six categories of techniques: Classification, Clustering, Dependency modeling, Regression, Summarization, Change and Deviation Detection [6]. All these categories can be found widely used in research and tools for intelligent crime data analysis. Oatley’s work investigates how a variety of data mining techniques can be exploited in forensic data analysis [4]. Pie charts and histograms are effective tools to visualize crime patterns. Geographic Information Systems (GIS), hotspot analysis and spatial analysis can be used for geocoded crime data. Link analysis and social network analysis are useful to find criminal network. Clustering analysis and Self Organizing Maps can be applied to attempt of matching crimes. Statistics, decision tree classification and association rule mining and their application in crime mining are also briefly discussed in their work.

Intelligent crime investigation systems emerged in 1980’s and evolved from simple data storage and visualization tools to comprehensive expert systems [11]. Depending on the nature of data as well as on functional requirements, different data mining techniques have been exploited. COPLINK is a database application to find links between data from different police data sources. To realize this goal a mix of techniques are used, including entity extraction, clustering, classification, association rule mining, sequential pattern mining, deviation detection and link analysis [12]. Other intelligent police systems include FLINTS [13], RECAP [14], CrimeStat [15], etc. Nissan surveyed the representative forensic systems in this paper [16] that currently have been in use. He concedes the main breakthrough of data mining in forensics is fraud detection and a major direction is identifying the crime network.

3.2 Crime patterns mining

A variety of studies have been carried out to find crime patterns based on Modus Operandi (MO) information.

A comprehensive expert system was developed to predict re-victimization of a victim and its time frame based on Modus Operandi information using a method based on a Bayesian Belief Network and Expectation Maximization [17]. Given a set of offender profiles, the system is able to find the connection between certain unsolved crimes and the possible criminals.

Yokota and Watanabe have developed an RCPA (random choice probability algorithm) model [7], which uses statistic methods to compute a similarity of Modus Operandi data of burglars, and with this similarity score an unsolved case can be matched to possible suspects. Later, Ewart
and other researchers compared the RCPA model with an algorithm based on recency, prolificness and actual location information (RPAL) and a model that is a combination of the RCPA and RPAL (COMBIN) [5]. While RCPA computes an MO-based similarity, RPAL seeks for possible suspects by making use of temporal and spatial information, and COMBIN combines both MO and temporal-spatial information. Each of the models is tested with a dataset of 966 domestic burglaries committed by 306 criminals. Each crime in the dataset is treated as a target, and other crimes in the dataset are ranked according to the likelihood of being committed by the same criminal as the target crime. The crimes considered more similar to the target are given smaller rank number. Results show that COMBIN model outperforms the other two and nearly all (94 per cent) of the 966 crimes have a rank of 50 or less. The research suggests that combining MO information with temporal and geographic information gives best performance in matching burglaries.

In Jau-Hwang and Chien-Lung’s research link analysis and entropy are used to establish links among crime cases and chronic criminals based on Modus Operandi data [18]. The model is tested with two dataset of chronic robbery and residence burglary respectively. Results show that 32.6%/35.7% of the cases retrieved for a chronic robbery/residential-burglary criminal are indeed committed by the criminal and 54.3%/43.4% of the cases committed by a chronic robbery/residential-burglary criminal can be retrieved using the proposed association model.

Adderley conducts a series of case studies on finding crime patterns for burglaries and sexual assaults based on MO information [11]. Here, Multi Layer Perceptron (MLP) and Self Organizing Map (SOM) are the most important techniques used.

MLP is a supervised learning method which classifies unsolved cases into known suspects. With a dataset of 387 solved crimes, 87 of which were committed by the same criminal (“OffenderX”) and others randomly chosen, an MLP network was built to classify the crimes made by OffenderX from others. Validated on another criminal “OffenderY” who committed 42 crimes, the model achieved 96.5% of accuracy.

In comparison SOM is an unsupervised learning method that finds clusters in a dataset. Cases within each cluster share some kind of similarity and thus a serial crime by the same criminal may reside. 2370 sexual offences were studied and grouped into a number of clusters using a Self-Organizing Map [19]. The clusters were used to form profiles and were verified by police experts. Some of the profiles were considered to contain offences committed by a same person, or have leading information for the investigation.

With a similar goal Nath performed clustering analysis with K-Means rather than SOM [20]. Enhanced with a weighting schema of attributes K-Means is applied to find groups of crime records which are also geographically close to each other. The results are claimed to be validated by looking into court dispositions on these crime incidents, but no clustering measurement or more details are given due to confidentiality.
Instead of any specific clustering methods, Yu designed a similarity Segmented Multiple Metric Similarity Measure (SMMSM) to find the most similar incidents given the MO attributes of a crime \[21\]. In this measurement, attributes are divided into different groups, and between different groups exists a compensation relationship: suppose a, and b, are the i-th attributes of A and B respectively. If the difference between a, and b, is larger, the i-th attribute will contribute a smaller portion to the total similarity. The researcher uses SMMSM and Euclidian distance respectively to cluster a crime dataset and compares the number of crimes associated with the target crime in the cluster. Results show SMMSM outperforms Euclidian distance, although neither of these two measurements gives good performance. Four of the closest 100 cases found with SMMSM share the criminal with the target crime, while using Euclidian this number is one.

Chandra and other researchers implement a query interface to a crime database and provides trend plots of crime patterns \[22\]. Besides, seeing crime data as continuous data stream, a stream-based clustering technique is applied to find potential geographic clusters. In their case study, eight districts are indentified as hot spots.

While a wide range of data analysis techniques are exploited to analyze Modus Operandi information, some of them are observed more frequently than others in relevant literature. Neural networks and Bayesian methods are widely accepted to do classification and prediction, while K-Means and SOM are the first choices for clustering due to simplicity and efficiency.

One thing worth mentioning is that for classification tasks \[11\] \[17\] a considerable sized dataset with solved cases or suspects is provided to train the classifier and the known criminals are the targets. In evaluating clustering, most of the studies ask experts (police officers or court dispositions) to evaluate the helpfulness of the clustering results \[19\] \[20\] \[22\].

Spatial information has been taken into account by many police tools, ranging from simple mapping to advanced Geographic Information Systems (GIS) techniques. Some researches focus on mining spatial patterns for crime data. Eck and his coworkers elaborate the significance of hot spot analysis, various algorithms to find hot spots and different methods to visualize hot spot areas \[23\]. Three major crime mapping techniques are introduced: point mapping, spatial eclipses and thematic mapping. In a point map, which is the simplest display of crime distribution, each point represents a crime event. A spatial eclipse is a shaded area in the map which suggests equivalent risks of crime throughout the area with a dramatic reduction in risk at the border. In a thematic map an area is covered by a gradient, which implies that crime risk is high in the center and reduces gradually from that center.

Five different techniques are applied to do hotspot mapping for four types of volume crimes and evaluated in \[24\]. The evaluated techniques are point mapping, spatial eclipses, thematic mapping of administrative units, grid thematic mapping and thematic mapping with KDE (Kernel Density Estimation). Evaluation results show that thematic mapping with KDE consistently outperforms the others. In this method, individual crime events are aggregated within a user
specified search radius. A continuous surface is calculated which represents the density of crime events, and a smooth surface map is produced across this area.

A novel method combining associate rule mining and clustering is introduced in [25]. Essentially it discovers rules by studying the overlays of clusters on different layers, with either a Vertical-View approach or a Horizontal-View approach. An autonomous pattern detector is implemented to reveal cause-effects association rules among multiple geographical layers. Evaluation against a real crime database shows that the tool manages to discover association rules between crime types and places.
4. Research questions

4.1 Research motivations
The most important goal of this thesis is to identify crime patterns which can bring about valuable clues for burglary investigation and prevention. According to the definition given by the International Association of Crime Analysts (IACA), seven types of crime patterns are interesting to police. They are:

- **Series**: A group of similar crimes thought to be committed by the same individual or group of individuals acting in concert.
- **Spree**: A specific type of series characterized by high frequency of criminal activity within a remarkably short time frame, to the extent that the activity appears almost continuous.
- **Hot Prey**: A group of crimes committed by one or more individuals, involving victims who share similar physical characteristics and/or engage in similar behavior.
- **Hot Product**: A group of crimes committed by one or more individuals in which a unique type of property is targeted for theft.
- **Hot Spot**: A group of similar crimes committed by one or more individuals at locations within close proximity to one another.
- **Hot Place**: A group of similar crimes committed by one or more individuals at the same location.
- **Hot Setting**: A group of similar crimes committed by one or more individuals that are primarily related by type of place where crimes occurred.

Apart from the above, police also demands the flexibility to look for correlating frequent patterns among specific attributes, for instance, a correlating pattern combining the type of residence and the methods to break in.

According to which analytic techniques are involved, these crime patterns defined by IACA can be categorized into four classes (Table 2).

<table>
<thead>
<tr>
<th>IACA Crime Patterns</th>
<th>Pattern Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spree</td>
<td>Simple Patterns</td>
</tr>
<tr>
<td>Hot Product</td>
<td>Spatial Patterns</td>
</tr>
<tr>
<td>Hot Spot</td>
<td>Correlating Frequent Patterns</td>
</tr>
<tr>
<td>Hot Place</td>
<td>Similarity-based Patterns</td>
</tr>
<tr>
<td>Hot Prey</td>
<td></td>
</tr>
<tr>
<td>Hot Setting</td>
<td></td>
</tr>
<tr>
<td>Series</td>
<td></td>
</tr>
</tbody>
</table>
Spree and Hot Product care about the distribution or frequency of single attribute which can be identified with statistical methods and by proper means of visualization. Hot Spot and Hot Place are patterns in geographical proximity. Geo-Information System, spatial clustering and heat map are useful for this category. Hot Setting and Hot Prey look for patterns in co-occurrence of multiple attributes. In data mining these patterns are essentially frequent itemsets in association rules mining. Series is a group of crimes which are similar to each other in a way, where clustering analysis will be used.

4.2 Research questions
After a sound analysis of the above patterns, two types of information are found to be particularly interesting. One is the information to help the police investigate the trend of crimes and allocate resources correspondently. This includes, for example, the frequency of burglary cases over different periods of year, the geographical distribution of crimes, statistics of stolen belongings, profiles of victims, etc.

The other type of information is demanded mostly to provide clues for solving specific cases. Previous psychological and criminological studies have demonstrated a positive connection between Modus Operandi and the identity of criminal. Suppose police have succeeded to catch a burglar [7]. Suppose police have solved one burglary case and have all MO information recorded. Since the similarity of Modus Operandi indicates a similarity of personality, police can picture a criminal profile based on this burglar and reuse the experience in investigation of other cases. Vice versa, when a new burglary case is reported, it is compared with the cases already in the database, especially the ones which have already been solved.

Therefore, this paper will try to address the following two broad questions:

RQ1: How can we find the crime patterns which can help police allocate resources?

RQ2: How can we aid the investigation based on the similarity between crime records?

In practice, a solution to RQ1 will be able to find Spree, Hot Product, Hot Spot, Hot Place, Hot Prey and Hot Setting of the IACA crime patterns; a solution to RQ2 will be able to find the Series pattern (Table 3).

<table>
<thead>
<tr>
<th>IACA Crime Patterns</th>
<th>Pattern Types</th>
<th>Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spree</td>
<td>Simple Patterns</td>
<td>RQ1</td>
</tr>
<tr>
<td>Hot Product</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hot Spot</td>
<td>Spatial Patterns</td>
<td></td>
</tr>
<tr>
<td>Hot Place</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hot Prey</td>
<td>Correlating Frequent Patterns</td>
<td></td>
</tr>
<tr>
<td>Hot Setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Series</td>
<td>Similarity-based Patterns</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Research questions and correspondent crime patterns
5. Research Methods

5.1 Methods for RQ1

In this section techniques and methods used to find crime patterns for police resource allocation are discussed. Techniques and methods to discover simple patterns, spatial patterns and correlating frequent patterns are discussed and presented respectively.

Simple patterns

Simple patterns are patterns in distribution or frequency of single attribute. While statistical calculation is simple, proper visualization is the key to finding the patterns efficiently. Various charts and graphs can be used, for example, time series chart, pie chart and bar chart.

Spatial patterns

Crime statistics have shown that crimes often concentrate on several subareas rather than disperse evenly over the whole area. These concentrations indicate higher need of police resources, as well as more potential series crimes. Areas of crime concentration are often referred to as hot spots, which have a greater than average number of criminal or disorder events, or where people have a higher than average risk of victimization [23]. Research on crime mapping greatly focuses on finding hot spots and presenting them on maps in an understandable fashion.

As the simplest crime mapping technique, Point Map simply maps every crime on the map according to its geographic location. Based on the distribution hot spots can be identified by clustering techniques or heat maps.

A simple geographic clustering algorithm

Start: clusters[] = empty, bound of each cluster

For each point p, do

- Compute the distance between p and the center of each cluster. Find the closest cluster C.
  - If p is within the bound of C, add p into C and update the center of C.
  - Otherwise create a new cluster C’ only containing p. p is the center of C’. Add C’ into clusters[].

End.

This algorithm has two advantages over K-Means. First, the number of clusters is not required. When using K-Means we need to define a “K” before the algorithm starts, which means the number of expected clusters in the dataset. In our case, however, we don’t have such an expectation. Second, burglars are believed to have an average active radius, which is
represented by the bound of cluster in this algorithm. This algorithm will generate a local optimum of spatial clustering.

Another powerful and popular hot spots analysis technique is heat map. From a geographic perspective, a heat map is a graphical representation of a phenomenon where frequency of occurrence is represented by colors. Using methods such as interpolation or kernel density estimation, heat map technique aggregates discrete data points and provides a continuous and smoothing surface which represents the density of distribution across an area [23]. Locations with higher density of crime are marked with deeper color, in which way hot spots easily emerge. Figure 2 gives an example in which heat map is applied to detect hot spots of vehicle crime. Red areas, which lie in the west and south of the map, are hot spots and have more vehicle crimes.

![Figure 2. An example of heat map representing vehicle crime](image)

**Correlating frequent patterns**

Unlike simple patterns which are usually related to single attributes, a correlating frequent pattern is the frequent co-occurrence of multiple attributes. From the perspective of data mining, a correlating pattern is essentially a frequent itemset in association rule analysis. Besides the value of frequent itemsets in crime analysis, the rules discovered by association rule mining also often provide valuable information for investigation, in that they reveal a deeper relation between data items [4]. For a rule X → Y, the occurrence of X (antecedent) implies the occurrence of Y (consequent).

The most popular algorithms are Apriori [27] and FP-Growth [28]. Compared with Apriori, FP-Growth can improve the performance since it reduces the database IO. With a condensed structure (pattern tree) it only needs to scan the DB twice. In this project the whole subset of dataset is cached in main memory (See Section 8.2.3), and using FP Growth will not gain obvious advantage since the main overhead is not scanning dataset any more. Apriori is simpler and easier to implement.
The frequent itemset mining process of Apriori algorithm is:

Start: k=1

Generate frequent itemsets of length 1

Repeat until no new frequent itemsets are identified

- Generate length (k+1) candidate itemsets from length k frequent itemsets
- Prune candidate itemsets containing subsets of length k that are infrequent
- Count the support of each candidate by scanning the DB
- Eliminate candidates that are infrequent, leaving only those that are frequent

End

The number of frequent patterns is contributed by the minsupp. With a low minsupp and a large dataset, many patterns are hard to visualize. In this project I introduce a novel prefix-tree based interactive representation for frequent patterns. The details will be discussed in 8.3.

For each itemset I with minsupp:

Let Li denotes a subset of I which contains i elements.

Find all minconf rules with a single consequent of the form I - L1 \(\rightarrow\) L1.

Repeat until no new rules are generated:

- Guess candidate consequents Ck by appending items from I - Lk-1 to Lk-1
- Verify confidence of each rule I - Ck \(\rightarrow\) Ck using known itemset support values

End.

### 5.2 Methods for RQ2

In this section, techniques and methods used for similarity search and cluster analysis are discussed.

**Similarity measurement**

The dataset used in this work is mixed with binary, temporal and geographical attributes. Theoretically an overall similarity measurement is designed to consider the influence of different categories of attributes:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} w_k s_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$  \hspace{1cm} (5)
Here $p$ and $q$ are two records. Given $n$ categories of attributes, $s_k$ is the similarity of the $k$th attribute, and $w_k$ is the weight of the $k$th attribute. If the $k$th attribute is an asymmetric attribute and the values of this attribute in both $p$ and $q$ are 0, or either of them is a missing value, then $\delta_k = 0$. Otherwise $\delta_k = 1$.

The weight of each attribute in (5) represents the significance of this attribute to finding a similar crime case. Ideally, an expert in burglary analysis can provide this information as a priori and assign the weight of each attributes. Alternatively, some machine learning methods, for example a Backpropagation neural network [29], can be applied to extract weighting knowledge from the dataset itself, if a reasonably big dataset is provided to validate the assigned weights.

Unfortunately neither of these approaches works in this project. The police have no idea about the significance of different attributes – they think all of them are important though. In the provided dataset, we only got about 20 crime cases with suggested suspects, which is not enough to infer which attributes give more information when we struggle to find similarity pattern between crime cases done by the same suspect.

In the provided dataset, all information except the exact dates and coordinates of burglary cases has already been presented as binary attributes (Section 6). In this project only binary attributes will be considered in the similarity comparison. Jaccard coefficient is selected as the similarity measurement.

**Clustering**

The initial impulse of cluster analysis lies in the attempt to find crimes done by the same suspects based on the similarity between crimes. More generally, each cluster may represent a common style of burglary. This information can be valuable in criminal profiling and investigation.

The standard K-Means [30] algorithm is very simple and widely used:

Start:

Select K points as initial centroids.

Repeat until centroids don’t change:

- Form K clusters by assigning all points to the closest centroid.
- Recompute centroid of each cluster.

End

As Jaccard coefficient is chosen as the similarity measurement, the distance function in K-Means is defined as:

$$D(p, q) = 1 - J(p, q)$$  \hspace{1cm} (6)
Here $p$ and $q$ are two records. $J(p, q)$ is the Jaccard coefficient which is defined as Equation (4).

Most of crime data analysis researches and tools choose K-Means as their clustering method due to the simplicity and speed of the algorithm. However, there are several disadvantages which limit the performance of clustering and usability of analysis tools.

First of all, K-Means is very sensitive to the selection of initial centroids. The convergence of the algorithm is highly dependent on the initial position in the searching space, which means that a local optimum rather than a global optimum is guaranteed. Thus a poor selection of initial centeroids will result in a bad clustering. In most cases a random selection is performed for the initial centroids, which leads to an unpredictable performance of clustering analysis.

In order to avoid bad initial centroids, an initialization method K-Means++ \[31\] can be exploited to get better performance. The intuition of K-Means++ is to spread out the initial centroids so that they are as far away from each other as possible.

Begin

Initialization:
- Choose one center uniformly at random from among the data points.
- Repeat until $K$ centers have been chosen.
- For each data point $x$, compute $D(x)$, the distance between $x$ and the nearest center that has already been chosen.
- Choose one new data point at random as a new center, using a weighted probability distribution where a point $x$ is chosen with probability proportional to $D(x)^2$.

Continue with standard k-means clustering.

End.

This seeding method has been proved to improve the final error of K-Means considerably. Although some extra time is needed for the initialization, K-Means++ finishes clustering faster and finds a better clustering compared to the standard K-Means \[31\].

Another drawback is the parameter $K$. The final clustering will contains $K$ clusters, which assumes the user should know or expect the number of clusters. A single cluster could be divided into two less optimal clusters only because of a large $K$, and on the other hand a small $K$ may result in an unwanted merging of two clusters. It’s also very inconvenient from the perspective of usability since tool users are asked to decide the number of clusters, which is actually one of their expectation of the tool.

An effective solution to this problem is to try out with different $K$ and find the one bringing the “best clustering”. Since the provided burglary dataset only have a few solved cases with
suspects, it is pointless to use supervised evaluation. On the other hand, evaluation of the
goodness of a clustering can be based on its structure, which means points in the same cluster
should be as close as possible while points in different clusters should be as separate as
possible. In this thesis Sum of Squared Error (SSE) is used to estimate the number of clusters.
Given a clustering of \( K \) clusters, SSE measures the overall error as follows:

\[
SSE = \sum_{i=1}^{k} \sum_{x \in C_i} ||m_i, x||
\]  

(7)

where \( x \) is a data point in cluster \( C_i \) and \( m_i \) is the centroid of \( C_i \). The error for each data point
is the distance between the point and the centre of the cluster it belongs to. One common
method to find the optimal \( K \) is to compare SSE for a number of clustering solutions with
different \( K \)'s. The “elbow” point in the SSE curve indicates the optimal \( K \), which means, before
this \( K \) the SSE curve drops dramatically while after this \( K \) the curve keeps stable or only declines
more slowly.

So the overall method for clustering using K-Means++ is described as follows:

Begin

Repeat from \( K=2 \) until \( K=\text{Max} \),

- Select initial centroids with K-Means++.
- Find \( K \) clusters using standard K-Means.
- Calculate the SSE.

Plot SSE against different \( K \). Find the \( K_{opt} \), which is the elbow point in the curve.

Find \( K_{opt} \) clusters using standard K-Means.

End
6. Data Understanding and Preprocessing

6.1 Data Understanding
A dataset, which includes 1065 burglary cases reported in the southern part of Sweden and Stockholm from June of 2011 till December of 2012, has been provided by police. Among these cases, 22 of them are provided with corresponding suspects, whose identities are anonymized and replaced by a unique identifier.

Table 4 summarizes all attributes of burglary cases. All of these attributes are considered somehow useful for the burglary investigation. Temporal attributes include the date when the crime happened, as well as season (only “winter” or “summer”) and time (only “day”, “evening” or “night”). Coordinates are provided for each case. Besides suspect’s identity, temporal and spatial attributes of each reported case, other collected attributes include residential area, type of residence, surroundings, activities of victims before crime, stolen possessions, methods of burglary, traces left by burglars, etc.

Among the attributes in Table 4, “pHash” is a string attribute which indicates the identity of the suspect of each case. All but the 22 cases with suspects have a NULL value in this attribute. “Date”, “City” and “Post code” attributes are strings representing the date, city and post code of a reported burglary case respectively. “Longitude” and “Latitude” are numeric coordinates of the case. All other attributes, which are marked with * in Table 4, are provided with value of either “YES” or “NO”.

For privacy protection reasons the latitude/longitude data are provided with an accuracy of 110 meters. Neither the names, nor any other personally identifiable information of suspects are included in the system or dataset. Instead each suspect is referred to by a unique but anonymized identifier. In fact, the accurate coordinates and more detailed information about suspects can be useful in finding interesting patterns. In this thesis, the design of analysis models and prototype tool only target the provided dataset.

6.2 Data preprocessing
The provided dataset is stored in an Excel sheet. All data are extracted from the Excel sheet and imported into a MySQL database, in that it is easier to maintain the data with a database management system as the burglary dataset increases in the future. In addition, the prototype tool, which is implemented in Java, can interact with the MySQL database through standardized JDBC interface.

In Table 4 the attributes with * only have value of “YES” or “NO”, so they are encoded as binary data. There is no necessity to do extra normalization for these attributes. “Longitude” and “Latitude” are processed as double numeric values. “Date” is processed as Timestamp. “pHash”, “City” and “Postal Code” are string data and used as they are.
<table>
<thead>
<tr>
<th>Category</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suspects</td>
<td>pHash</td>
</tr>
</tbody>
</table>
| Time                            | Season (summer/winter) *  
Date  
Day_or_night (day/evening/deep night) *  
Weekday_or_weekend (weekday/weekend) * |
| Location                        | Longitude **  
Latitude **  
City  
Postal code |
| Residential Area *              | enskilt/en_granne/flera_intilliggande_grannar/hörntomt_eller_tomt/tomt_ett_skogsparti_eller_allmänning |
| Residence Specification *       | villa/gård/kedjehus_eller_parhus_eller_radhus/lägenhet_bostadsrätt/lägenhet_hyresrätt  
mer_än_ett_plan/enplan/källare/lägenhet_bottenplan/lägenhet_längst_upp |
| Alarm *                         | larm_utlöst/larm_ej_utlöst/larm_saboterat/larm_inget_larm |
| Surroundings *                  | posten_tömd,tänt_inomhus,belysning_utomhus, belysning_på_gatan  
fordon_på_uppfart, gräsklipning_eller_snöskottning,  
hund_hemma_eller_skylt_tecken_på_hund |
| Victim *                        | Where_is_victim(målsägande_hemma_under_brottet/målsägande_borta_max_2_timmar/målsägande_borta_2_24_timmar/målsägande_borta_24_timmar)  
Entrepreneurs (egen_företagare, förekommer_bolagreg, ej_egenföretagare)  
Recurrence (målsägande2_återkommande/målsägande2_spontan)  
Planned Absence (målsägande3_planerat/målsägande3_oplanerat)  
Activity_beforeCrime(hushållsnära_tjänst/hembesök/uppringad_av_okänt_nummer_eller_person/dokumenterat_bortavaro_på_internet/hemmavarande_barn/annoserat_köp_eller_sälj/fordon_på_flygplats_eller_gräns |
| Entrance View *                 | ingång_objektet_insynsskyddat/ingång_objektet_je_insynsskyddat |
| Entry Point *                   | altan/balkong/entre_eller_framsida/källare/övervåning |
| Entry Method *                  | dörr/fönster bryt  
/glasskross/utgång_förberedd/samman_in_eller_utgång/verktyg_från_platsen_demonterat |
| Ransacking*                     | genomsök_genomsökt/genomsök_begränsat_genomsökt |
| Stolen Products *               | alkohol_eller_tobak/elektronik/guld_eller_smycken/kontanter/klädner/läkemedel/leksaker/vapen/kassa_eller_värde_eller_brandskåp/parfymer/nycklar_fordon/annat/pass_id/inget_av_oanstående_gods |
| Trace *                         | fingeravtryck/dna/handskar/skor/verktyg/färgavtäckning/childs_för_spårsökning/inget_av_oanstående_spår |

* Attributes to be encoded as binary values.

** Precision removed to the degree of 110 meters.
6.3 Data consistency
The police make sure that all records are consistent by a Javascript application along with the report form. For example, the “Season” attribute, which has optional values of “summer” and “winter”, cannot be both or neither of them. A few records in the provided dataset have missing values in “Date”, “City” or “Postal code”. The missing “Date” attribute will be assigned 1970-01-01, while other missing values are assigned “UNKNOWN”.

6.4 Data usage
All data attributes are used in the analysis. The provided dataset is used to evaluate Spree, Hot Produce, Hot Spot, Hot Place, Hot Prey and Hot Setting patterns discovered by the implemented tool. A subset of the provided dataset, which only includes 22 burglary cases with suspects, is used to evaluate Series patterns discovered by the implemented tool.
7. Analysis Models
In this section, models for the analysis of different crime patterns are. Analysis models for Spree, Hot Product, Hot Spot, Hot Place and Series patterns are described individually. Hot Prey and Hot Settings will be described under the analysis model for correlating frequent patterns, since the only difference in their analysis models is the selected set of attributes.

7.1 Analysis model for Spree
In the analysis of Spree patterns, as shown in Figure 3, user first selects the records to be analyzed by specifying a time period and a geographic area. These records are then aggregated by their “Date” value in three different levels: date, month and season. The aggregated records in month or season level go into groups by different months or seasons. Finally the records are displayed in time series charts.

7.2 Analysis model for Hot Product
In the analysis of Hot Product patterns, as shown in Figure 4, user first selects the records to be analyzed by specifying a time period and a geographic area. These records are then aggregated by the values of their stolen products attributes and then form groups. Each group stands for one type of products and includes all records associated with this type. Finally the groups are displayed in a pie chart.
7.3 Analysis model for Hot Spot
In the analysis of Hot spot patterns, as shown in Figure 5, a heatmap is produced using coordinates of all selected records. Spatial clusters are generated using the spatial clustering algorithm in section 5.1. In the end, all individual cases, the produced heatmap and all clusters are displayed on a map.

Figure 5. Analysis model for Hot Spot patterns

7.4 Analysis model for Hot Place
Figure 6 shows the analysis model for Hot Place patterns. For each record, the number of user specified places within the surrounding area of this record is calculated. An index map between record and places is created, with records as index keys and places as values. Then based on this map, a reversed index map is created, with places as index keys and records as values. The places, around which burglary frequently happens, can be found by sorting this reversed index map.
7.5 Analysis model for correlating frequent patterns

The analysis of correlating frequent pattern applies the Apriori algorithm to do the association rule mining, as discussed in Section 5.1. Features are selected by users depending on the goal of the mining task. In order to find Hot Setting patterns, which is a relation among features of the crime spots, features of residential area, residence type and surroundings should be selected. For Hot Prey patterns, on the other hand, features of victims and victims’ activities before crime are of most interest. Figure 7 shows the model for analysis of correlating frequent patterns.
7.6 Analysis model for Series
Two analysis tasks are involved in the analysis of Series patterns. One is to find records which are similar to a provided crime case. The other is to find clusters, which may indicate crimes committed by the same criminal, among a selected dataset.

Similarity search model

Figure 8 shows the model to do similarity search. Users are allowed to set time and geographic constraints: only the cases which happened within a specified time window and within a specified geographical area around the target case are considered “similar”. Users are also allowed to select specific features to compute the similarity. As discussed in Section 5.2, Jaccard coefficient is selected as the similarity measurement.

Cluster analysis

As shown in Figure 9, users are allowed to select features for cluster analysis. As discussed in Section 5.2, a method using K-Means++ is applied to do cluster analysis. K-Means++ are first run a number of times with different k. The corresponding SSE of each run is computed and recorded, and then an SSE curve is plotted. The k at the elbow point on the SSE curve is considered as the optimal k, which is used to do the final clustering using K-Means++.
Figure 9. Analysis model for cluster analysis.
8. Prototype design

8.1 Tools
In this section the tools used in the implementation of the prototype system are introduced.

WEKA

WEKA (Waikato Environment for Knowledge Analysis) \(^{32}\) is a free and open source Java-based platform for machine learning and data mining tasks. As an independent analysis tool WEKA integrates a number of machine learning and data mining algorithms, including data preprocessing, classification, clustering, association rule mining and visualization. Moreover, as an open source platform it doesn’t only offer various APIs for analysis, but allows developers to implement their own data mining algorithms and integrate them into WEKA.

The prototype takes classes “Instances”, “Instance” and “Attribute” from WEKA as basic data structures for data sets, data tuples and attributes respectively. Filters in WEKA are used to perform data processing.

Google Maps API

Google Maps API \(^{33}\) is a set of Javascript library provided to developers by Google. By using these libraries, Google maps with a variety of functionalities can be embedded into websites and applications.

This prototype uses Google Maps API to load a map and display crimes on the map. With the help of HeatmapLayer in visualization library a crime heat map is created from crime data set. By using the places library provided by Google Maps API, developers are able to get the location information of places which are supported by Google Maps API.

Java Swing

Swing \(^{34}\) is the primary Java GUI widget toolkit and it is easy to integrate other Java based visualizing libraries. The main interface of this prototype is implemented by Swing, together with Prefuse and JFreeCharts.

Prefuse

Prefuse \(^{35}\) is a Java based framework for data modeling and interactive visualization. Developers can easily build a rich, interactive visualization for structural or non-structural data by calling or extending APIs offered by Prefuse. In this prototype we use Prefuse to create a prefix-tree based interactive representation for frequent patterns.

JFreeChart
JFreeChart is a free Java chart library for developers to display data with high quality charts. In this prototype we use JFreeChart to create bar charts, pie charts and time series charts.

### 8.2 System architecture

The prototype system is designed as following (Figure 10).

![System Architecture Diagram](image)

#### 8.2.1 Query Interface

A graphical query interface allows users to interact with the system and view the analysis plots and reports. In the query panel user can set the time and location constraints, such as “From 2012-06-01 to 2012-09-01”, “Stockholm” and “Center: (Latitude, Longitude), Radius: 50km”. Then a query batch starts. Data are fetched from database and user can choose to mine different patterns.
8.2.2 Database
All data are stored in a MySQL database. For this project with the provided burglary dataset, the
data model only contains one table, which represents a burglary case and includes all attributes
in Table 4 as columns, because they are only attributes of the burglary case entity. In the future,
the data model needs to be extended in order to reflect more entities and relations between
them, if more data are involved.

8.2.3 Logic layer
The logic layer includes some common utilities and two major analysis modules: Temporal,
spatial and statistical analysis module and Similarity search module.

Common utilities
Common utilities mainly deal with import of data from a database and transform them to the
internal structures used by Weka.

After the database connection is established, the subset meeting the searching conditions is
fetched from database into a WEKA structure: Instances, which is composed of instances, which
refer to real records in the database. This structure resides in main memory until this analysis
batch is finished and destroyed.

For each specific type of pattern, a subset of the Instances is generated depending on which
attributes are needed for this type of pattern. This is with the help of Filters from WEKA. A filter
performs conditional selection on the Instances. Based on provided conditions a filter picks up
or removes instances and turns on or off attributes. For example, with a filter we can retrieve a
subset of Instances only containing stolen possessions and falling into the period between 2012-
6-1 and 2012-12-1. This is fast since all these happen in memory. This process is illustrated in
Figure 11.

A limitation of this “cache in memory” is that, a big result set may impair the performance since
too much memory is occupied for caching. With this provided dataset we don’t see such a
problem. But if more data are provided for analysis, this caching mechanism needs to be
improved or redesigned for better performance.
Temporal, spatial and statistical analysis module

This module is an implementation of analysis models in Section 7.1 to Section 7.5, which are models to discover Spree, Hot Product, Hot Spot, Hot Place and correlating frequent patterns respectively.

Statistical calculation is used to compute Spree and Hot Product patterns. For Hot Spot patterns, HeatmapLayer library from Google Maps API is used to compute the heat map. On the heatmap, an area with higher crime rate is colored, while a lower crime rate area is colored in green. The color gradually changes from red to green as crime rate degrades. An area with no crime is not colored at all. The spatial clustering algorithm in Section 5.1 is implemented to identify geographical clusters. The places library from Google Maps API provides interface to get coordinates of special types of places in an area, such as police stations, restaurants, schools, etc. The identification of Hot Place pattern uses this library to get specified places around burglary cases, and then the index map and reversed index map as described in Section 7.4 are populated. The Apriori implementation in Weka framework is used to do association rule mining. The identified frequent itemsets are used to generate a prefix tree for the representation described in Section 8.3.

Clustering and Similarity search module

This module is an implementation of analysis models in Section 7.5, which are models to find records similar to a given case and to find clusters in a dataset respectively. The knowledge of
which attributes are important for the similarity measurement is not provided, but the prototype allows the user select which attributes should be used for similarity search and cluster analysis.

Cluster analysis is divided into two steps. The first step is to decide the number of potential clusters in the data set. The prototype decides the optimal number by looking for the elbow points in the SSE curve. Then the tool performs the clustering with this optimal number. The SimpleKMeans class in Weka framework is extended to run K-Means++ clustering algorithm iteratively with different k values and calculating SSE for each iteration.

8.2.4 Visualization Layer
The visualization layer is composed of a number of views representing different types of crime patterns. JFreeChart time series charts, pie charts and bar charts are used to visualize Spree and Hot Product. Google Maps are used to visualize Hot Spot and Hot Place. To visualize correlating frequent patterns, a prefix-tree based interactive presentation is designed and introduced in Section 8.3.

8.3 A prefix-tree based interactive representation for frequent patterns
The number of frequent itemsets can be very big given a small minimum support. Usually an frequent itemset containing N attributes is represented as a tuple of N elements like \(<X_1, X_2, \ldots, X_n>\). When the number of itemsets is getting big, a better visualization is needed for users to find the itemsets which they are interested in.

A prefix-tree based representation is designed to visualize frequent itemsets in this system. The support of itemset is represented by depth of color. Interactive features are added to the representation in order to improve the understandability of the representation.

Given a set of frequent itemsets, for example:

\(<A>, <B>, <C>, \n<A, B>, <A, C>, <B, C>, \n<A, B, C>\)

They are arranged as in Figure 12. Each node and its precedents in its branch together represent a frequent itemset. So the first branch in Figure 12 represent \(<A>, <A, B> and <A, B, C>\).
In the system the nodes are colored according to their supports. A node with higher support gets deeper color. When user clicks on one node, the support of this frequent itemset will be displayed. Users are also allowed to zoom in and zoom out of the tree, and search for a specific attribute as illustrated in Figure 13.
8.4 Drill-down
The prototype allows user to perform further analysis on current results. When a user selects a
spatial cluster on the Hot Spot map, the records in this cluster are displayed in a list. Then the
user can do further analysis on these records, such as to identify the Spree or Hot Product
pattern, to do association mining, or to perform a cluster analysis. Similarly, if a user clicks at a
time point on the Spree time series graph, or a type of product on a Hot Product chart, the
records in this group can be further analyzed. If the user first does a cluster analysis, he can also
select a cluster and analyzing the spatial and time patterns on the records in this cluster. Figure
14 shows an example that, after getting Hot Product patterns user can continue to do cluster
analysis on all burglary records associated with gold and jewelry.

Figure 14. Further cluster analysis on the results of Hot Product


9. Evaluation

In this section, the prototype system is evaluated with the burglary dataset, which is provided by Swedish police and introduced in Section 6. The whole dataset with 1065 burglary records is used to evaluate against Research Question 1, which is to find the crime patterns which can help police allocate resources and includes identification of Spree, Hot Product, Hot Spot, Hot Place and correlating frequent patterns. A subset of the dataset with 22 solved cases is used to evaluate against Research Question 2, which is to help with the investigation based on the similarity between crime records. This includes similarity search and cluster analysis.

9.1 Evaluation of RQ1

The prototype is used to identify Spree, Hot Product, Hot Spot, Hot Place and correlating frequent patterns from the whole dataset with 1065 records, which are reported in south Sweden and Stockholm area. In real usage, the analysis may only need to focus on a specific period or area. This tool provides a query interface, through which users are able to fetch the records by specifying time and geographic conditions.

9.1.1 Evaluation of Spree patterns

In the Spree view, time series charts display the curve of the frequency of burglaries with respect to each day. By dragging the sliding bar user can easily see the trend of crime as time went by and find out in which period burglaries happened intensively (Figure 15(a)). From the graph we can see a large number of burglaries happened in the week before and after 2011-11-04. The cases which happened in this period are displayed as user drags the sliding bar.

The burglary records are also aggregated by Months and Seasons. As shown in Figure 15(b), the November of 2011 saw most burglaries, while burglaries were also reported massively in the October of 2011, the July of 2012 and the September of 2012. Figure 15(c) shows that many burglaries occurred in winter of 2011 and summer of 2012, which is a pattern in an even coarser granularity. The figures in Figure 15(a) and 15(b) also shows the distribution of the time (“day”, “evening” or “night”) when a burglary actually happened.
Figure 15. (a) Crime cases aggregated by day. (b) Crime cases aggregated by month. (c) Crime cases aggregated by season.
9.1.2 Evaluation of Hot Product patterns

Hot Product view displays the distribution aggregated by type of stolen belongings in a pie chart. Figure 16 shows that gold and jewelry, electronic products and cash are burglars’ favorites. The burglary cases in a group are displayed when the user clicks on that group in the chart.

Figure 16. Hot Product analysis shows that gold and jewelry, electronic products and cash are burglars’ favorites in south Sweden and Stockholm area from June of 2011 till December of 2012.

9.1.3 Evaluation of Hot Spot patterns

A Google map with marks of all reported cases is displayed on the tool’s interface. A heatmap is generated to cover all areas with burglaries. The color of areas degrades smoothly from red to green as burglary rate degrades. An area without any crimes is not colored. In order to give better illustration we zoom in the map and focus on Stockholm area, as shown in Figure 17. The area in the black rectangle is colored in red, which indicates that area (Spånga) has highest crime rate.

Spatial clusters are also displayed on the map, with the number of burglary cases marked on the cluster. The biggest cluster has 71 reported crimes, which is in the same area as the heat map suggests. Thus Spånga is a Hot Spot of burglaries in Stockholm area. Similarly, Hot Spots in other areas can be identified.
9.1.4 Evaluation of Hot Place patterns

The types of places supported by Google place library is listed on the interface. The user selects the type of places from the list, and the system ranks all these places by the frequency of crime in the surrounding area. In Figure 18, the user wants to rank the bank branches in the area centered <N59.3642, E17.8754> and within a radius of 20km according to the frequency of burglary. Result shows a branch of Nordea bank in Skårbygränd is the hottest place among all bank branches in this area, with 79 burglaries reported in its neighborhood of 3km.
Figure 18. A branch of Nordea bank in Skårbrygård has most burglaries in its neighborhood in the area centered <N59.3642, E17.8754> and within a radius of 20km.

9.1.5 Evaluation of Correlating frequent patterns

Hot Prey and Hot Setting patterns are essentially correlating patterns on victims and crime spots respectively. The prototype system extends the mining of correlating patterns to combination of any arbitrary attributes.

The identification of Hot Setting is shown in Figure 19. RESIDENCE_TYPE, RESIDENCE_AREA and RESIDENCE/store are selected as attribute sets. The min supp is 0.2 and the min conf is 0.4. The upper panel in Figure 19 displays frequent patterns and the lower panel lists association rules. The Hot Setting patterns represented by the prefix tree in Figure 19 are listed in Table 5.

Figure 19. Hot Setting pattern analysis, with min supp 0.2 and min conf 0.4.
Table 5. Hot Setting patterns represented by the prefix tree in Figure 19 and their supports

<table>
<thead>
<tr>
<th>Hot Setting patterns</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;typavbostadsomräde_flera_intilliggande_grannar&gt;</td>
<td>0.72</td>
</tr>
<tr>
<td>&lt;typavbostad_villa&gt;</td>
<td>0.64</td>
</tr>
<tr>
<td>&lt;typavbostad2_mer än ett_plan&gt;</td>
<td>0.55</td>
</tr>
<tr>
<td>&lt;typavbostad2_enplan&gt;</td>
<td>0.24</td>
</tr>
<tr>
<td>&lt; typavbostad_villa, typavbostadsomräde_flera_intilliggande_grannar&gt;</td>
<td>0.44</td>
</tr>
<tr>
<td>&lt; typavbostad_villa, typavbostadsomräde_flera_intilliggande_grannar, typavbostad2_mer än ett_plan&gt;</td>
<td>0.39</td>
</tr>
<tr>
<td>&lt;typavbostadsomräde_flera_intilliggande_grannar, typavbostad2_mer än ett_plan&gt;</td>
<td>0.37</td>
</tr>
<tr>
<td>&lt;typavbostad_villa, typavbostadsomräde_flera_intilliggande_grannar, typavbostad2_mer än ett_plan&gt;</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Association rules revealed from mining can also be significant to investigation. In Figure 20, STOLEN_PRODUCTS, RESIDENCE_TYPE and TRACE are chosen as attribute sets. In addition to frequent patterns such as <guld_eller_smycken, typavbostad_villa, verktyg>, the tool also finds an association rule “guld_eller_smycken=1, typavbostad_villa=1 => verktyg=1, conf(0.75)”, which indicates 75% of the burglars who stole gold or jewelry from a villa left their tools at the spot. This rule suggests that in those cases which happen in villa and gold and jewelry are reported stolen, police should pay attention to the tools left by burglars when they are collecting evidence. Some interesting rules are listed in Table 6 as examples.

Figure 20. Correlating frequent patterns mining also finds an association rule “guld_eller_smycken=1, typavbostad_villa=1 => verktyg=1, conf(0.75)”
Table 6. Association rules identified by combining different attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Rule</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTRANCE_BREAKIN_METHODS, TRACE</td>
<td>fönster=1 =&gt; skor=1</td>
<td>0.41</td>
</tr>
<tr>
<td>RESIDENCE_TYPE, SEASON</td>
<td>typbostads_lägenhet_hyresätt=1 =&gt; årstid_sommar=1</td>
<td>0.81</td>
</tr>
<tr>
<td>RESIDENCE_TYPE, SEASON</td>
<td>typbostads_villa=1 =&gt; årstid_sommar=1</td>
<td>0.65</td>
</tr>
<tr>
<td>RESIDENCE_TYPE, PRODUCTS</td>
<td>typbostads_villa=1 =&gt; guld_eller_smycken=1</td>
<td>0.5</td>
</tr>
<tr>
<td>RESIDENCE_TYPE, RESIDENCE_STORE, ENTRANCE_BREAKIN_METHODS</td>
<td>typbostads_villa=1, typbostad2_enplan=1 =&gt; föster=1</td>
<td>0.63</td>
</tr>
</tbody>
</table>

9.1.6 Results of evaluation

The identified crime patterns provide valuable information which help police allocate resource and thus contribute to burglary prevention. Series patterns, for instance, reveal in which period of the past years burglaries occurred intensively. Hot Product patterns reflect that which products are more popular among burglars. As shown in spatial patterns, some areas have higher rates of burglary. With this information warning of burglary can be sent to citizens before this period, and more policemen can be appointed to those areas. Correlating frequent patterns describe the correlating relations among specific attributes, for example, the correlation between time and residence type. Some of the discovered identified association rules provide indications which may contribute investigation. The tool allows user to select arbitrary attributes for the association analysis, thus providing flexibility to probe interesting patterns based on user’s experience.

9.2 Evaluation of RQ2

The 22 cases with suspects in the provided dataset are used to evaluate against Research Question 2, which is to help with the investigation based on the similarity between crime records. The 22 cases are associated 17 different suspects. The distribution of cases associated with suspects is shown in Table 7. Among them, SuspectA committed 4 crimes, SuspectB and SuspectC committed 2 crimes respectively. The other 14 cases are associated with 14 different suspects. Each case has a Record ID as an identifier. SuspectA is associated with record 1197, 1209, 1876 and 1891. SuspectB is associated with record 1864 and 1865. SuspectC is associated with record 1272 and 1828. Ideally, similarity search will find the cases associated with the suspect of the target case, and the clusters generated by cluster analysis will include cases committed by the same suspect.
Table 7. Suspects and their associated case records in evaluation dataset

<table>
<thead>
<tr>
<th>Suspect</th>
<th>Record ID’s of cases associated to the suspect</th>
<th>Number of cases connected to the suspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuspectA</td>
<td>1197, 1209, 1876, 1891</td>
<td>4</td>
</tr>
<tr>
<td>SuspectB</td>
<td>1864, 1865</td>
<td>2</td>
</tr>
<tr>
<td>SuspectC</td>
<td>1272, 1828</td>
<td>2</td>
</tr>
<tr>
<td>SuspectD</td>
<td>1138</td>
<td>1</td>
</tr>
<tr>
<td>SuspectE</td>
<td>1215</td>
<td>1</td>
</tr>
<tr>
<td>SuspectF</td>
<td>1217</td>
<td>1</td>
</tr>
<tr>
<td>SuspectG</td>
<td>1251</td>
<td>1</td>
</tr>
<tr>
<td>SuspectH</td>
<td>1293</td>
<td>1</td>
</tr>
<tr>
<td>SuspectI</td>
<td>1313</td>
<td>1</td>
</tr>
<tr>
<td>SuspectJ</td>
<td>1768</td>
<td>1</td>
</tr>
<tr>
<td>SuspectK</td>
<td>1769</td>
<td>1</td>
</tr>
<tr>
<td>SuspectL</td>
<td>1879</td>
<td>1</td>
</tr>
<tr>
<td>SuspectM</td>
<td>2127</td>
<td>1</td>
</tr>
<tr>
<td>SuspectN</td>
<td>2144</td>
<td>1</td>
</tr>
<tr>
<td>SuspectO</td>
<td>2146</td>
<td>1</td>
</tr>
<tr>
<td>SuspectP</td>
<td>2151</td>
<td>1</td>
</tr>
<tr>
<td>SuspectQ</td>
<td>2153</td>
<td>1</td>
</tr>
</tbody>
</table>

9.2.1 Evaluation of Similarity search

Given a target record, Similarity Search view finds the user-specified number of most similar records in database and display them on the map. The user can select which attributes are used for the similarity search.

SuspectA is associated with record 1197, 1209, 1876 and 1891. During the evaluation, each record associated with SuspectA is used as target record, and the tool is used to find 5 and 10 most similar records respectively. All attributes are selected for the similarity computation. Then we calculate how many of the other three cases appear in these 5 and 10 records respectively. The results are shown in Table 8 and Table 9 respectively. The results show that the tool provides useful information to identify the cases which are associated with SuspectA, given a target case which is done by SuspectA.

Table 8. Five most similar records found by similarity search

<table>
<thead>
<tr>
<th>Target Record</th>
<th>Records appearing among 5 most similar records</th>
<th>Match rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1197</td>
<td>1891</td>
<td>33%</td>
</tr>
<tr>
<td>1209</td>
<td>1876, 1891</td>
<td>67%</td>
</tr>
<tr>
<td>1876</td>
<td>1891, 1209</td>
<td>67%</td>
</tr>
<tr>
<td>1891</td>
<td>1876</td>
<td>33%</td>
</tr>
<tr>
<td>Average match rate</td>
<td></td>
<td>50%</td>
</tr>
</tbody>
</table>
Table 9. Ten most similar records found by similarity search

<table>
<thead>
<tr>
<th>Target Record</th>
<th>Records appearing among 5 most similar records</th>
<th>Match rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1197</td>
<td>1876, 1891</td>
<td>67%</td>
</tr>
<tr>
<td>1209</td>
<td>1876, 1891</td>
<td>67%</td>
</tr>
<tr>
<td>1876</td>
<td>1891, 1209, 1197</td>
<td>100%</td>
</tr>
<tr>
<td>1891</td>
<td>1876, 1197</td>
<td>67%</td>
</tr>
<tr>
<td>Average match rate</td>
<td></td>
<td>75.3%</td>
</tr>
</tbody>
</table>

9.2.2 Evaluation of Cluster analysis

We use the prototype system to perform cluster analysis on the evaluation dataset and all attributes are selected for the analysis. Ideally the system would find 17 clusters, each of which matches a suspect and contains all crimes and only crimes committed by him.

The prototype system eventually finds 18 clusters within the data set. The cluster results are shown in Table 10. Each row in the table represents the cluster ID and the case records assigned to this cluster. The values in column “A” to “Q” represent the number of cases that belong to SuspectA to SuspectQ in this cluster, respectively.

Since the suspects are provided as class labels, to evaluate the cluster analysis results, we choose entropy and recall to measure the correspondence between the cluster labels and the class labels [10].

Entropy is defined as:

$$entropy = \sum_{i=1}^{K} \frac{m_i}{m} \left( - \sum_{j=1}^{L} \frac{m_{ij}}{m_i} \log_2 \frac{m_{ij}}{m_i} \right)$$  \hspace{1cm} (8)

Here, $K$ is the number of clusters, $L$ is the number of classes, $m$ is the total number of data points, $m_i$ is the number of points in cluster $i$, and $m_{ij}$ is the number of points of class $j$ in cluster $i$. A low entropy indicates the clusters tend to contain data points of single classes, thus meaning better performance.

Using the previous terminology for entropy, the recall of cluster $i$ with respect to cluster $j$ is defined as:

$$recall(i,j) = \frac{m_{ij}}{m_i}$$  \hspace{1cm} (9)

A high recall indicates that the cases of a single class tend to be clustered into the same group.

The entropy of the overall clustering, as well as of each cluster, are calculated and presented in Table 10. The overall entropy is 0.307.
Table 10. Results of cluster analysis on the evaluation dataset

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cases</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1864</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1865</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2151</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1828</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1138</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1209, 1879, 2153</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.585</td>
</tr>
<tr>
<td>7</td>
<td>1215</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1217</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1197, 1272</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1293</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1313</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1768</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1769</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2127</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2144</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2146</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1876, 1891</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall   | 0.307

The recall measurements of each cluster are shown in Table 11.

Incorrect clusters in the clustering are evaluated. It occurs in Table 10 that records associated to different suspects are grouped into the same cluster (Cluster 6 and 9), and records associated to the same suspects are separated (SuspectA, SuspectB and SuspectC). These miss-grouped cases in the same cluster, however, show more similarity to each other, than other cases associated to the same suspects. The user can select the records in one cluster and do further analysis on these records using the tool’s “drill-down” function. More information can be found that, for example, these cases happen during the same period, be associated with the same type of residence or target the same kinds of victims.
### 9.2.3 Results of evaluation

During the evaluation of similarity search, which finds records similar to a given target record, 50% of the records associated to the same suspect of the target record are listed in the 5 most similar cases, and 75.3% of the records associated to the same suspect of the target record are listed in the 10 most similar cases. The evaluation results demonstrate that when an unknown record is used as an input of similarity search, it is promising to find the records which are done by the same criminal.

The evaluation of cluster analysis on the evaluation dataset shows that the clustering accomplished by the prototype reaches low entropy (0.307), and high recall for most of the clusters. This indicates that, for the provided dataset, the cluster analysis succeeds in finding a good mapping between the suspects and their associated records.

Due to the size of this evaluation dataset, it is too early to draw a conclusion whether or not the prototype is able to find burglary cases committed by the same suspect. However, possibilities...
of using clustering analysis techniques to match suspects with cases are demonstrated, and important information can be gained from the analysis results.
10. Conclusion and future work

10.1 Conclusion
This project investigates methods of identifying, analyzing and visualizing crime patterns from burglary databases, and supporting crime investigations based on the similarity between burglary records. The definition of crime patterns from International Association of Crime Analysts (IACA) is adopted as domain requirements.

The crime patterns from IACA are categorized into four types and data analysis and mining methods are introduced for each type. A prototype system is designed and implemented to analyze a burglary dataset provided by Swedish police.

Hot Product and Spree are simple patterns based on the frequency or distribution of single attribute. Using statistical analysis and proper graphs, these patterns are easy to be identified by analysts. Spatial pattern analysis identifies patterns in geographic distribution of crime cases. With the help of heat map technique, spatial clustering and Google Maps API, Hot Spot and Hot Place are identified and presented to analysts and can support decisions of allocating police resources and preventing crime. Association rules mining methods are applied to mine correlating frequent patterns such as Hot Prey and Hot Settings. A prefix-tree based representation is implemented to provide understandability and efficiency for visualization of the correlating patterns. A dataset including 1065 records is used to evaluate the prototype. The evaluation shows that the prototype can serve as a useful tool to identify the simple statistical patterns, spatial patterns and correlating frequent patterns.

Series patterns are identified using similarity search and cluster analysis techniques. Similarity search is applied to find crimes which share the same suspects with a given case. Cluster analysis is applied to find clusters from the burglary dataset, each of which ought to be mapped to one suspect. Evaluations on both similarity search and cluster analysis show good performance. However, since the evaluation dataset is too small (22 records), we are not able to draw the conclusion that the prototype succeeds in finding the Series patterns from a more general burglary dataset. Nevertheless, as a pilot study, the evaluation results indicate that similarity based search and cluster analysis are promising techniques for mining Series patterns, and further research and development can be based on this prototype.

10.2 Future work
In the follow-up study, a crime expert should be involved to provide information on attribute selection and the importance of each attribute to identification of burglars. Based on this information a similarity measurement can be designed and applied to the similarity search and cluster analysis. A larger burglary case dataset with identified criminals should be provided for evaluation. The similarity measurement and analysis model are evaluated against the evaluation dataset, and they can be adjusted to achieve better performance.
A better understanding of the needs of crime analysts is necessary, and new features can be implemented continuously based on this prototype.

Scalability issues should be taken into account in the selection of mining methods and development of the tool. More attributes may be included into the burglary dataset and the tool’s data model needs to be extended accordingly. In the current design there exists only one table including all MO attributes, because all these attributes are tightly related to the burglary case. If more information about the victim is included, for example, more personal information about the victim, a new table may need to be added to represent victims. The number of records in the burglary dataset is also increasing as new records are added continuously. The current prototype keeps the selected records in main memory using the data structures from WEKA, which is a data mining framework widely used in many large scale data mining tasks [32]. The architecture may need to be modified to cope with the growing data.

In terms of usability of the system, more interactive elements can be introduced to simplify the analysis process. To make the mining results more understandable, future work can enhance the representation of correlating frequent patterns and association rules. Experts in Human Computer Interaction can be involved in the design of interface. Crime analysts can be invited to use the system and give feedback on the usability.
References

[28] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD Record, 2000, pp. 1-12.


