An Interactive Human Centered Data Science Approach Towards Crime Analysis

Nadeem Qazi*, B.L.William Wong

Middlesex University,London UK

Abstract

The key challenge during crime analysis is to identify plausible linkages in unstructured crime text for the hypothesis formulation. Crime analysts painstakingly perform directed, time-consuming searches of many different structured and unstructured databases to collate these associations without any proper visualization.

To tackle these challenges and aiming towards facilitating the crime analysis, in this paper, we examine unstructured crime reports through text mining to extract plausible associations. Specifically, we present associative questioning based searching model to elicit multi-level associations among crime entities. We coupled this model with partition clustering to develop an interactive, human-assisted knowledge discovery scheme. Our proposed KDD scheme handles the issue of categorical data in clustering through the bag-of-words approach and measures cluster quality utilizing silhouette analysis. It is able to extract plausible associations identifying crime pattern, clusters of similar crimes, co-offender network and suspect list based on spatial-temporal and behavioral similarity. We quantified these similarities through calculating Cosine, Jacquard, and Euclidean distances. Additionally, each suspect is also ranked by a similarity score in the plausible suspect list.

The proposed KDD also inspect grand challenge of integrating effective human interaction with machine learning algorithm. It offers intuitive visualization allowing the analyst to feed his domain knowledge including choosing of similarity functions for identifying associations, assigning weight to each crime pattern component for suspect ranking towards unsolved crime and dynamic feature selection for interactive clustering of similar crimes. A two-dimensional re-configurable crime cluster space along with bipartite knowledge graph is used for visualizing associations.

We demonstrate the proposed scheme through a case study using the Anonymized burglary dataset.The proposed scheme is found to facilitate human reasoning and analytic discourse for intelligence analysis

*Corresponding author.

Email addresses: n.qazi@mdx.ac.uk (Nadeem Qazi), w.wong@mdx.ac.uk (B.L.William Wong)

Keywords: `Interactive Clustering`, Linkage Analysis, Crime matching, Text mining, Human-centered searching, Network Visualization, Knowledge Graph, Data Science

1. Introduction

The intuition of human experts plays a vital role towards solving complicated problems. Researchers have emphasized human role towards the setting of particular issue and dataset for the extraction of relationships in structured and unstructured content[1]. This integration of computer and human, where computer assists the human design process and human being is the in charge of an algorithmic process is termed as interactive data mining in information retrieval literature. It offers three benefits over black box algorithms of machine learning i.e. understanding (why one ML is different than others), diagnosis (reasons for the failure of a ML process) and refinement (factors affecting the performance of a ML such as changing feature vector)[2]. However data mining/machine learning application are mostly automatic without or limited human intervention and hence increases the risk of modeling artefacts. The grand challenge is to integrate effective human interaction with powerful machine intelligence through visual analytics to support both human insight and decision making [3]. In another research [4] have emphasized that data mining techniques should keep the domain expert intelligence into loop.

This grand challenge is handled in many data data-intensive science applications including health informatics [5],social media [6],[7],image processing [8], web page analysis [9] etc. We in this research has focused towards exploring the solution of this challenge in the crime analysis domain,that mostly deals with unstructured text content. We report our approach of utilizing interactive data mining in uncovering criminal associations to assist the crime investigation in general and crime matching in particular.

1.1. Problem statement

Researchers [10] have defined crime matching as the process of "assigning crimes or criminals to [previously] solved or unsolved crime incidents", while [11] describe crime matching as the ability to link or connect crimes in ways that enable the identification of potential suspects. Despite subtle differences, both refer to the use of machine learning based algorithms to (i) find similarities among crimes to discover potential suspects; and (ii) to develop offender profiles in a way that can be used to find matches with the profiles of offenders in unsolved crimes. The information-intensive querying process of crime matching requires establishing multi-level associations among crime entities to discover and reconstruct crimes through analysis of the evidence left at the crime scene.

More often during crime matching process analysts spend a large amount of time reading crime reports looking clues for criminal associations among criminal entities such as criminals, vehicles, weapons, bank accounts, and organizations. They ask a variety of questions based on associative questioning [12] to learn more about the diverse nature of the context in which the crimes were committed for making sense of the situation that would help solve the crime.

Some analysts apply the 5WH (Who, What, When, Where, Why and How) structured analytic model [13] to discover who else might have been involved in the crime, what other factors or events could be relevant and how was the crime and other similar crimes committed? They seek information that could lead them to make associations with other concepts to create plausible hypotheses that can lead to solving a criminal case. In addition to this, they analyze the commonalities between criminal cases and compare solved crimes with an unsolved crime to generate a new hypothesis during a criminal investigation. It requires them to group the solved crimes on the basis of similar characteristics and can examine the changes in grouping caused by different attributes of a crime. Currently, the investigator has to painstakingly perform directed, time-consuming searches of many different databases to collate, such a comprehensive picture of the crime.

The keyword and semantic base searches do not leverage the power of associations of concepts in the search domain, as the former does not consider the meaning of the given query and later though looks for the meaning of the question, however, lacks to understand any association in the data. Additionally, the resultant search data due to lack of proper visualization causes analysts to face a number of significant difficulties including making sense of collated data, distinguishing the relevance or similarities among the cases, identifying and understanding associations between criminal entities. This indeed suffers the efficiency of the crime matching process, therefore it is valuable important and challenging to work towards a human centered searching model with efficient visualization for grouping similar cases, finding associations between them to facilitate hypothesis formulation.

1.2. Solution

Therefore identifying the need for linkage based search mechanism, in our earlier related research, we have introduced associative search[14] and demonstrated its use through formal concept analysis for criminal investigation. We defined associative search as the cognitive thinking process consisting of associative questioning based on crime triangle and the routine activity theory[15]. It searches along the networks of associations between objects such as people, places, organizations, products, events, services, and so forth. In this work, utilizing this concept, we proposed an interactive, human-centered knowledge discovery scheme inspired from [16] to extract criminal associations from the unstructured text of crime reports using temporal, spatial and behavior characteristics of the committed crime.

In our proposed scheme, we employed vector space model and partition clustering to group the unlabeled text of the crime data using dynamic features selection and validate the crime clusters quality through silhouette analysis. Later we visualize these multi-level associations using graph theory. Our framework enables analysts to interact directly with machine learning models to integrate domain knowledge into the analysis process. It is provided through setting required number of clusters, dynamic features selection, choosing associative questions for criminal network creation and setting parameters during preprocessing of the text mining process.

The contributions of this paper include 1) a detailed literature review showing data science contributions towards crime analysis, 2) an association discovery scheme incorporating a proposed multi-level associations model for identifying criminal linkages, 3) interactive clustering distinguishing the relevance or similarities among the cases, 4) our approach to handle categorical data in the clustering and lastly visualization of multidimensional associations of crime entities through knowledge graph identifying criminal groups.

The rest of paper is organized as follows. Section 2 presents related research. We unfold the proposed knowledge discovery scheme for crime matching in multiple sections, describing association miner, interactive clustering and visualization in sections 3,4 and 5 respectively. A case study using the Anonymized data is also presented in section 6 and conclusion is drawn towards the end of the paper.

2. Related Work

Our proposed knowledge discovery scheme for criminal analysis integrates multiple data mining techniques under a single framework. Following this,we have performed a detailed literature review towards every component of our proposed KDD, including text mining,association extraction, clustering and identifying criminal network from unstructured text specifically focusing on crime domain, and present it with examples in the next section.

2.1. Interactive Knowledge Discovery Scheme

Knowledge discovery is an interactive and iterative process starting from acquiring domain knowledge, followed by selecting, preprocessing and cleaning the target dataset. The other stages of Knowledge discovery process consist of dimensional reduction and projection of the data, implementation of appropriate data mining algorithm for the required task and finally interpreting the mined pattern and extracting knowledge from it [16]. Several researchers have proposed different theoretical variations of KDD models such as interaction model [17],[18], and sense-making models [17] in order to recognize and integrate human role with analytic process. More recently an interactive visualization prototype[19] demonstrated “analyst is in the loop” approach to extract crime signature from modus operandi description for event detection.

Text mining is an important part of the KDD to extract information from the unstructured content. Researchers [20] have developed a text mining framework “TexRap” to handle scientific challenges of using unstructured text data from online media. They used Random Forests (RF), Support Vector Machines (SVM) and Multilayer neural network to identify entities of interest and classify sentiment polarity and intensity. In crime based text mining research, researchers [21] have used SVM to classify news articles of Sri Lankan English newspaper as crimes or no crimes articles. In another similar research, Crime Profiling System[22] using Arabic text, extracts crime-related information from given crime document. It employs N-gram model to extract crime type, location, and nationality of a person involved in an event from the Arabic text and utilizes Self Organizing Map (SOM) to cluster crime texts. The work presented in this paper, however, integrates N-gram model of text mining with effective human interaction and elicits tempo, spatial and behavioral associations for crime matching through associative questioning.

2.2. Associations Extraction for Crime analysis

Association analysis utilizes data mining methods to extract the relationships/affinity patterns/rules among various items objects or events.In recent years several implications for association analysis has been demonstrated, primarily focusing on support-confidence theory, in various domains including market analysis for extracting consumer purchase patterns with buying products [23], social media mining for identifying topics in tweets [24], recommendation systems [25],[26] and health-care [27]. [27] performed association analysis on electronic medical records of diabetes patients and proposed a new assessment metric to identify rare items/patterns without over-generating association rules. Researchers [28] have developed a hierarchical matrix-based visualization technique employing Apriori algorithm for mining association rules in categorical datasets.

However in recent past not very much work is reported towards extracting crime linkages from textual data. The more recent example includes linking serial crimes [29] through behavioral information. They developed a decision support system consisting of similarity algorithms, a classification model, a feature selection and parameter learning algorithm to link the serial crimes. However, their approach does not deal with textual data. Researchers [30] developed an algorithm based on the modus operandi similarity of crime pattern using Jacquard coefficient to link burglaries to a serial offender. The result shows that crime series with the same offender on average had a higher behavioral similarity than random crime series. [31] utilized association analysis concept including color, brand, and type of vehicles to detect suspect that are potentially involved in criminal activity. They integrate journey path analysis techniques together with the association rule mining to analyze such criminal behavior.

The growing trend for association analysis in crime domain is towards the use of Nave Bayes algorithm.[32] utilized Bayesian networks to link evidence in crimes. In another research [33] has utilized Bayesian networks for modeling multiple offenders for two separate offenses. Another example is from Reference [34] who employed crime date, location, and the criminal name and criminals acquaintances as clues to predict the posterior probability of a criminal to be associated towards an unsolved crime. Earlier, researchers however also have demonstrated the use of other machine learning algorithms such as logistic regression [35][36], probability inference [37] using behavioral features of the crime pattern to elicit associations between crime and criminals. Another example is PrepSearch [38] an integrated system to detect the rank list of suspects for a given crime scene. It combines geographic profiling with social network analysis for crime patterns detection.

In this work, we have utilized the concept of associative search from our previous work [14], incorporating associative questioning through a 5-WH model for the elicitation of spatial, temporal,and behavioral associations of criminals from crime reports.

2.3. Clustering similar crimes

Clustering allows similar objects to be organized into groups. In recent years it has applied in a wide range of applications including in sentiment classification [39], electricity load management [40], active learning [41] ,tourism industry [42] etc. However, most of the clustering techniques in text mining are based on vector space model or its different variant.

In a digital forensic domain, [43] have introduced subject-based semantic document clustering algorithm based on vector space model to groups documents on a suspect's computer into a set of overlapping clusters, each corresponding to one unique subject. Reference [10] has proposed a framework for crime matching combining crimes classification and clustering through multilayer neural networks and K-Mean algorithm respectively. [44] proposed a Bayesian model, utilizing crime locations and offenders modus operandi for burglary crime series identifications. In another approach [45], minimum cut based graph clustering is demonstrated to detect residential burglaries series. They used a feature vector consisting of modus operandi, residential characteristics, stolen goods, spatial similarity, to group similar crimes. Reference [46] construct crime cluster zones of Indian crime dataset using K-means method, however, they used numerical data. Researchers [47] detected crime patterns in news articles through K-mean clustering over multiple crime types. They employed affinity propagation algorithm for determination of the number of clusters.

Most of the work cited above have employed fixed features in clustering algorithms. However, textual data involves high dimensional features, that due to curse of dimensionality not only leads to high computational cost but also affects the algorithm performance. Feature selection methods reported as solution are mostly automatic including filter[48], wrapper[49], and hybrid [50].

In recent years interactive clustering emerges as a potential solution towards this problem. The data visualization community has produced a number of interactive approaches where users are integrated into the analysis process. For example, iVisClustering tool[51] perform interactive document clustering through topic modeling, allowing users to guide the process. I-TWEC [6] is an interactive web-based clustering tool for twitter data that utilized suffix tree based algorithm to cluster user uploaded tweets using their semantic. Some other examples includes Cluster Sculptor [52],INFUSE [53],radial axes method for visual backward feature selection [54] etc .These tools facilitate analyst to steer the feature selection process according to their domain knowledge and specification.

We in this work present an interactive clustering through dynamic feature selection for grouping solved and unsolved crimes along with their associated offenders based on tempo spatial and modus operandi similarity. In our work, both numbers of clusters and feature vector are not fixed and are chosen by users to dynamically cluster the crimes. We use multidimensional scaling technique to visualize the hidden relationship between crime KPIs.

2.4. Criminal network Analysis

Another important characteristic of our framework is the extraction of a criminal network. One of the important features in constructing and analyzing network is the detection of the groups or communities. Community detection falls in two categories [55], the first method relies on the structure of the network graph and mainly involves some variant of divisive [56] and agglomerative [57] algorithms. The second category, however, employs similarity matching between each couple of nodes to extract communities. A variant of agglomerative algorithm ,called as TRIBASE [55] is demonstrated for the extraction of communities from twitter data,LOGANALYSIS [58], however utilized divisive algorithm [56] to detect communities. It employs force-directed node-link layout to constructs criminal networks from phone call records. Other examples include HICODE [59] that uses disjoint and overlapping community detection algorithms for finding the hidden communities.

In crime based data mining, researchers [60] employed name entity recognition along with a modified Apriori algorithm to extract prominent criminal community, from unstructured textual data of chat log. Their method uses frequency of the interaction between two people to measure the strength of linkages. Another example,VISFAN [61],a visual analytics framework, incorporated enhanced graph drawing techniques with hierarchical clustering to visualize financial activity networks. It extracts entities like bank accounts, addresses, amount and types of the transactions, motivations from financial reports and visualizes it in form of a network. Reference [62] has used bipartite model for extracting hidden ties in both traditional and cyber crimes over pharmaceutical crime and underground forum data set respectively. Some earlier examples related to our work are commercial tools like COPLINK Explorer [63], Dynalink [64], JIGSAW [65], However, either most of these tools lack proper visualization or do not have the ability to extract criminal relationship from textual data.

it is also evident from the above literature that, in crime analysis, there is a need for a unified framework that should offer crimes clustering, criminal association extraction and community elicitation with proper visualization under a single envelope.

3. Data Mining Framework for Crime matching

Aiming towards facilitating analyst for hypothesis creation,and following the need for a unified framework for crime analysis, we now present a human-centered discovery pipeline for crime analysis in general and crime matching in

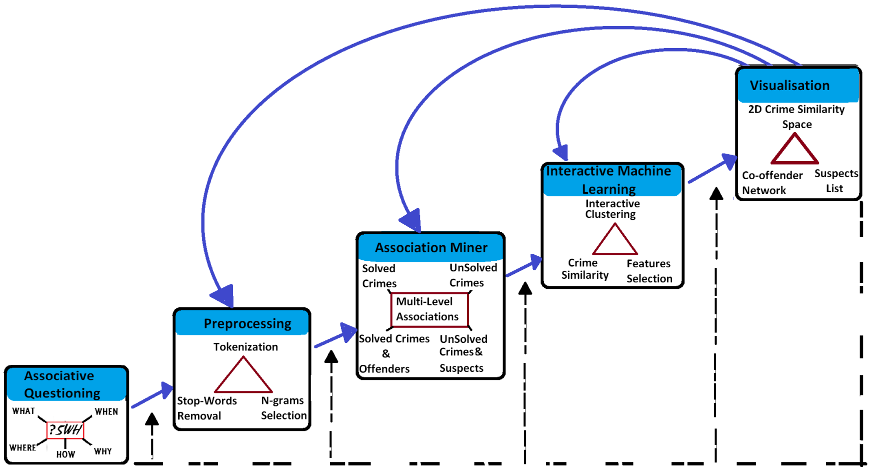


Figure 1: Knowledge Discovery Pipeline for Crime Analysis.

particular. The architecture of proposed pipeline, shown in Figure 1 is inspired by the general process of KDD [16]. It takes a crime pattern as input, elicits the multi-level associations on the basis of temporal, spatial and behavioral characteristic, and group crime similarities in a 2D interactive clustering space. The associations in each cluster are then hierarchically visualized through a bipartite tree based structure called as knowledge graphs in our framework, to depict the co-offender network and plausible suspect list using graph theory. We now describe each component of the pipeline in sections below.

3.1. Associative Search Engine

We employs an associative query engine that during retrieval and integration phase generates the spatial, temporal and modus operandi based associative queries presented in Table 1.0, for a user-specified crime pattern extracting data from the knowledge base for association extraction. The generated data after pre-processing is then fed into association miner which elicit the multi-dimensional associations to group criminal as shown in the Figure 1 and is described below.

3.2. Association Miner

The association miner unit Figure 1 of our KDD scheme elicits multi-level associations to unfold similar crimes, criminal community and plausible suspect list from a given crime dataset. Following the methodology used by previous researchers [66] and [37], we compared modus operandi similarity, geographical and temporal proximity to establish and visualized these associations. The proposed model through rule-based heuristic and similarity matching extracts

Table 1: Associative Questioning

Where else crimes like this have been committed?
Who else in past have committed the crimes like this?
What are other modi operandi that has been used in committing crimes like this?
Who are the known offenders operating in an area and what is their modus operandi to commit crimes
What are the additional details of the associated offenders/victims his past history etc?
What are the geo-spatial profiles of the offenders, including its temporal, spatial and other similar criminal activities resembling with the given crime pattern?
How many times the offender has committed the similar crimes and what are its temporal and spatial details?
What is his/her pattern of modus operandi?
Where an offender mostly likes committing an offense and who else has committed the same crime at this location?
What are the other offenses that have occurred with a similar given crime pattern?
How often offenses like the given crime pattern have occurred?

associations in three levels. The Level 1 and Level 2 as shown in Figure 2 reveal the relationship between solved and unsolved crimes and the associated offenders or victims respectively. The employed heuristic rule to distinguish crimes is that a solved crime is the one which has been solved and a perpetrator/offender has been identified/sentenced for a crime, and unsolved crime is the one, for which the goal is to identify potential/probable offender responsible for committing this crime. The feedback loop as shown in Figure 1, from the visualization module to association miner, allows analyst to steer the associations by feeding his/her domain knowledge including putting up questions, choosing similarity functions and setting weight for each component of the crime pattern.

Focusing towards the human-centered approach and based on the given input question, level 3 of our model elicit associations between offenders and solved crime, between suspects and unsolved crimes through measuring the similarity of modus operandi, time and location of crime occurred with that of given input crime entities i.e. offender or unsolved crime. These associations are represented as an undirected heterogeneous graph [67], consisting of multiple nodes of crime entities connected with spatial-temporal and behavioral associations.

The root node of this graph is the user given question which could be a criminal name or unsolved crimes and based on this, the constituent or children nodes may be the perpetrator, location, offense, time and modus operandi, which are created dynamically based on the given input question. The edges in the network are made of the associations connecting nodes on the basis of similarities, highlighting a subset of nodes having a similar characteristic. Thus for a given

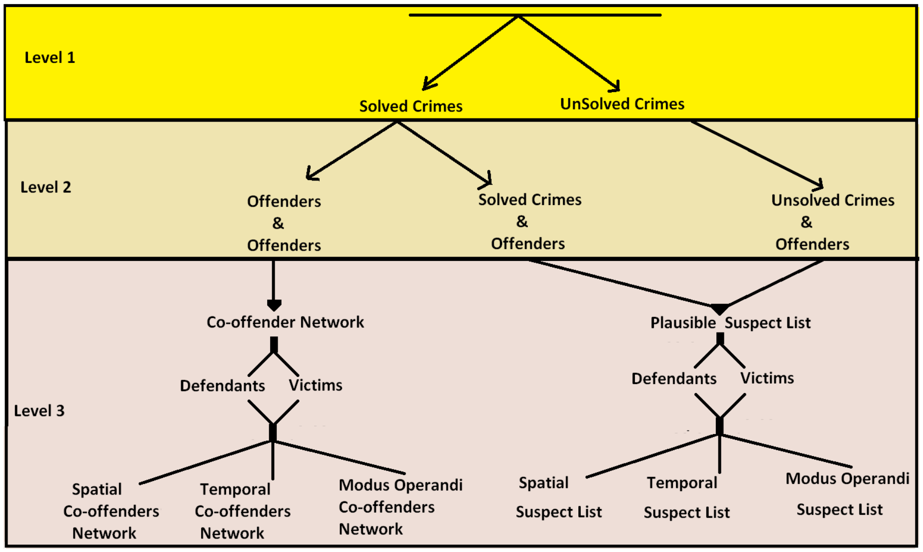


Figure 2: Spatial,Temporal, and Modus Operandi based Multi-level Associations model.

input node of a criminal name, it compares the crime pattern similarities of the root node with that of all the offenders/victims in other crime reports and generates a graph for the co-offender network. On the other hand, if the input node is an unsolved crime, then the crime pattern comparison extract associations between offenders of the similar solved crimes with the given unsolved crime, generating a list of possible suspects. We now describe these graphs separately in the following sections.

3.2.1. The Co-Offender Network

We modeled two types of co-offender network. The first model is based on crime associations i.e. when two or more offenders are reported together in a crime report for committing a crime. Our second network model is based on the spatial, temporal and modus operandi similarity and we have named it as (STM) model in our framework. These similarities are quantified through calculating Cosine, Jacquard or Euclidean distances. However, following the co-reasoning approach between human and machine, choice of distances is dynamically set through user input. This thus implicitly steer the model each time an analyst chooses different distance function to create similarity based associations.

We adopted the idea of k-Nearest Neighbor query to set the maximum number of the offenders in a network graph. It is implemented by setting the desired number of the retrieved offender in the graph and is defined by the user. Following this, we retrieved all the offenders that matched with the given similarity, rank them in descending order according to the value of chosen distance function. However, instead of selecting all, we selected only K number of offenders.

This thus enables to weave the graph of crime entities having shortest distance and hence largest similarity among them based on the chosen distance function.

Mathematically Let A_t represents type of association where $t \subseteq (C, L, T, MO)$ and P, C, L, T, MO represent the set of distinct values of the offenders, crimes, locations, time of the event, and modus operandi respectively. We define a criminal co-offender network model having nodes of offenders $P_s \subseteq P$ connected with the chosen association type $A_t \subseteq (C, L, T, MO))$. Any two offenders in the network P_n and P_m are said to be connected with each other if they share same association type so that following is true.

$$A_k = \langle P_n, P_m \rangle$$

where $k \subseteq (C, L, T, MO)$

Alternatively if two offenders P_n and P_m have a spatial association, for example they have committed a crime together at same location and P_n also has a spatial association with another offender P_j, for example they have committed a crime together at some other place, then we elicit an indirect association between P_m and P_j as:

$$A_l = \langle P_n, P_m \rangle$$

$$A_l = \langle P_n, P_j \rangle$$

$$A_l = \langle P_m, P_j \rangle$$

P_m and P_j are thus supposed to be indirectly connected with each other in our proposed association model. An algorithm is written to create an adjacency matrix to show total number of common associations between two offenders. The first row and first column of this matrix contains the offender P_1 to P_n and rest of the matrix elements bears total number of common associations between two offenders P_{ij}. We represent each P_{ij} offender of the matrix as: P_{ij}=

$$\left\{ \begin{array}{c} \text{=n; When there is n number of associations of the offenders} \\ \text{=0 When there is no association exist between two offenders} \end{array} \right\}$$

3.2.2. Plausible Suspect List

In this paper, we also extended our earlier work [68] of eliciting suspect list for a given unsolved crime. For extracting plausible suspect list, we resolve given a pattern of unsolved crime into its modus operandi, temporal and spatial component, and compare each of this with that of the suspect. Each member of this list has at least committed one crime exhibiting similarity in any or all the components of the given unsolved crime.

The similarity of each component is measured through cosine function, however, following human-centered data mining approach, each of these components is dynamically weighted between 0 and 1 set by the analyst. The final similarity score would be the weighted sum of modus operandi, temporal and spatial component of the crime pattern.

Each member is thus ranked based on the similarity score of their committed crime with that of given unsolved crime. Mathematically Let C_1, C_2, C_3, C_n be the solved crimes committed by perpetrators $P_1, P_2, P_3, ...P_n$ having similarity to any or all the components of the given unsolved crime and hence may be considered as suspects. Let S_1, S_2, S_3, S_nbe the corresponding similarity scores of unsolved crime with these solved crimes. Lets also suppose P_1 has committed the crimes $C_1$and C_2, P_2 has committed the crime C_2, C_3 and P_3 has committed the crimes $C1, C2, C3$,the similarity ranking of each of these suspects then can be calculated using following equations:

$$\text{Rank of } P_1 = S_1 + S_2$$
$$\text{Rank of } P_2 = S_1 + S_3$$
$$\text{Rank of } P_3 = S_1 + S_2 + S_3$$

Now suppose if $S_1 < S_2 < S_3$ which means that given unsolved crime is more similar to the C_3 and least similar to the C_1, then based on the similarity score, though P_1 has appeared in two crimes so is the P_2, but since solved crime P_3 is more similar to the given unsolved crime due to high value of the S_3, therefore it will rank high on plausible suspects list.

A crime having same similarity score may have a different combination of similarity in a various component of the crime pattern. It is represented by an adjacency matrix of order nXm with n numbers of suspects and m numbers of the attributes of the crime pattern. The top row of this matrix contains all the attributes of a crime pattern i.e. spatial component including district and street, temporal component(time of the event) and lastly all the elements of modus operandi. The first column of this matrix contains the name of all the suspects in the list. We tag each cell of the matrix S_{ij} corresponding to each suspect and the crime pattern component either 0 or 1 to record the presence or absence of the similarity. The prime objectives in extracting all these associations are to help analyst in identifying a group of crimes having common crime pattern to reason about the unsolved crime.

The level 1 and level 2 associations distinguishing solved and unsolved crimes generate complex and high dimensional data. Therefore aiming towards both explanatory and exploratory data criminal data analysis, we represent this multidimensional information through an interactive dynamic clustering process to group crimes similarities described in next section.

4. Interactive Clustering for 2-D Crime Space

We incorporated unsupervised machine learning technique of clustering to group and the analyze crimes on the basis of the similarity to fulfill the need of an analyst. We employed spatial, temporal and modus operandi as an investigative lens, to group crimes into a 2D crime space, such that the similarity in a cluster is larger than among the clusters. However, focusing our objective towards a human-machine collaboration and avoiding fully automated or manual system, we incorporated dynamic feature selection for interactive and iterative clustering. A feedback loop from visualization module (Figure 1), drives the analyst to redefine the feature vector selecting or de-selecting any or all of temporal, spatial or modus operandi variables to perform iterative clustering. In this way, the analyst may examine any change in grouping caused by a spatial, temporal and behavioral characteristic of the crimes.

4.1. Dynamic Feature Selection

For a feature vector, we represented location through postcode, street, and town, modus operandi through twelve variables including Entry position in the premises, Entry type, Fixture, Fixture type, Search type, Exit type, and Exit Fixture etc. Each of these modus operandi variables has a set of predefined values. We represented temporal information through month, day, and time of the offense occurred. For time of the offense, we, adopted the idea of conceptual scaling to transform 24 hours of the day into its symbolic value which resulted in four periods of the day: morning (from 6 am to 12 am), afternoon (from 12 pm to 6 pm), evening (from 6 am to 12 pm), and night (from 12 pm to 6 am). Any or all of these attributes may be selected or deselected by the analyst for clustering crimes.

4.2. Categorical Data Handling through VSM

The Feature vector described above involves categorical data. However clustering algorithms works on numerical data, some researchers have demonstrated solution towards this critical challenge of clustering. [10] have illustrated the use of categorical data into clustering algorithm by converting these variables into binary attributes and used 0 or 1 to indicate the categorical value either absent or present in a data record. This approach, however, is not suitable for high dimensional categorical data. Therefore in order to tackle this issue, we employed Vector Space Model (VSM) and through the process of vectorization created a bag of words or (crime terms in this case) from the crime dataset.

The process of vectorization was accomplished following sequences of simple tasks including removing delimiters, converting all words to lower case, removing stop words and stemming words to their base. We also made this preprocessing step interactive as shown in Figure 1 through the use of n-gram models which allow the user not to just use uni-gram models, but also bi-gram and trigram models. The basic idea is to extract unique content-bearing words from the set of crime documents, assign weights to every term measured through Term Frequency-Inverse Document Frequency (TF-IDF) and treat these words as a numerical representation of the features to the clustering algorithm.

Mathematically, Let $C = C_1, C_2, C_3, C_n$be the crime space consisting of N crimes. Each crime $C_i; i = 1, 2...n$ is consisted of n numbers of terms t_1, t_2, t_3, t_n, representing the spatial-temporal and modus operandi information of a crime. We represent a crime C_i through the n-dimensional feature vector in the term space as $C_i = W_1t_1, W_2t_2, W_3t_3.W_nt_n$, where $W_n$is the weight assigned to each term t_j in the crime document C_i through the following relationship.

$$W_n = tf - idf(t_j, C_i) * IDF$$

Where; $tf - idf(t_j, C_i)$ is the frequency of the term j in a crime document i and IDF is the inverse document term frequency calculated as

$$IDF = (1 + log) * \frac{\text{Total crimes documents}}{\text{Total crimes reports in which term j has appeared } t_n}$$

The numerical representation of the crime space C was thus represented through this weighted crime terms matrix consisting of rows as crime documents and columns containing weighted crime terms and was fed into the K-mean clustering algorithm. However, we used cosine similarity as the distance function in the K-mean algorithm rather than using Euclidean. Silhouette analysis was employed to calculate the optimal number of the clusters as required by the K-Mean algorithm.

4.3. Dynamic Numbers of Clusters

The proposed clustering mechanism calculates average silhouette score for a list of predefined values of a number of clusters starting from 3 to 24 with an increment of 3 i.e(3,6..,18,21,24) and chooses the one having the largest value of average silhouette score. However, like the dynamic feature selection, we also provided an interactive user interface to set the required number of the clusters from the user. It thus provides the analyst to incorporate his/her background knowledge in choosing a number of clusters. This process of clustering thus group solved and unsolved crime along with the associated offenders on the basis of the similar characteristic of the crimes.

4.4. Dynamic Configuration of 2D Crime space

Another feature of this interactive clustering is the creation of dynamic 2D cluster space having reconfigurable axes for visualizing the implicit relationship of KPIs with each other. It enables the analyst to observe the relationship between two KPI with respect to each other revealing more insight of the data. Thus for example, if the analyst wishes to examine how crimes (either solved or unsolved) are distributed on the streets of a town, s/he may choose to set these two KPI on either X or Y axis, to see their hidden relationship on 2-dimensional crime space. Likewise, if s/he sets cluster global similarity on X-axis and total offenders on Y-axis, then the clusters would arrange themselves revealing how offenders are distributed over cluster space as a function of clusters global similarity.

We employed multi-dimensional scaling to map the distances of clusters on either of X or Y axis. We first calculated a nxn distance matrix of centroids of each cluster and map each element of this matrix to the configuration points x_1, x_2, x_3, x_n in such a way that the given distances D_{ij} between any two clusters are well approximated by the distances $|x_i - x_j|$. Following this, either X or Y axis of the configuration space, when set to the cluster distance, would arrange clusters in a fashion, such that clusters those perceived to be very similar to each other will be placed near to each other and those are perceived to be very different from each other would appear far away from each other on the chosen axis. This enables the user to easily tag a cluster based on its crime pattern across the generated global similarity map.

5. ASSOCIATIONS VISUALIZATION

The visualization module of our KDD present generated 2D crime space as aggregated or detailed view as shown in Figure 4a.

5.1. Aggregated View

The aggregated view Figure 4a represented through visual Doughnuts, presents a summary of the crime Key process indicators (KPIs) inside a cluster illustrating the associations between unsolved, solved crimes and associated offender. Hovering on each of these Doughnuts shows the crime reports contents inside the clusters as tool-tip. The arcs length of the Doughnut is kept proportional to the unsolved, solved crimes whereas their associated offenders are represented in the center of the Doughnuts.

5.2. Detailed View

The detailed view of the crime space Figure 4b depicts how crimes are related to each other on basis of the similarity inside the cluster. Each cluster is represented as a big gray circle, showing three types of the associations including crime objects i.e. crime and offenders, the type of theses objects i.e. solved and unsolved crime, roles of offenders such as defendant, suspect and victim etc, associations of solved crimes with the offenders and lastly local similarity of the crimes. Hovering on each of these circles shows the information of the crime such as crime reference numbers as tool-tip. An offender association with crimes is visualized through focus and context technique. When an offender is focused or hovered through mouse its association with all of its associated solved crimes in any cluster is highlighted through increasing the size of the related solved crime circles in the clusters, which goes back to normal when hover is off as shown in Figure 4b.

5.3. Offender Space:Knowledge graph

According to visualization literature nodes and links in a tree can signify relations among objects without the constraint of mapping variables onto multidimensional axes. Consequently, we have employed the notion of a dynamic hierarchal bipartite tree called as knowledge graph, to visualize the crime linkages, extracted in level 3 of the association miner. Each node represented through iconic graphic is collapsible and expandable which means a user can click a node of interest to view its underlying children while closing any other node so that only relevant/desired information is placed on the screen.

6. Case study

We also have demonstrated the developed scheme to the police analysts and received positive comments. We tested our developed scheme over Anonymized burglary dataset, consisting of over 1.6 million crime reports and associated offenders or victims information, collected from UK Law Enforcement Agency. The crime reports contain textual data consisting of the crime reference, nominal information, modus operandi description, offense category, location, time and date of the crime occurred along with other related information. Twelve modus operandi variables each having a set of predefined values were used to specify several modus operandi. Several use cases were tested to demonstrate the performance of the scheme, however here we present a use case where offender entered the premises through either "UPVC Door" or "UPVC Window". This information in our Anonymized data is represented in the field "MoFixtureMaterial" and contains the value "Plastic". The objective of this use case is set to find out the crime pattern and other associations. For this crime pattern, the search engine generates a list of solved and unsolved crime having similarities with the given crime pattern.

6.1. Effect of Crime Features on Clusters

For the resulted search data, the first step is to find out the best feature vector that generates the good quality cluster. To answer this, we first examined the effect of feature vector on the number of clusters through silhouette analysis and measured average silhouette coefficients for each pair of the chosen number of clusters and feature vectors. Four sets of feature vectors were taken.The (Full FV) was consisted of all features i.e. temporal, spatial and modus operandi, while other three feature vectors were made of using spatial, temporal and modus operandi features separately. The result presented in Figure 3 shows that the silhouette coefficient value and hence the cluster quality decreases with the increasing number of clusters for all type of feature tested in making clusters. The highest value of silhouette coefficient value also suggests that feature vector consisting of only modus operandi information generated good quality clusters than other sets feature vectors.

6.2. Clusters Similarities

The generated 2D crime cluster space was then studied to examine how the global and local similarity of clusters varies with proximity. Figure 4a and 4b shows the Aggregated and Detailed view of the generated 2D crime cluster space. Both these views compare the global similarities of the clusters (X-axis) with proximity represented by "Town" on (Y-axis). In Figure 4a, it can be seen that, in the town "DEWMAPLE" two clusters i.e. cluster 3 and cluster 4 are more similar(due to less distance between them) as compared to cluster 2 which is at a farther distance than these two clusters. A small cluster (i.e. cluster 5) consisting of five crimes (2 solved and 3 unsolved) at "Yarnforth" town seems to be similar the bigger cluster 1 having 12 crimes in the town "Carsington". When hovering on any of these clusters in the Aggregated view,it shows the

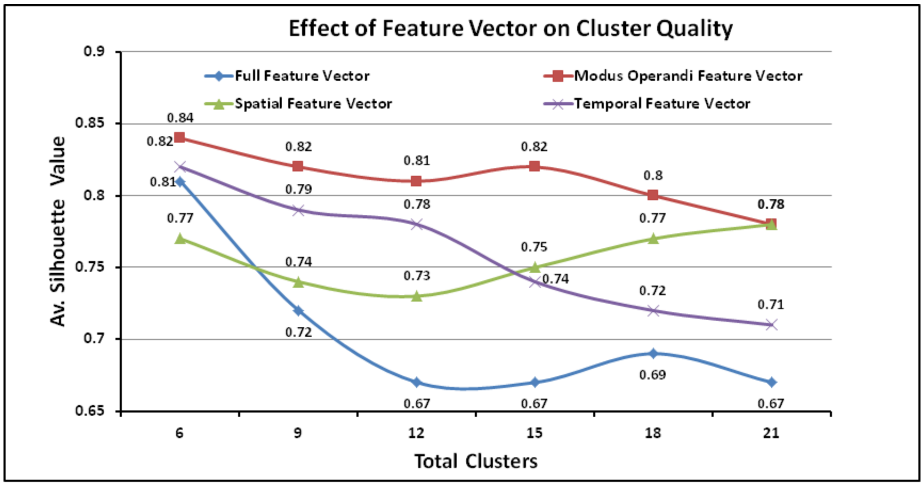


Figure 3: Effect of Feature Vector on Cluster Quality

content of that cluster. For example, when cluster 4 hovers, the tooltip shows the centroid and statistic of the crimes. This cluster has six crimes, five of them committed in the town “DEWMAPLE”, and one crime happened in town “YARNFORTH”. Other trends can also be seen in the Figure 4a.

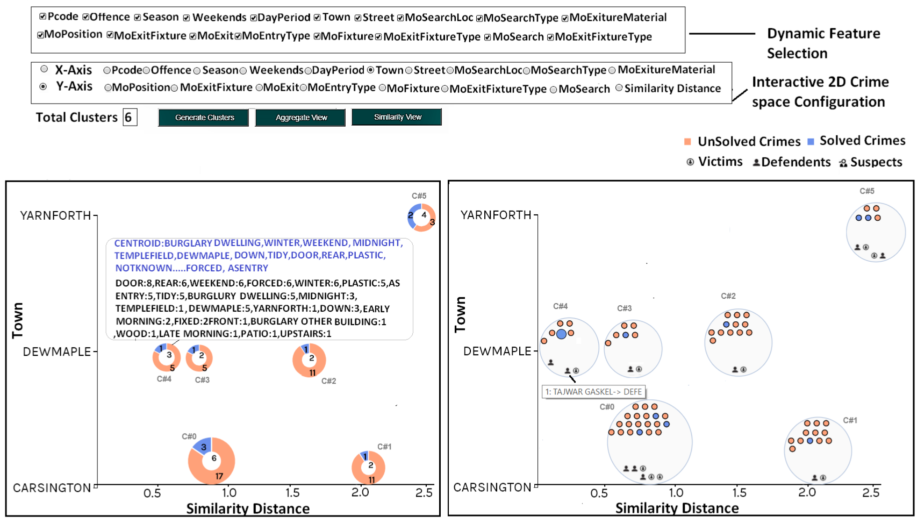
Figure 4b shows the local similarity of the crimes within a cluster along with associated offenders and victims. The offenders of the solved crime are highlighted when hovered as shown in Figure 4b. Additionally, when any of this offender is clicked it generates a knowledge graph of all of his/her associated offender/s as shown in Figure.

4.

6.3. Crime pattern

We then examined generated 2D crime cluster space as shown in Figure 5 to answer three basic questions i.e. what are the hot spots of the given crime pattern, what are associated temporal and modus operandi information. Figure 5a,Figure 5b, Figure 5c,Figure 5d,Figure 5e,and Figure 5f shows the aggregated views of generated 2D cluster space with Y-axis representing (Street,Day of period and modus operandi information including Exit from the premises i.e. MOExit,Fixture material used and search locations in the premises respectively),while for each of these Y-axis, the X-axis is set on proximity represented by Town i.e. where these crime have occurred.

Figure 5a shows that the major crime hot spots for the given crime pattern are found in towns namely “CARSINGTON”, “DEWMAPLE”, “YARNFORTH”. The Figure 5a shows three clusters of similar crimes in the town “DEWMAPLE” at “PAVEMENT ROW”,“LINGSTON CLOSE” and “TEMPLEFIELD” streets. The two big clusters;one having 20 solved and unsolved crimes and other with 12 crimes showing similarity with given crime pattern



☑Pcode ☑Offence ☑Season ☑ Weekends ☑DayPeriod ☑ Town ☑Street ☑MoSearchLoc ☑MoSearchType ☑MoExitureMaterial ☑MoPosition ☑MoExitFixture ☑MoExit☑MoEntryType ☑MoFixture☑MoExitFixtureType ☑MoSearch ☑MoExitFixtureType

○ X-Axis ○Pcode○Offence ○ Season○ Weekends○DayPeriod ● Town○Street○MoSearchLoc○MoSearchType○ MoExitureMaterial

● Y-Axis ○MoPosition○ MoExitFixture ○MoExit○MoEntryType ○MoFixture○MoExitFixtureType ○MoSearch ○Similarity Distance

Total Clusters 6

CENTROID:BURGLARY DWELLING,WINTER,WEEKEND, MIDNIGHT, TEMPLEFIELD,DEWMAPLE, DOWN,TIDY,DOOR,REAR,PLASTIC, NOTKNOWN.....FORCED, ASENTRY

DOOR:8,REAR:6,WEEKEND:6,FORCED:6,WINTER:6,PLASTIC:5,AS ENTRY:5,TIDY:5,BURGLURY DWELLING:5,MIDNIGHT:3, TEMPLEFIELD:1 , DEWMAPLE:5,YARNFORTH:1,DOWN:3,EARLY MORNING:2,FIXED:2FRONT:1,BURGLARY OTHER BUILDING:1 ,WOOD:1,LATE MORNING:1,PATIO:1,UPSTAIRS:1

1: TAJWAR GASKEL-> DEFE

(a) Aggregated View

(b) Detailed View

Figure 4: Crime cluster Space (Aggregated & Detailed View).

are happened at “EASON CLOSE” street of the town “CARSINGTON” and finally one cluster in town “YARNFORTH”

Another projection of the same clusters Figure 5b, rearrange the clusters revealing temporal information of these crime clusters. It can be seen in the Figure 5b that three clusters in town “DEWMAPLE” contain crimes that mostly occurred at mid night, while the two crime clusters in town “CARSINGTON” at “EASON CLOSE” contains crimes that are occurred in early mornings. Additionally The Figure 5c, Figure 5d,Figure 5e and Figure 5f respectively, reveal days of the week and modus operandi information (Exit from premises, fixture material and search locations) respectively, of committed crimes present in these clusters. The crime clusters in town “CARSINGTON” occurred on weekdays and the offenders in most of these cases used UPVC plastic in committing the crimes and they escaped from the premises through “REAR”. Other patterns are also visible in the Figure 5.

Hovering on any of these clusters shows the details of the crimes inside the cluster, for example, in Fig 4a the cluster No 4 in town “DEWMAPLE” have five crimes of “BURGLARY DWELLING” and one of “OTHER BUILDING”, and out of these six crimes five have reported to used “Plastic” as fixture material. These extracted patterns thus indicate that for most of the burglary dwelling crime in the town “DEWMAPLE” offenders have used “UPVC DOOR or WINDOWS” to enter in the premises. These patterns may be used as anchors in generating a hypothesis to facilitate reasoning process towards matching solved and unsolved crimes.

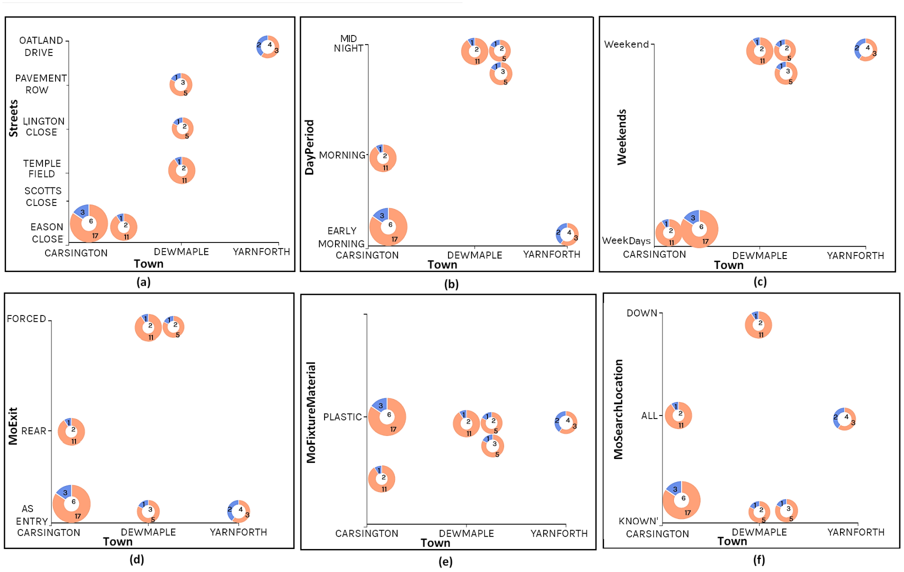


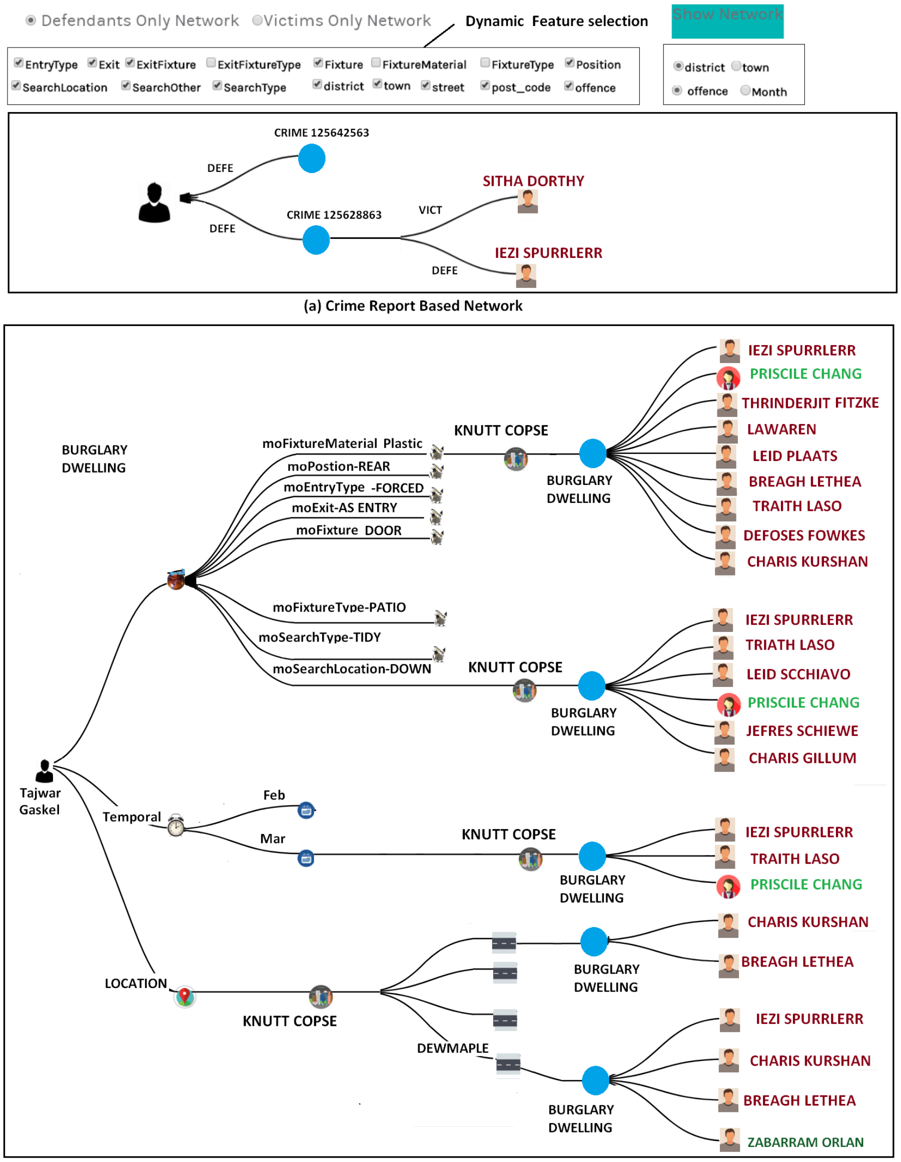
Figure 5: Detailed views of the Interactive Cluster Space.

6.4. Knowledge-Graph

After inspecting the crime pattern, the next anchor is to explore the details of associated offenders, which might be helpful to link unsolved crimes with a plausible suspect/s. It involves answering the questions such as who are the other offenders who have committed similar crimes showing similar characteristics in modus operandi, proximity and time. These associations are examined through Knowledge graph of our proposed scheme and shown in the Figure 6.

Like the clustering Knowledge graph is also made interactive through a selection of a set of similarity attributes for modus operandi, proximity and temporal information to set the associations specifications for the associated perpetrators for a given offender. For example in Figure 4b, the knowledge graph of the associated offender (“TAJWAR GASKEL”) of crime “BURGLARY DWELLING” in cluster 4 is shown in Figure 6. Figure 6a shows the group of offenders with whom “TAJWAR GASKEL”has committed crimes together as they have appeared in the same crime report. “TAJWAR GASKEL”is connected with offender “IEZI SPURRLERR ”, as a defendant through crime report “125628863”,however he is nominated as a single defendant in the crime report “125642563”.

The Figure 6b highlights his spatial, temporal and modus operandi similarities with other defendants. The top section Figure 6b represents the group of offenders having behavioral similarity. For simplicity and sake of the space we have only expanded two nodes of modus operandi i.e. i) the (moFixtureMaterial) i.e. Fixture material used to entered into premises, which in this case is “PLASTIC”, and ii) the (moSearchLocation) i.e. where did he searched in the premises, which in this case is“DOWN”. These two nodes are further expanded



● Defendants Only Network ○Victims Only Network

Dynamic Feature selection

Show Network

☑EntryType ☑Exit ☑ExitFixture □ExitFixtureType ☑Fixture □FixtureMaterial □FixtureType ☑Position
☑SearchLocation ☑SearchOther ☑SearchType ☑district ☑town ☑street ☑post_code ☑offence

●district ○town
● offence ○Month

(a) Crime Report Based Network

(b) STM based Criminal Network

Figure 6: Knowledge graph of an offender: TAJWAR GASKEL.

to narrow the association towards district and offense. These nodes thus answer the questions who are the other offenders who have committed these crimes using same modus operandi(MO).

The “moFixtureMaterial” node shows a group of nine defendants consisting of eight men all of them are white skinned Europeans represented by red text and one woman of Asian origin represented by green text. The “moSearchLocation” node, however, shows six offenders who like “TAJWAR GASKEL” while committing the crime they have searched the “DOWN ”portion of the premises. It can be seen that the three offenders i.e.“IEZI SPURRLERR”,“PRISCILE CHANG”and “TRAITH LASO”bear the similarities in these modi operandi. This means these persons because of the similarity in their modus operandi signature, may be thought to have more close associations with “TAJWAR GASKEL ”.

The temporal node of the graph Figure 6b shows that “TAJWAR GASKEL” has committed in the month of the “MAR”and “FEB”. When the“MAR”node is further expanded in the tree, it shows he has committed the offense “BURGLARY DWELLING”in the district “KNUTT COPSE”. “IEZI SPURRLERR”, “PRISCILE CHANG ”, and “TRAITH LASO ”have also committed crimes in the same month, however separately.

The last node i.e. the spatial node Figure 6d highlights the criminal groups with the spatial similarity. Two other persons one Asian namely“PRISCILE CHANG ”and an European namely “IEZI SPURRLERR” have committed similar crimes in the town “DEWMAPLE” of the “KNUTT COPSE” district. This graph thus shows that Both “IEZI SPURRLERR ”,“PRISCILE CHANG ”have exhibited very high similarity in crime pattern with the “TAJWAR GASKEL”.

6.5. Plausible Suspect List Graph

An analyst may also be interested to see who could be possible suspect of an unsolved crime. In Figure 4b, for example, unsolved crime having crimeref “127553987”, which is grouped with a solved crime in cluster 4, when clicked, another knowledge graph representing a plausible suspect list for the clicked unsolved crime is weaved and shown in the Figure 7.

The root node for the plausible suspect list is the unsolved crime represented through the dark orange circle, showing a list of the offenders of the similar solved crimes. Like the knowledge graph of the offender, The root node is branched into three child nodes each for the spatial, temporal and modus operandi component and for simplicity we have expanded spatial and two MO nodes. The fourth level shows suspects along with their gender ethnicity and age group.

In the Figure 7 the offender “LEID SCCHIAVO” has committed the similar crime of burglary dwelling in the same town “DEWMAPLE”, and modus operandi i.e. the Mofixturematerial attributed is also similar to that of unsolved crime. In addition to this, “LEID SCCHIAVO” is also present in the co-offender group of the “TAJWAR GASKEL”who has committed a similar Burglary Dwelling in the town “DEWMAPLE” as can be seen in the cluster 4

Unsolved Crime Details:

Crime_Ref:127553987

Offence Burglary Dwelling reported in district Morrbridge in town Dewmaple on Fri,3rd Feb 2007 . Modus operandi used MO_Position:REAR,MO_Fixture:,MO_FixtureType:CASEMENT,MoFxtureMaterial:PLASTIC,MO_EntryType:GLASS, MO_SearchLocation:ALL,MO_SearchType:TIDY,MO_Exit:AS ENTRY

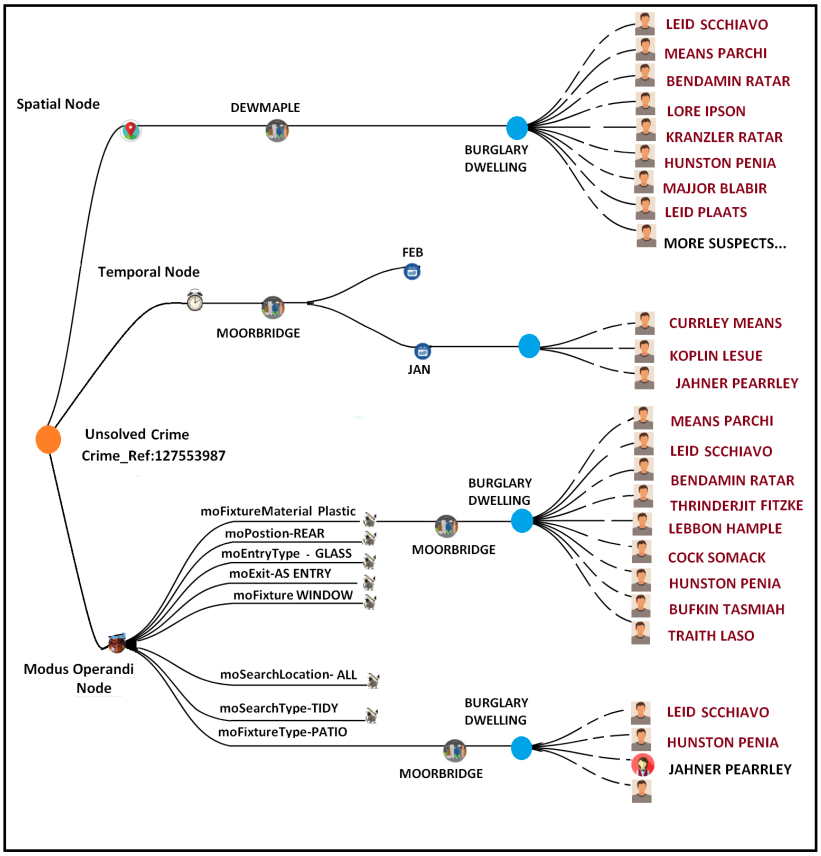


Figure 7: Knowledge Graph:Plausible Suspects List for Un-solved Crime

of Figure 4b. The suspect list also extracts three offenders who have committed a similar crime in the month of “JAN” in the same district.

The suspect list widget thus along with Co-offender network knowledge graph gives the insight to facilitate analyst in making hypotheses revealing the interesting hidden relationship.

7. Conclusion

We presented an human centered discovery scheme using text mining of the crime reports for extracting multi-level associations among crime entities. It integrates human interaction for analyst feed towards model building and fine-tuning of machine learning algorithm through the interactive user interface. It is able to extract plausible associations identifying crime pattern, clusters of similar crimes, co-offender network and suspect list based on spatial-temporal and modus operandi similarity, computed through multiple distance functions including euclidean cosine and Jacquard functions. We demonstrated the use of graph theory to weave a heterogeneous associations network for a given type of input node, highlighting either offender network, plausible suspect list depending upon the nature of given input node.

We also have addressed the issue of categorical variables in partition clustering through vector space model to group the similar crimes. The analyst is able to create a 2D re-configurable crime space to see the hidden pattern in the crime text data, implemented through Dynamic feature selection and Multidimensional scaling.The Clusters quality measured through silhouette analysis has shown that that modus operandi is a better feature vector to cluster the similar crimes.

We have demonstrated the use of this scheme for a given crime pattern where ”UPVC Door” or windows” is reported as modus operandi in committing a burglary, our scheme has visualized the similarity of solved and unsolved crimes in a 2D crime space revealing temporal and spatial crime pattern along with co-offender network and suspect list for unsolved crimes that involve the given crime pattern. The police analysts during a preliminary user feedback have given positive feedback indicating that this prototype has potential to improve the efficiency of the criminal investigation process.

Such associations can provide the basis for activating ideas/thoughts/tentative or plausible conclusions, that could trigger new lines of inquiry.We do think that the scheme with the proposed visualized widgets may be helpful to uncover the interesting aspects of the reasoning for crime matching. However, we do acknowledge that it does not capture all the problems. Our framework thus enables crime analysts to see the possibility of linkages between data and to make assessment rather than a recommendation.

8. Acknowledgments

The research leading to the results reported here has received funding from the European Union Seventh Framework Program through Project VALCRI, European Commission Grant Agreement N FP7-IP-608142, awarded to Middlesex University and partners

References

[1] L. Cao, T. Joachims, C. Wang, E. Gaussier, J. Li, Y. Ou, D. Luo, R. Zafarani, H. Liu, G. Xu, Z. Wu, G. Pasi, Y. Zhang, X. Yang, H. Zha, E. Serra, V. S. Subrahmanian, Behavior informatics: A new perspective, IEEE Intelligent Systems 29 (4) (2014) 62–80.

[2] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (1) (2017) 48 – 56.

[3] A. Holzinger, Human-computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together?, in: International Conference on Availability, Reliability, and Security, Springer, 2013, pp. 319–328.

[4] A. Holzinger, I. Jurisica, Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions, in: Interactive knowledge discovery and data mining in biomedical informatics, Springer, 2014, pp. 1–18.

[5] A. Holzinger, Machine learning for health informatics, in: Lecture Notes in Computer Science, 2016.

[6] nan Arn, M. K. Erpam, Y. Saygn, I-twec: Interactive clustering tool for twitter, Expert Systems with Applications 96 (2018) 1 – 13.

[7] S. Amershi, J. Fogarty, D. Weld, Regroup: Interactive machine learning for on-demand group creation in social networks, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 21–30.

[8] F. M. F. Gonalves, I. R. Guilherme, D. C. G. Pedronette, Semantic guided interactive image retrieval for plant identification, Expert Systems with Applications 91 (2018) 12 – 26.

[9] T. Kulesza, S. Amershi, R. Caruana, D. Fisher, D. Charles, Structured labeling for facilitating concept evolution in machine learning, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, pp. 3075–3084.

[10] M. R. Keyvanpour, M. Javideh, M. R. Ebrahimi, Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science 3 (2011) 872 – 880.

[11] D. G. C. Oatley, P. J. Zeleznikow, D. B. W. Ewart, Matching and predicting crimes, in: In Proceedings of the Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence, 2004.

[12] B. W. Wong, N. Kodagoda, How analysts think, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 60 (1) (2016) 178–182.

[13] A. of Chief Police Officers, N. P. I. Agency, Practice advice on analysis (2008).

[14] N. Qazi, B. L. W. Wong, N. Kodagoda, R. Adderley, Associative search through formal concept analysis in criminal intelligence analysis, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 1917–1922.

[15] L. E. Cohen, M. Felson, Social change and crime rate trends: A routine activity approach, American sociological review (1979) 588–608.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The kdd process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34.

[17] M. Brehmer, T. Munzner, A multi-level typology of abstract visualization tasks, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2376–2385.

[18] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, D. A. Keim, Visual interaction with dimensionality reduction: A structured literature analysis, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 241–250.

[19] W. Jentner, G. Ellis, F. Stoffel, D. Sacha, D. Keim, A visual analytics approach for crime signature generation and exploration, in: IEEE VIS2016 Workshop on Temporal & Sequential Event Analysis, 2016.

[20] P. Saleiro, E. M. Rodrigues, C. Soares, E. Oliveira, Texrep: A text mining framework for online reputation monitoring, New Generation Computing 35 (4) (2017) 365–389.

[21] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, A. Wijayasiri, Crime analytics: Analysis of crimes through newspaper articles, in: 2015 Moratuwa Engineering Research Conference (MERCon), 2015, pp. 277–282.

[22] M. Alruily, A. Ayesh, H. Zedan, Crime profiling for the arabic language using computational linguistic techniques, Information Processing & Management 50 (2) (2014) 315 – 341.

[23] M. A. Valle, G. A. Ruz, R. Morrs, Market basket analysis: Complementing association rules with minimum spanning trees, Expert Systems with Applications 97 (2018) 146 – 162.

[24] F. Zarrinkalam, M. Kahani, E. Bagheri, Mining user interests over active topics on social networks, Information Processing & Management 54 (2) (2018) 339 – 357.

[25] S. hsien Liao, H. ko Chang, A rough set-based association rule approach for a recommendation system for online consumers, Information Processing & Management 52 (6) (2016) 1142 – 1160.

[26] I. Viktoratos, A. Tsadiras, N. Bassiliades, Combining community-based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems, Expert Systems with Applications 101 (2018) 78 – 90.

[27] S. Piri, D. Delen, T. Liu, W. Paiva, Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications, Expert Systems with Applications 94 (2018) 112 – 125.

[28] W. Chen, C. Xie, P. Shang, Q. Peng, Visual analysis of user-driven association rule mining, Journal of Visual Languages & Computing 42 (2017) 76 – 85.

[29] H. Chi, Z. Lin, H. Jin, B. Xu, M. Qi, A decision support system for detecting serial crimes, Knowledge-Based Systems 123 (2017) 88 – 101.

[30] A. Borg, M. Boldt, J. Eliasson, Detecting crime series based on route estimation and behavioral similarity, in: 2017 European Intelligence and Security Informatics Conference (EISIC), 2017, pp. 1–8.

[31] U. Thongsatapornwatana, W. Lilakiatsakun, A. Kawbunjun, T. Boongoen, Analysis of criminal behaviors for suspect vehicle detection, in: 2017 Twelfth International Conference on Digital Information Management (ICDIM), 2017, pp. 15–20.

[32] J. de Zoete, M. Sjerps, D. Lagnado, N. Fenton, Modelling crime linkage with bayesian networks, Science and justice 55 (3) (2015) 209–217.

[33] J. de Zoete, M. Sjerps, R. Meester, Evaluating evidence in linked crimes with multiple offenders, Science and justice 57 (3) (2017) 228–238.

[34] M. S. Vural, M. Gök, Criminal prediction using naive bayes theory, Neural Computing and Applications 28 (9) (2017) 2581–2592.

[35] C. Bennell, D. Canter, Linking commercial burglaries by modus operandi: tests using regression and roc analysis, Science & Justice 42 (3) (2002) 153 – 164.

[36] M. Tonkin, J. Woodhams, R. Bull, J. W. Bond, Behavioural case linkage with solved and unsolved crimes, Forensic Science International 222 (1) (2012) 146 – 153.

[37] J.-H. Wang, C.-L. Lin, An association model based on modus operandi mining for implicit crime link construction, in: Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM '11, IEEE Computer Society, Washington, DC, USA, 2011, pp. 548–550.

[38] L. Ding, D. Steil, M. Hudnall, B. Dixon, R. Smith, D. Brown, A. Parrish, Perpsearch: an integrated crime detection system, in: Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on, IEEE, 2009, pp. 161–163.

[39] A. Onan, S. Korukolu, H. Bulut, A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification, Information Processing & Management 53 (4) (2017) 814 – 833.

[40] F. Biscarri, I. Monedero, A. Garca, J. I. Guerrero, C. Len, Electricity clustering framework for automatic classification of customer loads, Expert Systems with Applications 86 (2017) 54 – 63.

[41] M. Wang, F. Min, Z.-H. Zhang, Y.-X. Wu, Active learning through density clustering, Expert Systems with Applications 85 (2017) 305 – 317.

[42] Y.-H. Hu, Y.-L. Chen, H.-L. Chou, Opinion mining from online hotel reviews a text summarization approach, Information Processing & Management 53 (2) (2017) 436 – 449.

[43] G. G. Dagher, B. C. Fung, Subject-based semantic document clustering for digital forensic investigations, Data & Knowledge Engineering 86 (2013) 224 – 241.

[44] B. J. Reich, M. D. Porter, Partially supervised spatiotemporal clustering for burglary crime series identification, Journal of the Royal Statistical Society: Series A (Statistics in Society) 178 (2) (2015) 465–480.

[45] A. Borg, M. Boldt, N. Lavesson, U. Melander, V. Boeva, Detecting serial residential burglaries using clustering, Expert Systems with Applications 41 (11) (2014) 5252 – 5266.

[46] L. S. Thota, M. Alalyan, A. O. A. Khalid, F. Fathima, S. B. Changalasetty, M. Shiblee, Cluster based zoning of crime info, in: 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), 2017, pp. 87–92.

[47] Q. Bsoul, J. Salim, L. Q. Zakaria, An intelligent document clustering approach to detect crime patterns, Procedia Technology 11 (2013) 1181 – 1187, 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013.

[48] S. Alelyani, J. Tang, H. Liu, Feature selection for clustering: A review., Data Clustering: Algorithms and Applications 29 (2013) 110–121.

[49] K.-C. Lin, K.-Y. Zhang, Y.-H. Huang, J. C. Hung, N. Yen, Feature selection based on an improved cat swarm optimization algorithm for big data classification, The Journal of Supercomputing 72 (8) (2016) 3210–3221.

[50] K. K. Bharti, P. Singh, A three-stage unsupervised dimension reduction method for text clustering, Journal of Computational Science 5 (2) (2014) 156 – 169.

[51] H. Lee, J. Kihm, J. Choo, J. Stasko, H. Park, ivisclustering: An interactive visual document clustering via topic modeling, in: Computer Graphics Forum, Vol. 31, Wiley Online Library, 2012, pp. 1155–1164.

[52] P. Bruneau, P. Pinheiro, B. Broeksema, B. Otjacques, Cluster sculptor, an interactive visual clustering system, Neurocomputing 150 (2015) 627 – 644.

[53] J. Krause, A. Perer, E. Bertini, Infuse: Interactive feature selection for predictive modeling of high dimensional data, IEEE Transactions on Visualization and Computer Graphics 20 (12) (2014) 1614–1623.

[54] A. Sanchez, C. Soguero-Ruiz, I. Mora-Jimnez, F. Rivas-Flores, D. Lehmann, M. Rubio-Snchez, Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions, Expert Systems with Applications 100 (2018) 182 – 196.

[55] Y. Abdelsadek, K. Chelghoum, F. Herrmann, I. Kacem, B. Otjacques, Community extraction and visualization in social networks applied to twitter, Information Sciences 424 (2018) 204 – 223.

[56] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2) (2004) 026113.

[57] A. M. Fiscarelli, A. Beliakov, S. Konchenko, P. Bouvry, A degenerate agglomerative hierarchical clustering algorithm for community detection, in: Intelligent Information and Database Systems, Springer International Publishing, Cham, 2018, pp. 234–242.

[58] E. Ferrara, P. D. Meo, S. Catanese, G. Fiumara, Detecting criminal organizations in mobile phone networks, Expert Systems with Applications 41 (13) (2014) 5733 – 5750.

[59] K. He, Y. Li, S. Soundarajan, J. E. Hopcroft, Hidden community detection in social networks, Information Sciences 425 (2018) 92 – 106.

[60] R. Al-Zaidy, B. C. Fung, A. M. Youssef, F. Fortin, Mining criminal networks from unstructured text documents, Digital Investigation 8 (34) (2012) 147 – 160.

[61] W. Didimo, G. Liotta, F. Montecchiani, Network visualization for financial crime detection, J. Vis. Lang. Comput. 25 (4) (2014) 433–451.

[62] H. Isah, D. Neagu, P. Trundle, Bipartite network model for inferring hidden ties in crime data, in: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2015, pp. 994–1001.

[63] J. Schroeder, J. Xu, H. Chen, M. Chau, Automated criminal link analysis based on domain knowledge: Research articles, J. Am. Soc. Inf. Sci. Technol. 58 (6) (2007) 842–855.

[64] A. J. Park, H. H. Tsang, P. L. Brantingham, Dynalink: A framework for dynamic criminal network visualization, in: 2012 European Intelligence and Security Informatics Conference, 2012, pp. 217–224.

[65] J. Stasko, C. Görg, Z. Liu, Jigsaw: Supporting investigative analysis through interactive visualization, Information Visualization 7 (2) (2008) 118–132.

[66] F. Ozgul, M. Gok, Z. Erdem, Y. Ozal, Detecting criminal networks: Sna models are compared to proprietary models, in: 2012 IEEE International Conference on Intelligence and Security Informatics, 2012, pp. 156–158.

[67] Y. Sun, Mining heterogeneous information networks.

[68] N. Qazi, B. L. W. Wong, Behavioural tempo-spatial knowledge graph for crime matching through graph theory, in: 2017 European Intelligence and Security Informatics Conference (EISIC), 2017, pp. 143–146.