Received: 25 January 2023 | Revised: 31 October 2023 | Accepted: 4 December 2023

DOI: 10.1111/exsy.13524

ORIGINAL ARTICLE

Expert Systems WILEY

An AI knowledge-based system for police assistance in crime investigation

Carlos Fernandez-Basso[1,2] | Karel Gutiérrez-Batista[1] | Juan Gómez-Romero[1] | M. Dolores Ruiz[1] | Maria J. Martin-Bautista[1]

[1]Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

[2]Causal Cognition Lab, University College London, London, UK

Correspondence
Carlos Fernandez-Basso, Causal Cognition Lab, University College London, London, UK.
Email: carlos.basso@ucl.ac.uk, cjferba@decsai.ugr.es

Funding information
EU-funded Margarita Salas programme NextGenerationEU; European Union's Horizon 2020 Research and Innovation Programme, Grant/Award Numbers: 786687, 101121309; NextGenerationEU/PRTR; MCIN/AEI, Grant/Award Number: PID2021-123960OB-I00; ERDF, Grant/Award Number: TED2021-1289402B-C21

Abstract

The fight against crime is often an arduous task overall when huge amounts of data have to be inspected, as is currently the case when it comes for example in the detection of criminal activity on the dark web. This work presents and describes an artificial intelligence (AI) based system that combines various tools to assist police or law enforcement agencies during their investigations, or at least mitigate the hard process of data collection, processing and analysis. The system is an early warning/early action system for crime investigation that supports law enforcement with different processes to collect and process data as well as having knowledge extraction tools. It helps to extract information during the investigation of a criminal case or even to detect possible criminal hotspots that may lead to further investigation or analysis of a criminal case Abu Al-Haija et al. (2022, *Electronics*, 11, 556). The functionality of the proposed system is illustrated through several examples using data collected from the dark web, which includes advertisements offering firearms-related products.

KEYWORDS

artificial intelligence, association rules, crime detection, knowledge repository

1 | INTRODUCTION

Artificial intelligence (AI) tools have gained prominence over the last decade to assist in very diverse activities. The fight against crime, delinquency and terrorism is no exception, although it poses some challenges: for instance, the functioning of tools must be transparent for the law enforcement agencies (LEAs) and must avoid discrimination in the obtained results. In addition, the AI algorithms, their functioning and the obtained results should be understandable by LEAs in order to assist them in their daily work.

The use of AI is particularly interesting when big data volumes have to be analysed, as in the case of social media or the dark net. The increasing use of these channels for criminal or illegal activities and the massive number of publications to be analysed is one of the main challenges facing police forces.

Moreover, the inherent difficulty of extracting information from natural language is an important factor in automatically using textual information for advanced knowledge discovery Rawat et al. (2021). The combination of natural language processing (NLP) technologies with knowledge discovery (KD) tools seems to be the new direction to follow to extract more information, not only from numerical data but also from textual data

All authors contributed equally to this study.

as it can be seen for instance in Griol-Barres et al. (2020). In this regard and focusing on the general ambit of crime, the existing analysis/forecasting techniques have been used in an isolated way as can be derived from the literature (see for instance Hassani et al. (2016), Li (2021) and Qayyum and Dar (2018) where reviews for data mining (DM) tools in the ambit of crime can be found).

In this paper, we propose and describe a system based on some AI tools (Fernandez-Basso et al., 2016, Fernandez-Basso, Ruiz, et al., 2019) to assist LEAs during their investigations, helping or at least mitigating the hard process of data collection, processing and analysis. The proposed system is not intended to detect crime automatically but to provide assistance with new insights and findings that may assist in the investigation of a criminal case or even in the identification of possible criminal spots that can derive from a criminal case for further investigation. This methodology is known as the EW/EA (Early Warning/Early Action), which is related to the SOCTA[1] methodology used by EUROPOL. The EW encompasses the identification of "weak signals", warnings and new insights that can be used for support at both strategic and operational levels, to develop, in a further step, an EA which includes the different measures/decisions to mitigate, prevent or prepare future security policies.

The system is composed of different parts utilising several AI tools and algorithms starting with automatic crawling and NLP extraction tools, followed by a Knowledge Repository (KR) component containing processed knowledge that the user (in our case law enforcement agencies) has to provide with technical assistance, and ending with the application of KD tools that analyse the collected data to find new relations and insights. This extracted new knowledge could be used to (1) further enrich the knowledge repository, (2) help LEAs in their situation assessment process, and (3) find new information to consider in the criminal case under investigation.

The main novelty of proposed system architecture is that (1) it implements a complete system capable of starting the analysis from the raw data improving these data through the knowledge base including provenance, expert knowledge, and privacy by design to the last step consisting in the visualisation of the results and (2) it allows a continuous evolution by incorporating not only the initial expert knowledge provided by LEAs but also the knowledge provided by the KD tools such as the Association Rules (see Section 5). We have also compared the proposal with other similar systems in Section 2 to highlight its main features.

The proposed system architecture has been employed in the EU-funded COPKIT[2] project. The COPKIT project addresses the problem of analysing, investigating, mitigating and preventing the use of new information and communication technologies by organised crime and terrorist groups. To this end, COPKIT proposes an intelligence-led early warning (EW)/early action (EA) system for both strategic and operational levels. It should be noted that, due to security restrictions, some details cannot be provided.

The main system has been summarised in Section 3, highlighting the two components that we will explain in more depth: (a) the Knowledge Repository and (b) a KD tool called Association Rules. We also present a use case for the system based on the dataset called "Grams"[3] which is a structured dataset containing a collection of advertisements from different darknet markets. In particular, we have focused on those ads that offer weapons-related items. Using this dataset, we will describe the functionality of KR and KD components.

The paper is structured as follows. Next section provides a brief overview about other projects and proposals that aim to assist LEAs in the fight against crime. Section 3 presents the system architecture. Section 4 describes the knowledge repository component and some of its functionalities. Section 5 explains the knowledge discovery tool for the extraction of frequent itemsets and association rules. Section 6 develops a use case based on firearms trafficking to illustrate the operation of the system. Finally, the paper finishes with the conclusions.

2 | RELATED WORKS

One of the European Council objectives is the fight against crime while preserving the citizens' rights. Therefore, we can find several EU funded projects pursuing this goal.

One of these projects is the TITANIUM (2020) project, which researches, develops, and validates novel data-driven techniques and solutions designed to support law enforcement agencies. In this way, it helps to investigate criminal or terrorist activities involving virtual currencies and/or underground markets on the darknet. The result of TITANIUM is a set of services and forensic tools, that operate within a privacy and data protection environment that is configurable according to local legal requirements.

Another project with the same theme as the COPKIT (2021) project is TENSOR (2019) project. This is another example where the primary goal is to keep people safe. The project, which is funded by the EU under the Horizon 2020 programme, seeks to develop a platform offering Law Enforcement Agencies fast and reliable planning and prevention capabilities for the early detection of terrorist activities, radicalisation and recruitment. This project focuses on the collection of data in order to extract interpretative information such as the correlation of information, validation of sources, extraction of events, hidden meanings and high-level interpretations. All of this aggregated and summarised in a final application.

The i-LEAD project (2023) project also covers the fight against crime, focusing on the needs of LEA users who are the main sources of expertise. These are divided into five groups of experts. Each group defines the current situation regarding the use of technology by law enforcement agencies, assesses baseline capabilities through a capability map, identifies capability gaps and opportunities for innovation, defines priority areas for improving performance through innovative methods, and identifies potential areas for standardisation at EU level. This project is more focused on the management and analysis of the requirements and capabilities of each LEA, as well as the improvement and development of new technological capabilities for them. The MAGNETO (2021) project and Pourhabibi et al. (2021) aim to establish a continuously improving crime

TABLE 1 Comparison of main features of different EU-financed projects for the fight against crime.

Name	Data collection	Big Data	AI tools	Data enrichment	Early detection	Statistic tools	HMIs/visualisation tools
COPKIT	✓	✓	✓	✓	✓	✓	✓
TENSOR (2019)	✓	×	×	×	✓	✓	×
e-POOLICE	×	×	×	✓	×	✓	×
CC-DRIVER (2023)	×	×	✓	×	✓	✓	×
AIDA (2020)	×	✓	✓	×	×	✓	×
ANITA (2021)	×	✓	✓	×	×	✓	×
TITANIUM (2020)	✓	×	×	×	×	✓	✓
i-LEAD project (2023)	✓	×	×	×	×	×	✓
MAGNETO (2021)	✓	×	✓	✓	×	×	✓

prevention and investigation scheme. For that, heterogeneous data streams are transformed into knowledge bases according to a sophisticated representation model, which are then processed and fused using semantic technologies. The results are visually represented by immersive HMIs (Human Machine Interfaces), enabling timely and accurate decision-making, situational awareness and court-proof evidence extraction.

We can also distinguish another group of projects such as: ANITA (2021), CC-DRIVER (2023) and AIDA (2020) projects. All of them are based on the use of Data Mining and Artificial Intelligence technologies to extract information of interest to law enforcement agencies. A less recent project with similar objectives is the e-POOLICE project. In it, we worked on a tool that allowed an effective and efficient exploration of raw and open information sources, developing an intelligent environmental radar that uses a knowledge repository to enrich the exploration. A key part of this process is a semantic filtering to identify data elements that may constitute weak signals of emerging organised crime threats, fully exploiting the concept of crime hubs, crime indicators and enablers as understood by its user partners.

In this paper, we describe the system architecture employed in the COPKIT project, as well as a set of tools to fight crime using different data collection, enrichment, and extraction of hidden knowledge. These tools have been applied to the analysis of dark net data, supported by a knowledge repository module. This repository is capable of enriching the data to obtain better results when applying the different knowledge discovery (KD) tools. This system has great capabilities and encompasses many tools that allow it to be distinguished in several aspects from the projects discussed above. As can be seen in Table 1, our system allows the use of innovative tools using Big Data technology, as well as having a repository of knowledge that enriches the data and the results, thus allowing LEA users to obtain more interesting results.

3 | SYSTEM ARCHITECTURE

The developed system comprises different artificial intelligence technologies that are able to extract potentially new and useful knowledge in an automatic and unsupervised way, that is, no intelligence has to be given by the end-user in advance about the data that is being analysed. However, the knowledge repository module can collect the intelligence in advance and be re-utilised and updated by the different KD tools under the LEAs supervision. In this way, this system can be seen as the initial step in the investigation process to gather new information, which can be subsequently examined in order to discard irrelevant data.

Therefore, the components included in the system are placed in the first step of the analysis procedure, and they help to describe the collected data offering new insights that cannot be inferred from a glance to the data or the volume of data is so big, that cannot be analysed manually. In any case, human intervention is necessary to avoid any automatic decision by the tool. In this regard, it is important to remember that the system assists the user in their posterior process of decision-making and does not make any decision by itself.

According to the architecture depicted in Figure 1, it is assumed that certain kind of data has been previously crawled, such as, for instance, a number of advertisements in the dark web, or a network among users in some forums on the dark web. Typically, these collected datasets must be pre-processed before applying any discovery knowledge tools. Natural Language Processing tools such as Part of Speech and Entity Recognition can be used in this pre-processing stage. Additional data structuring processes are performed to organize data in a structured way, such as a table or an Entity-Relation database.

Once the datasets have been processed, they are stored in a Knowledge Repository (see Section 4), where they are enriched with a priori and/or expert knowledge. This enhanced knowledge representation is used by the components of the KD tools to empower the semantic capabilities of the mining processes. The output of these processes assists the LEAs in the discovery of new relationships and insights to assess the situation.

In the next section, the Knowledge Repository and the Knowledge Discovery components are explained in detail.

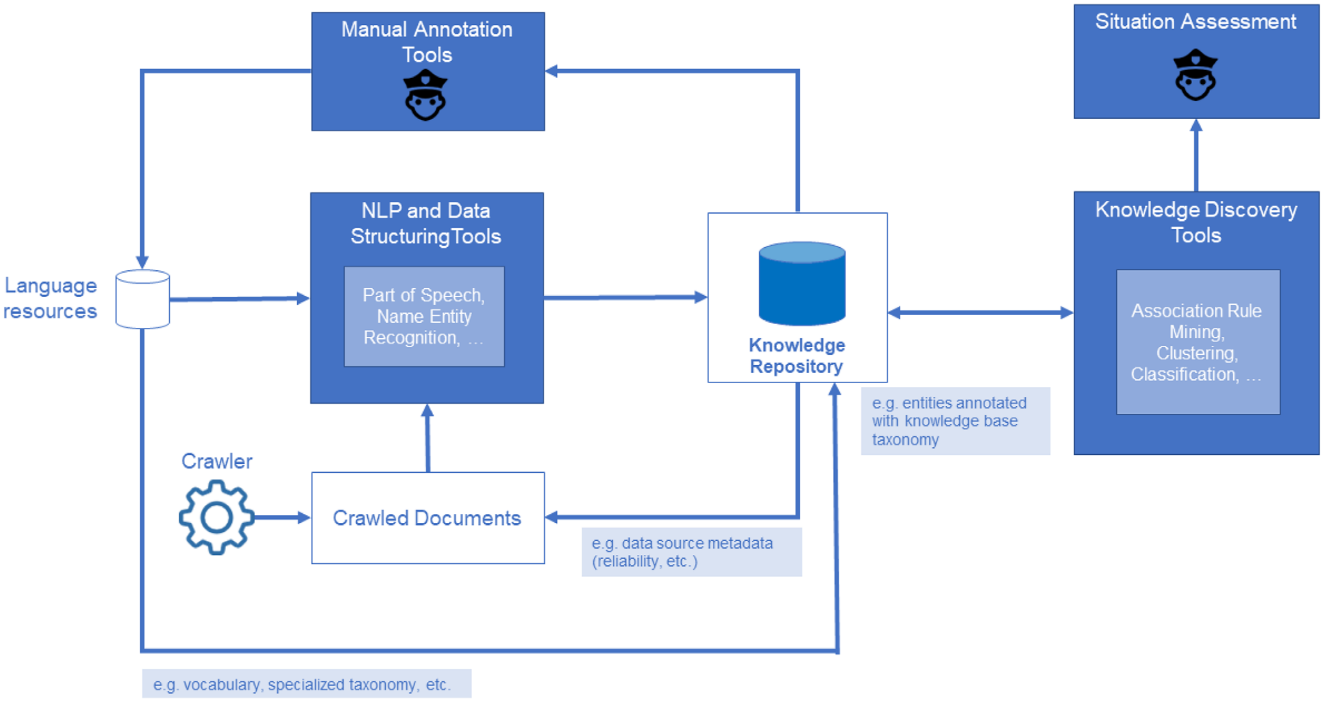

Manual Annotation Tools
Situation Assessment
NLP and Data StructuringTools
Part of Speech, Name Entity Recognition, …
Language resources
Knowledge Repository
Knowledge Discovery Tools
Association Rule Mining, Clustering, Classification, …
e.g. entities annotated with knowledge base taxonomy
Crawler
Crawled Documents
e.g. data source metadata (reliability, etc.)
e.g. vocabulary, specialized taxonomy, etc.

FIGURE 1 Architecture of the system.

4 | KNOWLEDGE REPOSITORY COMPONENT

This section introduces the main concepts and functionalities of the KRC (Knowledge Repository Component). The KRC aims to represent and manage expert and learnt information relevant to the Early Warning/Early Action ecosystem.

4.1 | The concept of the KRC

The KRC implements a formal model in the form of an ontology for knowledge storage and distribution, improving the capabilities of the LEAs at the investigative and strategic level by supporting analysts involved in information management, offering additional metadata for the contextualisation of LEAs investigations.

The KRC serves several functions: (1) it provides the knowledge that drives the execution of some technical components (e.g., starting points for the crawler); (2) it enriches input data used by another component with a priori and previously obtained knowledge; (3) it acts as a storage repository for the outputs of other components.

4.2 | Semantic data models

The KRC is defined using the RDF (resource description framework) and the OWL (ontology web language) languages. These languages are standards proposed by the W3C (world wide web consortium)–in the context of the Semantic Web initiative– that respectively allow representing semi-structured data and logical restrictions in a flexible and normalized way.

RDF (resource description framework) is the W3C standard language to describe resources in the Semantic Web Klyne and Carroll (2004). RDF allows metadata to be asserted in the form of triples, that is, statements relating an object, a property, and a value. For instance, it can be stated that [John] (subject) [has email address] (predicate) [john@doe.com] (object). Subjects, predicates and objects are identified by their URI, a generalization of URLs (uniform resource locator) with a similar structure, but which can be used to identify things or concepts in an unambiguous way without necessarily returning an electronic representation of them as it happens with web pages. Objects can also be literals, that is, strings of characters that can have a type, like a number or a date, or be untyped, like in free text.

RDFS (RDF schema) defines an RDF vocabulary that can be used to express logical relations between resources Manola et al. (2014). For instance, a resource can be declared as a class and another as an instance of that class, or a class can subsume another class; for example, citizens are human beings, and cocaine is a drug, respectively. OWL extends RDFS with additional vocabulary to express more complex logical statements and inferring new facts (W3C OWL Working Group, 2012).

SPARQL (Harris & Seaborne, 2013) is a W3C standard query language that allows retrieving information from an RDF triplestore. The most basic feature of SPARQL is specifying a set of triples where variables can appear as subjects, predicates or objects of any triple. Using the SELECT form in the query, the SPARQL engine will return either the set of mappings of variables to values that conduct the query inside the KB (i.e., the possible ways the KB can satisfy the query), or an empty set if no match occurs.

4.3 | Knowledge included in the KRC

This section describes the knowledge included in the KRC. To address the system requirements, the KRC includes knowledge regarding:

- Metadata to describe the characteristics of knowledge
- General knowledge
- Domain knowledge about each specific use case

Metadata in the KRC is used to characterise the "chain of custody" of a piece of data. The most important metadata in the KRC is provenance, which is essential to support result traceability and trust assessment. By provenance, we mean which process has led to a conclusion and which actors (human or artificial) have been involved. LEAs and authorities commonly use this concept in criminal prosecution because they need to precisely identify the individuals who have provided information to solve a case.

Similarly, as it is done in these situations, the KRC considers credibility, reliability and similar confidence assessment values of sources and processes that affect data quality and results validity. Provenance is represented in the KRC by using the standard publicly available PROV-O ontology (Belhajjame et al., 2012).

The general knowledge in the KRC is composed of several knowledge bases imported from the Linked Open Data web. Specifically, we have considered the following sources:

- DBPedia knowledge base, a structured graph extracted from Wikimedia data publicly available (Graua et al., 2008).
- YAGO2, an open knowledge base automatically built from Wikipedia, GeoNames and WordNet (Hoffart et al., 2013).
- GeoNames, a free geographical database containing all countries and over 11 million place names (Wick, 2015).
- NUTS (Nomenclature of Territorial Units for Statistics), the RDF version of the classification defined by Eurostat office (Correndo & Shadbolt, 2013).

The domain-specific knowledge in the KRC is obtained from expert end-users and according to the opportunities identified by the technical partners for enriching the information managed by each component in the system.

A general methodology for knowledge acquisition has been developed to minimise the burden of LEAs, who are not expected to be able to directly formalize their expertise into the knowledge base. Instead, a simple knowledge acquisition process encompassing the following steps has been carried out:

1. Identification of knowledge sources.
2. Summarisation into a document, which is circulated among involved partners.
3. Formalisation into the knowledge base.
4. If not finished, go back to point 2. To refine the base document, additional requests can be placed to the LEAs; for example, pointers to external knowledge bases, case reports, informal taxonomies, and so forth.

Figure 2 depicts the firearms taxonomy for the firearms trafficking use case. The firearms taxonomy has been built with the support of the LEAs. Specifically, we create a knowledge model that will define firearms features concerning national laws to automatically identify the category and the regulations associated with a firearm in different countries.

5 | KNOWLEDGE DISCOVERY COMPONENT

This section is devoted to explaining the knowledge discovery module. As already mentioned in the system architecture, different methods, such as classification, clustering, association rules, etc., can be used in this module. In this work, we focus on a non-supervised automatic tool that discovers tendencies and relationships among the different objects and attributes that may appear in a data collection. In particular, this tool is

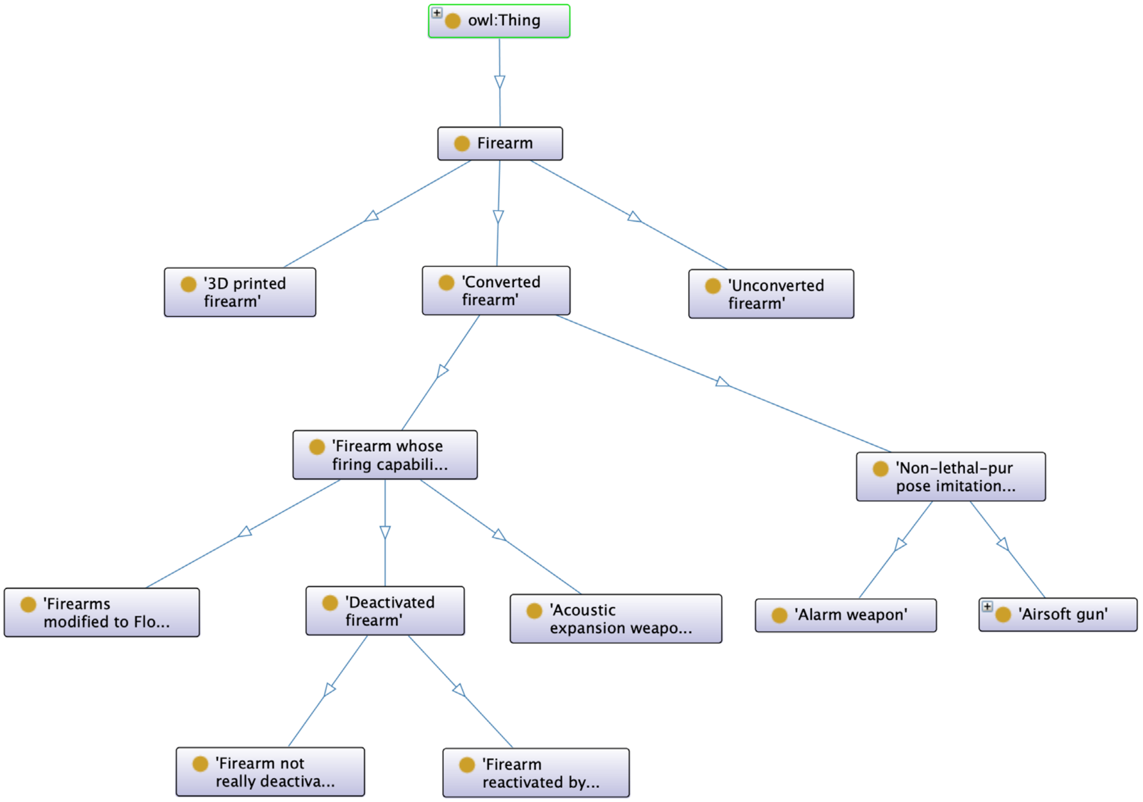

owl:Thing
Firearm
'3D printed firearm'
'Converted firearm'
'Unconverted firearm'
'Firearm whose firing capabili...
'Non-lethal-pur pose imitation...
'Firearms modified to Flo...
'Deactivated firearm'
'Acoustic expansion weapo...
'Alarm weapon'
'Airsoft gun'
'Firearm not really deactiva...
'Firearm reactivated by...

FIGURE 2 Summarised view of the firearms taxonomy.

described using, for exemplary purposes, a database example of a fictitious dataset consisting of advertisements from the dark net offering items related to firearms.

5.1 | Frequent itemsets and association rules discovery

Association rules have been employed to discover meaningful and easy-to-interpret information that can be utilised for different scopes within the crime field (Cheng et al., 2019; Englin, 2015; Ruiz et al., 2014). This component aims to find the most frequent items in a structured dataset and some relationships between these items, measuring their frequency and accuracy.

Formally, for a set of items $I = \{i_1, i_2, ..., i_n\}$ and a set of transactions $D = \{t_1, t_2, ..., t_N\}$, where each transaction may contain or not some of the items, *an association rule* is defined as the relation between two disjoint ($X \cap Y = \emptyset$) itemsets $X, Y \subseteq I$ and is noted by $X \rightarrow Y$.

The itemset X is often referred as the antecedent (or left-hand side of the rule) and Y as the consequent (or right-hand side of the rule). The most commonly used measures to extract frequent itemsets and association rules are the support and the confidence, defined as follows:

- The *support* measures the frequency of appearance of an itemset in the database.

$$Supp_D(X) = \frac{|t_i \in D : X \subseteq t_i|}{|D|}. \tag{1}$$

In particular, the *support of an association rule* is the support of the union of itemsets X and Y:

$$Supp_D(X \rightarrow Y) = Supp_D(X \cup Y) = \frac{|t_i \in D : (X \cup Y) \subseteq t_i|}{|D|}. \tag{2}$$

In general, the most interesting association rules are those with a high support value.

- The confidence of the rule $X \rightarrow Y$ measures the percentage of transactions that containing X, also contain Y. This is measured by the conditional probability of Y given X as follows:

$$Conf_D(X \rightarrow Y) = \frac{Supp_D(X \cup Y)}{Supp_D(X)}. \tag{3}$$

For example, in the case of having a database like the one depicted in 3 where transactions (rows) represent advertisements found in the dark net selling something related with a firearm (e.g. gun, ammunition, etc.) and columns are different types of attributes (e.g. firearm, location, price, nickname, webpage, etc.); the items, in this case, will be pairs of the form $< attribute,\ value >$ like for instance $< location,\ London >$ or $< market,\ Alphabay >$.

In this way, in this kind of dataset, the extracted frequent itemsets and association rules will be relations among the locations and the arms, the day of the week, the market and so forth. For instance, if the following frequent itemsets were obtained:

$$\{ < firearm, Crossbow\_80Lbs > ,\ < location, Madrid > \}$$

$$\text{with} Supp = 0.13$$

this means that *Crossbow*_80*Lbs* and *Madrid* appear together in the 13% of transactions, that is, in the 13% of collected advertisements. If the following association rule were obtained:

$$< firearm, Crossbow\_80Lbs >$$

$$\rightarrow$$

$$< location, Madrid > ,\ < market, Nucleus >$$

$$\text{with} Supp = 0.11, Conf = 0.82$$

that means that when *Crossbow*_80*Lbs* appears in a transaction, it is more likely (82%) that the location of the arm is *Madrid* and it is sold in the *Nucleus* market, having the 11% of transactions supporting this (i.e., satisfying the three items at the same time).

The problem of uncovering association rules (see Figure 4) is usually developed in two steps Agrawal et al. (1994):

- **Step 1: Frequent Itemset Mining (FIM)**. Finding all the itemsets above the minimum support threshold, called *MinSupp*. These itemsets are known as frequent itemsets.
- **Step 2: Association Rule Extraction (ARE)**. Using the frequent itemsets, association rules are discovered by imposing a minimum threshold for an assessment measure, such as confidence, called *MinConf*.

In these steps, some pruning strategies can be applied in order to reduce the complexity of the algorithm. However, these strategies are not enough when the data to be processed scales in an exponential way. This usually happens when analysing social media data searching, for instance, for new insights related to some criminal activity. Therefore, there is a growing need to analyse these very large data sets, which with traditional association rule mining algorithms, often lead to memory overflow errors or extend the processing up to several days.

In the proposed system, we have employed new implementations based on the MapReduce paradigm to extract frequent itemsets and association rules in a more efficient way. To achieve this, this component has been developed using the Spark framework, which enables a distributed computation of data and avoids the memory overflow problems of classic frequent itemset and association rule algorithms available in the literature (see Fernandez-Basso et al., 2023; Fernandez-Basso et al., 2016; Fernandez-Basso, Ruiz, et al., 2019 for more details). In particular, we have used the Apriori-TID proposal using MapReduce functions in Spark This algorithm enables in-memory computations and obtains a complete set of association rules very fast (more details about the advantages of this algorithm on a MapReduce paradigm can be found in Fernandez-Basso et al. (2023)).

Once the association rules are obtained, they are stored and displayed using the following visualisation module.

5.1.1 | Visualisation of results

This component also incorporates a module for storing and visualising the obtained results. This module transforms the obtained frequent itemsets and association rules into an intermediate form (Fernandez-Basso, Ruiz, et al., 2019). For this, we proposed a custom JSON representation of association rules through which the results can be visualised using a wide spectrum of libraries and methods, including graphical and interactive matrices or graphs. This module is of great importance when end users have to inspect the results, thus facilitating the review process to find new knowledge that may be of interest in the case under study. Some of these visualisations are explained in the following use case.

6 | USE CASE: FIREARMS TRAFFICKING IN DARK NET ADVERTISEMENTS

This section contains an overview of the system's performance through its different components, showing how it works in a real use case. In particular, we describe an example that analyses dark net advertisements offering firearms, their components or ammunition for sale. It is worth mentioning that sensible data that cannot assure privacy rights have been conveniently removed or anonymised to comply with the General Data Protection Regulation.

In Figure 5, we have depicted a possible workflow that can be followed in our system. This flow is in line with the proposed architecture (Section 3) for the system. In our particular case, we are going to focus on the two presented components that are applied to a set of dark net processed advertisements. For instance, in Figure 6, there is an example of the data type that can be conveniently extracted from an advertisement about a firearm. To illustrate the performance of the components, we will use a pre-formatted data set that fulfils the specifications of our problem, whose structure is explained in the next section. However, it should be noted that the Knowledge Repository Component can be conveniently employed to enhance the NLP processing, for example, by feeding the classification algorithms for named entity extraction with new terms or slang employed for firearms, neighbourhoods or geographical places, or any other type of knowledge that may exist in the Knowledge Repository. Although this possibility exists, we are going to describe other applications of the KRC to enrich the information obtained after a knowledge discovery process. An example of the pre-processed and transformed database can be seen in Figure 3.

6.1 | Data

The functionality of the different components will be described using a dataset called “Grams” which is a structured dataset containing a collection of advertisements from various dark net markets. It comprises information about the market name (i.e., the name of the market where the advertisement was posted), the vendor name (i.e., the username of the user who posted the advertisement), the name of the item being sold (usually the title of the advertisement), the country from which the advertised item will be shipped, the time at which the advertisement was published, some keywords, some columns containing internal identifiers and information obtained by applying automatic classification techniques to classify the type of firearm offered in the ad Heistracher et al. (2020). An example of the information contained in each of the rows of the dataset can be seen in Figure 3.

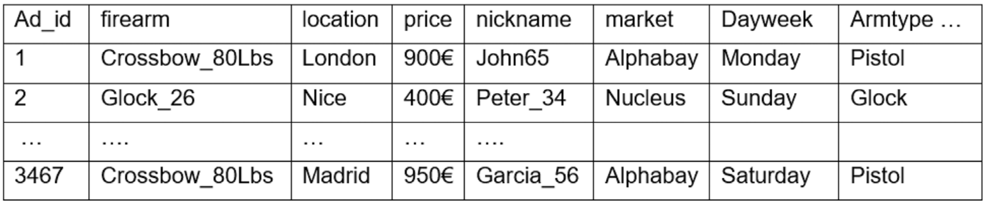
We have also conveniently disaggregated the time attribute (when the ad was posted) into more meaningful attributes giving the month, day of the week, and whether it was posted during the day or at night. At the end, the resulted database has the following attributes (columns) which are used in the subsequent Knowledge Discovery process: Market, Vendor, SoldItem, ShipFrom, Keywords, ArmType, DateTime, Category, Month, DayWeek and DayTime.

Additionally, text fields such as description have been processed by extracting keywords. These terms and other characteristics such as location, were enriched using the Knowledge Repository component in order to have the same granularity level in all the fields of the location-related columns such as ShipFrom.

Finally, the database has to be transformed into a transactional dataset in order to apply the Frequent Itemset and Association Rule mining component. For this purpose, transactions with items using the format *attribute_value* were created in each of the fields.

6.2 | Application of the frequent itemset and association rule component

Following the workflow, the component for the extraction of frequent itemsets and association rules 2019 has been applied. This component allows the extraction of hidden relationships in the data, for which a database of transactions obtained from the processes explained above must first be obtained. The example of weapons advertising data in the dark web is followed by an example to illustrate the usability and understanding of the component.



Ad_id	firearm	location	price	nickname	market	Dayweek	Armtype …
1	Crossbow_80Lbs	London	900€	John65	Alphabay	Monday	Pistol
2	Glock_26	Nice	400€	Peter_34	Nucleus	Sunday	Glock
…	….	…	…	….			
3467	Crossbow_80Lbs	Madrid	950€	Garcia_56	Alphabay	Saturday	Pistol

FIGURE 3 Example of a transactional database with some advertisements.

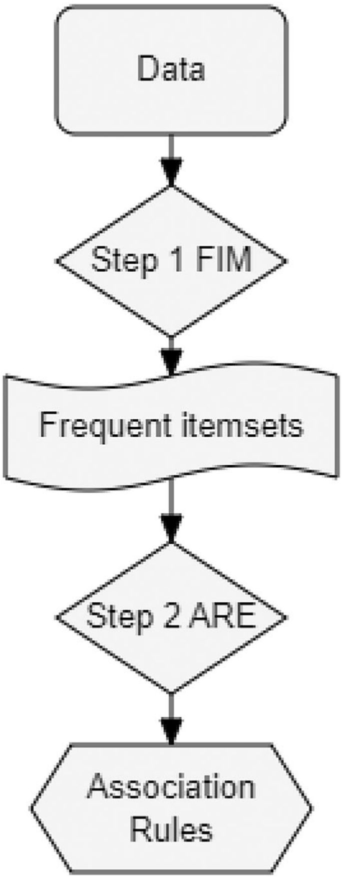


FIGURE 4 Association rule mining process.

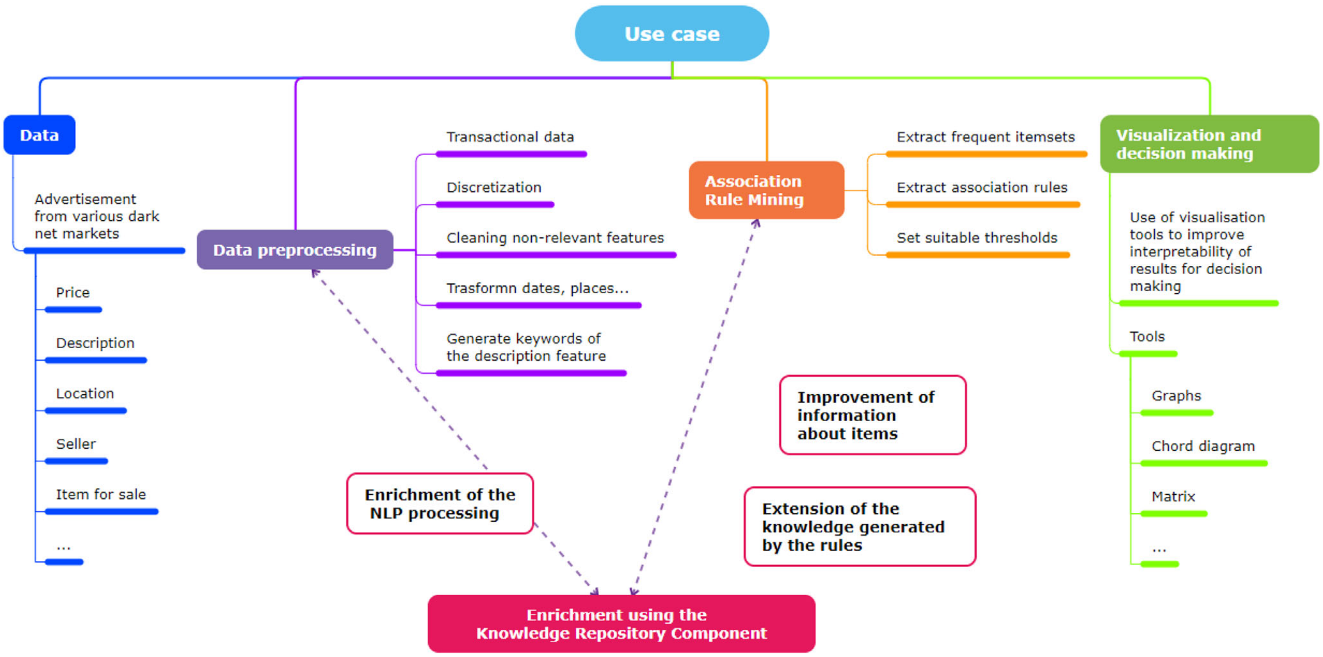
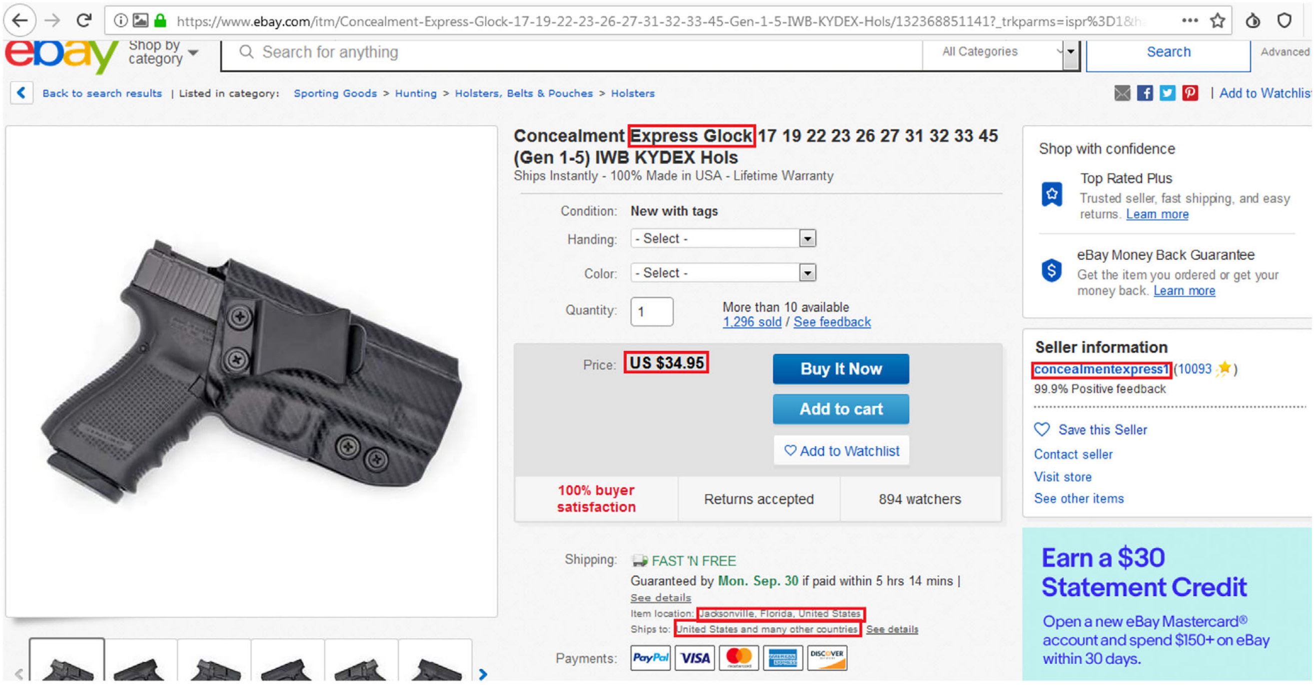

Use case
Data
Advertisement from various dark net markets
Price
Description
Location
Seller
Item for sale
...
Data preprocessing
Transactional data
Discretization
Cleaning non-relevant features
Trasformn dates, places...
Generate keywords of the description feature
Association Rule Mining
Extract frequent itemsets
Extract association rules
Set suitable thresholds
Visualization and decision making
Use of visualisation tools to improve interpretability of results for decision making
Tools
Graphs
Chord diagram
Matrix
...
Improvement of information about items
Extension of the knowledge generated by the rules
Enrichment of the NLP processing
Enrichment using the Knowledge Repository Component

FIGURE 5 Use case workflow.



https://www.ebay.com/itm/Concealment-Express-Glock-17-19-22-23-26-27-31-32-33-45-Gen-1-5-IWB-KYDEX-Hols/132368851141?_trkparms=ispr%3D1&h

Back to search results | Listed in category: Sporting Goods > Hunting > Holsters, Belts & Pouches > Holsters

Concealment Express Glock 17 19 22 23 26 27 31 32 33 45 (Gen 1-5) IWB KYDEX Hols

Ships Instantly - 100% Made in USA - Lifetime Warranty

Condition: New with tags

Handing: - Select -

Color: - Select -

Quantity: 1 More than 10 available 1,296 sold / See feedback

Price: US $34.95

100% buyer satisfaction | Returns accepted | 894 watchers

Shipping: FAST 'N FREE

Guaranteed by Mon. Sep. 30 if paid within 5 hrs 14 mins | See details

Item location: Jacksonville, Florida, United States

Ships to: United States and many other countries See details

Seller information

concealmentexpress1 (10093 ★)

99.9% Positive feedback

Shop by category
Search for anything
All Categories
Search
Advanced
Add to Watchlist
Buy It Now
Add to cart
Add to Watchlist
Payments:
PayPal
VISA
DISCOVER
Shop with confidence
Top Rated Plus
Trusted seller, fast shipping, and easy returns. Learn more
eBay Money Back Guarantee
Get the item you ordered or get your money back. Learn more
Save this Seller
Contact seller
Visit store
See other items
Earn a $30 Statement Credit
Open a new eBay Mastercard® account and spend $150+ on eBay within 30 days.

FIGURE 6 Example of advertisement selling a firearm.

In this component, we will proceed to the application of the algorithm described in Section 5.1. As explained, we will first extract the frequent items, for which we will make use of the measure of interest called support, and the user will determine the minimum value for the threshold *MinSupp*. This measure indicates the percentage of transactions that contain the given item or itemsets.

From these extracted frequent itemsets, we can proceed to the second step of the algorithm, the extraction of association rules 2023. For this purpose, we will use the frequent elements extracted in the previous step and the internal measure named confidence. Again, the user will establish the minimum value for the threshold *MinConf*.

The results extracted from the processed transactional data from *Grams* (see Section 6.1) can be seen below. In the execution of this use case, we have experimented with different thresholds for *MinSupp* and *MinConf* in order to find the parameters that allow extraction of association rules and frequent itemsets that can be of interest to end users. Interesting relationships between the 11 features of the dataset have been obtained for a minimum support value of 0.1 and minimum confidence equal to 0.5. In this case, 62 association rules were obtained.

In general, the FIS/ARD component will find many relationships in a given dataset which can be overwhelming for the analyst using the tool. The component has a visualisation module to find better, comprehend and interpret the results (for end users). In this module, specific graphical methods have been developed to facilitate visual analysis, using the insights provided by LEAs. Depending on the task at hand, the analyst can be interested to focus his or her attention on the relationships with certain certainty and important characteristics:

- Being able to focus his or her attention on the relationships involving certain attributes. This is particularly interesting when looking for patterns that can be used for ulterior situation assessment.
- Being able to restrict to the most frequent/confident itemsets or relationships. This is interesting to find trends or spots about different criminal activities.
- Being able to reason using a subset of discovered relationships. This is particularly interesting when the analyst is trying to understand the underlying phenomena that cause the discovered relationships.

We are going to highlight some of the visualisation graphs that can be interesting for the project. The first type of visualisation aims at providing an overview of the relationships in terms of frequency (support) and importance (confidence). With this visualisation, the analyst would be able to notice and localise groups of discovered rules by confidence and support. For that, the system offers different types of visualisations like the one in Figure 7.

In this example, we can find the most frequent rules located at the right part of the graph and the most confident located at the top part of the graph. Additionally, the intensity of the colour indicates higher confidence. This graph is interactive and shows the rule when the mouse is positioned over the point. For instance, in Figure 7, we can see an association rule that relates Mondays and advertisements posted during the day (defined as ads posted between 7:00 and 22:00).

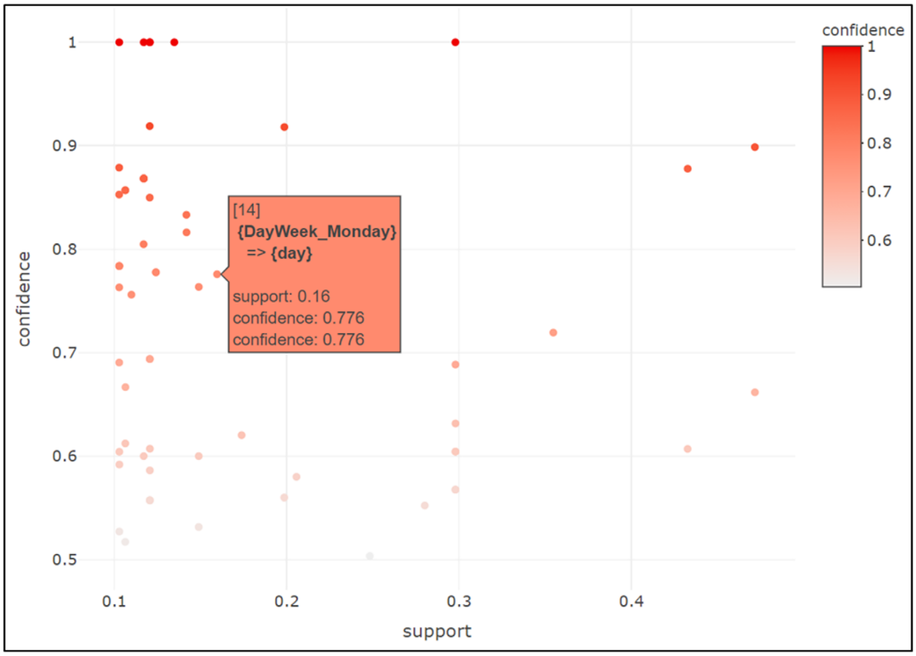


FIGURE 7 Example of association rule found using a 2D-graphical interactive visualisation.

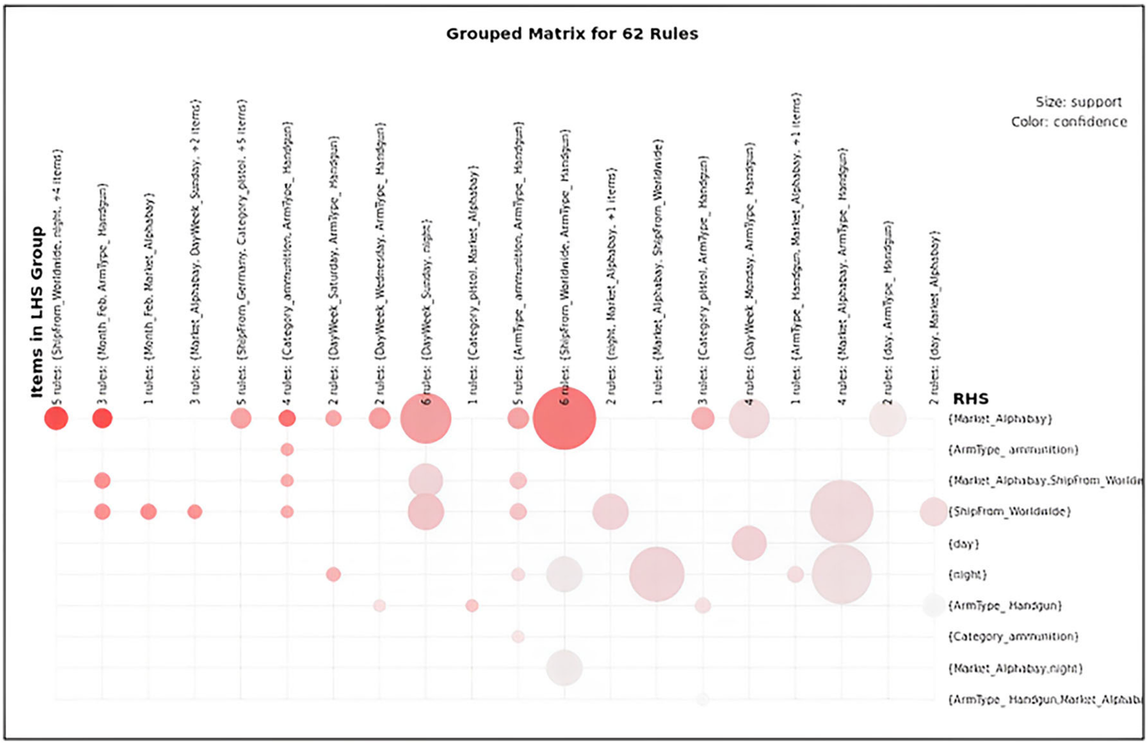
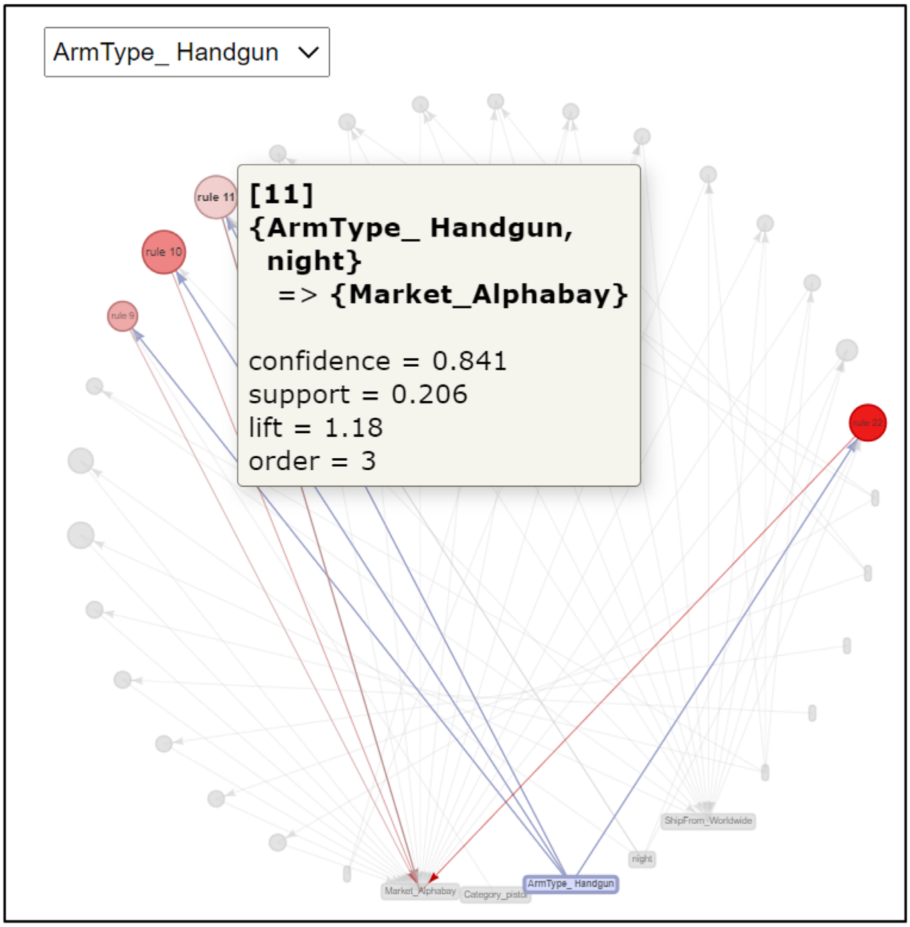


FIGURE 8 Example of association rule found using a 2D-graphical group visualisation.

In Figure 8, we can see the plot of the rules with the highest support and confidence. We can see that there are rules with high support that relate Sunday nights to the Alphabay market (up to 6 rules with a high confidence level, red colour). Other rules that can be seen in Figure 8 are the high support rules, where we can see how the ammunition rules relate to shipping from anywhere in the world and at night.

The second type of visualisation (Figure 9) makes it easier to search for relationships involving specific attributes, that is, domain concepts can be selected in the top left drop-down menu. This allows the analyst to select an attribute/item of interest, and the tool will provide a graphical overview of the attributes strongly related to the selected one, eliminating in the visualisation those less relevant attributes. For instance, in

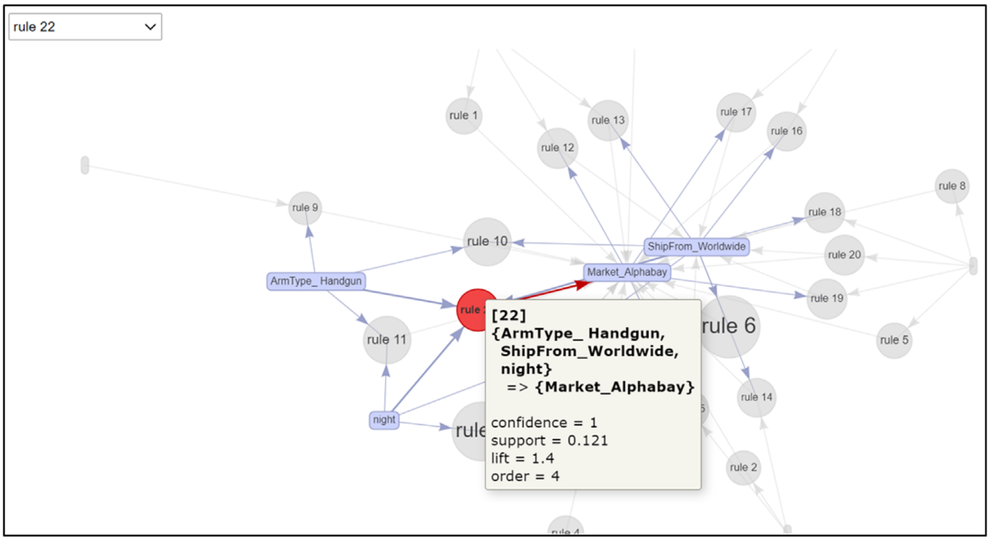


ArmType_ Handgun

[11]
{ArmType_ Handgun,
 night}
  => {Market_Alphabay}

confidence = 0.841
support = 0.206
lift = 1.18
order = 3

FIGURE 9 Example of association rule that was found using an interactive chord diagram.



rule 22

[22]
{ArmType_ Handgun,
 ShipFrom_Worldwide,
 night}
  => {Market_Alphabay}

confidence = 1
support = 0.121
lift = 1.4
order = 4

FIGURE 10 Example of visualisation of rules obtained in an experiment where rule 22 is highlighted joint with its connections.

Figure 9, we can find another interesting example of association rule, where we can see that in those advertisements posted at night selling a Handgun arm type tend has been posted in the Alphabay market.

The third type of visualisation we want to highlight provides an overview of relationships similar to that in Figure 9, but in this case, the drop-down menu enables the selection of an association rule of interest, and the graph shows all the association rules that are related to some of the

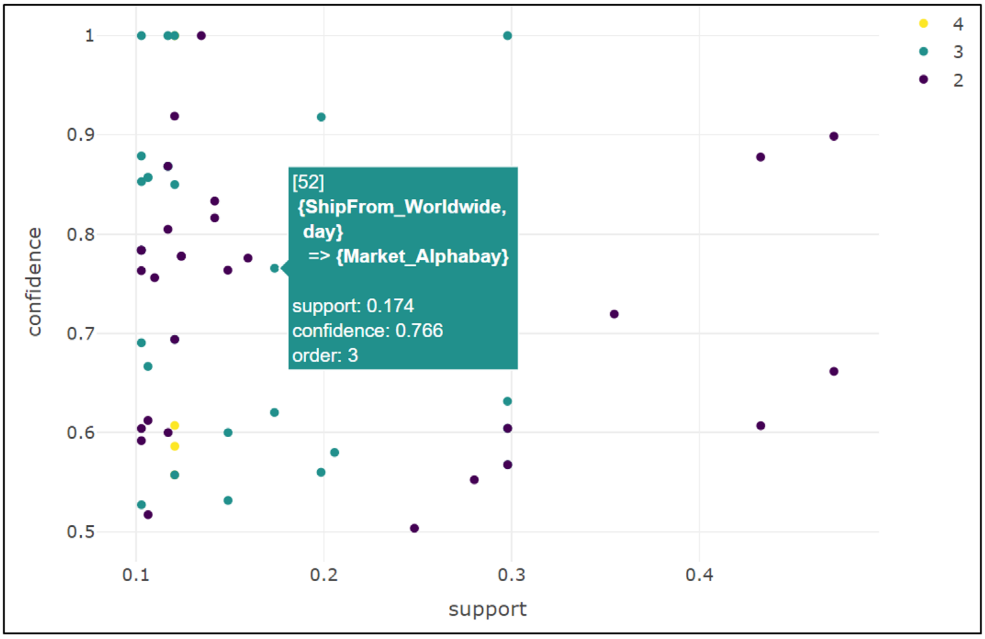


FIGURE 11 Example of association rule found using a 2D interactive graph visualisation.

attributes of the selected association rules (see Figure 10). If no rule is selected, the entire set of rules is displayed. In this case, the analyst can therefore discover chain of relationships that helps him/her reasoning about possible causes of relationships.

The last type of visualisation we show here provides an overview of relationships similar to that in Figure 7, but in this case, the colour represents the number of items of the association rule. With this type of visualisation, the user can see at a glance which associations involving more items but also have high support and confidence (see Figure 11. This can be interesting to find those associations that relate different itemsets in order to analyse the relationship among different groups of items. In this particular figure, we can find the association rule that relates Worldwide international shipping with ads posted at night and the Alphabay market.

6.3 | Enrichment using the knowledge repository component

In addition to the application of enrichment in data pre-processing and transformation, this enrichment process can be applied to the results obtained by the knowledge discovery tool. To better explain this process, we will use an example of an association rule obtained from the dark web advertisement dataset.

For instance, for association rules obtained like this one:

$$< firearm\_3Dprinted, Market\_alphabay > \rightarrow < night > .$$

The user may be interested in knowing if there is any legislation regarding the sales of 3D printed firearms in a certain country. For this, the Knowledge Repository component can be of interest to consult that, in order to research into this type of advertising. In this particular case, the KRC should be populated with such type of information.

Furthermore, as the KRC provides a firearm taxonomy, the user can be interested in focusing the analysis on those firearms within a particular category, like the different types of converted firearms. For that, the results obtained after the Association Rule Mining component can be processed to cluster those rules and frequent itemsets that contain items pertaining to that category, such as the following association rules:

$$\begin{gathered}< firearm\_modified\_shotgun, Market\_alphabay > \rightarrow < night > ,\\ \vdots \\ < firearm\_deactivated\_glock > \rightarrow < shipFrom\_WorldWide > .\end{gathered}$$

In particular, the Association Rule mining process can also be modified by changing the transactional input database by considering a different granularisation level (according to a category in the KRC) in some of the attributes (columns of the dataset). For instance, the type of arm appearing in a transaction could be generalized to its category, obtaining, for instance, an association rule like:

$$< \mathit{firearm\_handgun} > \rightarrow < \mathit{shipFrom\_WorldWide} > \tag{4}$$

instead of finding

$$< \mathit{firearm\_glock26} > \rightarrow < \mathit{shipFrom\_WorldWide} > . \tag{5}$$

Or a rule like:

$$< \mathit{firearm\_assault\_rifle} > \rightarrow < \mathit{Market\_alphabay} > \tag{6}$$

instead of finding

$$< \mathit{firearm\_AK47} > \rightarrow < \mathit{Market\_alphabay} > . \tag{7}$$

Additionally, this change in the hierarchy or the granularisation may help to find new associations that will not be found with the original data, for example, because the minimum support is not achieved. But merging different firearms into the same category will help to have higher support, and therefore the association rule like the one in (4) will emerge whilst the rule in (5) will not because there are not enough transactions to exceed the *MinSupp* threshold.

7 | CONCLUSIONS

This article presents an artificial intelligence application to assist Law Enforcement Agencies during their investigations. Specifically, the paper has focused on two different components that offer many possibilities to enhance the research process during a criminal case.

To illustrate the functioning of the system, we have described different possibilities using a use case based on the analysis of advertisements published on the Darknet. Additionally, we have described different examples for visualising the results obtained by the FIS/ARD tool and combining the information stored in the KRC to enrich the LEAs research process. Needless to say, the data used during the development of the system have been conveniently anonymised and processed in order to comply with the General Data Protection Regulation. The use case was demonstrated with real-life data and validated by several LEAs who gave positive feedback on the use and outcome of the system. In addition, the innovative use of the knowledge base has yielded very positive results.

The developed system can be of interest not only to support LEAs in their investigation but also in other scenarios where knowledge enrichment plays an important role in enhancing the obtained results of a Data Mining or Machine Learning algorithm.

ACKNOWLEDGEMENTS

This research paper is part of the COPKIT project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 786687 and European Union (BAG-INTEL, 101121309. This paper was also supported by grant PID2021-123960OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe and by grant TED2021-1289402B-C21 founded by MCIN/AEI/10.13039/501100011033 and, by the European Union NextGenerationEU/PRTR. In addition, this research has been partially supported by the Ministry of Universities through the EU-funded Margarita Salas programme NextGenerationEU.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Carlos Fernandez-Basso https://orcid.org/0000-0002-8809-8676
Karel Gutiérrez-Batista https://orcid.org/0000-0003-2711-4625
Juan Gómez-Romero https://orcid.org/0000-0003-0439-3692

M. Dolores Ruiz https://orcid.org/0000-0003-1077-3173
Maria J. Martin-Bautista https://orcid.org/0000-0002-6973-477X

ENDNOTES

[1] https://www.europol.europa.eu/socta-report.

[2] https://copkit.eu/.

[3] Grams data set is publicly available at https://www.gwern.net/DNM-archives#grams.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases* (Vol. 1215, pp. 487–499). VLDB.

AIDA. (2020). Advanced european infrastructures for detectors at accelerators. https://www.project-aida.eu/

ANITA. (2021). Advanced tools for fighting online illegal trafficking. https://www.anita-project.eu/

Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2012). Prov-o: The prov ontology. *Tech. rep*.

CC-DRIVER. (2023). Understanding the drivers of cybercriminality, and new methods to prevent, investigate and mitigate cybercriminal behaviour a research. https://www.ccdriver-h2020.com/

Cheng, B., Li, W., & Tong, H. (2019). Prediction of criminal suspects based on association rules and tag clustering. *Journal of Software Engineering and Applications*, *12*, 35–50.

COPKIT. (2021). Technology, training and knowledge for early-warning/early-action led policing in fighting organised crime and terrorism. https://copkit.eu/

Correndo, G., & Shadbolt, N. (2013). Linked nomenclature of territorial units for statistics. *Semantic Web*, *4*, 251–256.

Englin, R. (2015). *Indirect association rule mining for crime data analysis*. Ph.D. thesis. EWU Masters Thesis Collection http://dc.ewu.edu/theses/331

Fernandez-Basso, C., Francisco-Agra, A. J., Martin-Bautista, M. J., & Ruiz, M. D. (2019). Finding tendencies in streaming data using big data frequent itemset mining. *Knowledge-Based Systems*, *163*, 666–674.

Fernandez-Basso, C., Ruiz, M. D., Delgado, M., & Martin-Bautista, M. J. (2019). A comparative analysis of tools for visualizing association rules: A proposal for visualising fuzzy association rules. In *Proceedings of the 11th Conf. of the European Society for Fuzzy Logic and Technology, EUSFLAT 2019, Prague, Czech Republic, September 9-13, 2019, vol. 1 of Atlantis Studies in Uncertainty Modelling* (pp. 520–527). Atlantis Press.

Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2016). Extraction of association rules using big data technologies. *International Journal of Design & Nature and Ecodynamics*, *11*, 178–185.

Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2023). New spark solutions for distributed frequent itemset and association rule mining algorithms. *Cluster Computing*.

Graua, B. C., Horrocksa, I., Motika, B., Parsiab, B., Patel-Schneiderc, P., & Sattlerb, U. (2008). Web semantics: Science, services and agents on the world wide web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*, 309–322.

Griol-Barres, I., Milla, S., Cebrián, A., Fan, H., & Millet, J. (2020). Detecting weak signals of the future: A system implementation based on text mining and natural language processing. *Sustainability*, *12*, 7848. https://www.mdpi.com/2071-1050/12/19/7848

Harris, S., & Seaborne, A. (2013). SPARQL 1.1 Query Language.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*, 139–154.

Heistracher, C., Mignet, F., & Schlarb, S. (2020). Machine learning techniques for the classification of product descriptions from darknet marketplaces. In G. Kovásznai, I. Fazekas, & T. Tómács (Eds.), *Proceedings of the 11th Int. Conf. on Applied Informatics (ICAI 2020), Eger, Hungary, January 29-31, 2020* (Vol. 2650, pp. 128–137). Springer. CEUR-WS.org

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, *194*, 28–61.

i LEAD project. (2023). Innovation–law enforcement agencies dialogue. https://i-lead.eu/

Klyne, G., & Carroll, J. J. (2004). *Resource description framework (RDF): Concepts and abstract syntax*. World Wide Web Consortium.

Li, J. (2021). Threats and data trading detection methods in the dark web. In *In 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA)* (pp. 1–9). IEEE.

MAGNETO. (2021). Multimedia analysis and correlation engine for organised crime prevention and investigation. http://www.magneto-h2020.eu/

Manola, F., Miller, E., & McBride, B. (2014). RDF 1.1 Primer. https://www.w3.org/TR/rdf11-primer/

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2021). Darknetexplorer (DNE): Exploring dark multi-layer networks beyond the resolution limit. *Decision Support Systems*, *146*, 113537.

Qayyum, S., & Dar, H. (2018). A survey of data mining techniques for crime detection. *University of Sindh Journal of Information and Communication Technology*, *2*, 1–6.

Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R. N., & Ghosh, A. (2021). Analysis of darknet traffic for criminal activities detection using tf-idf and light gradient boosted machine learning algorithm. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021* (pp. 671–681). Springer.

Ruiz, M. D., Martin-Bautista, M. J., Sánchez, D., Vila, M. A., & Delgado, M. (2014). Anomaly detection using fuzzy association rules. *International Journal of Electronic Security and Digital Forensics*, *6*, 25–37.

TENSOR. (2019). Retrieval and analysis of heterogeneous online content for terrorist activity recognition. https://tensor-project.eu/

TITANIUM. (2020). Tools for the investigation of transactions in underground markets. https://titanium-project.eu/

W3C OWL Working Group. (2012). *OWL 2 Web Ontology Language Document Overview* (2nd ed.). W3Standards. https://www.w3.org/TR/owl2-overview/

Wick, M. (2015). Geonames ontology.

AUTHOR BIOGRAPHIES

Carlos Fernandez-Basso received the degree in computer science, the M.Sc. degree in data science, and the Ph.D. degree in computer science from the University of Granada, Granada, Spain, in 2014, 2015, and 2020, respectively. He is currently a Postdoctoral Fellow with Causal Cognition Lab, University College London, London, UK. He was a Lead Developer in the EU FP7 Project Energy IN TIME in the topics of building simulation and control, data analytics, and machine learning, and in the COPKIT Project in the topics of cybercrime, Big Data, and machine learning. From 2016 to 2018, he collaborated with the Data Science Institute, Imperial College London, London, UK, where he has carried out. He is now a postdoctoral researcher at University College London applying artificial intelligence in social contexts and researching explain ability in artificial intelligence (XAI).

Karel Gutiérrez Batista was born in 1984. He received a degree in computer science and M.Sc. degree in data science from the University of Camagüey, Cuba. He received his PhD in computer science in 2018 from the University of Granada, Spain. He works as a postdoc fellow at the Department of Computer Science at the University of Granada. He is an Associated Research of Intelligent Data Bases and Information Systems (IDBIS) research group at the University of Granada. His research interests are Multidimensional Data Analysis, Deep Learning, Data Mining, Knowledge graphs, and Natural Language Processing.

Juan Gómez-Romero received the B.Sc. degree in computer science and the M.Sc. and Ph.D. degrees from the University of Granada, Granada, Spain, in 2004, 2006, and 2008, respectively. He was a Lecturer with the Applied Artificial Intelligence Group, Universidad Carlos III de Madrid, Madrid, Spain, from 2008 to 2013, and a Research Associate in the EU FP7 Project Energy IN TIME with the University of Granada, from 2013 to 2017. Since 2019, he has been an Associate Professor with the Computer Science and Artificial Intelligence Department, Universidad de Granada. His research interests include machine learning for control optimization and simulation of power systems. He is currently the Principal Investigator of the projects PROFICIENT: Deep learning for energy-efficient building control and DeepSim: Deep learning of building simulation models.

M. Dolores Ruiz received the degree in mathematics and the European Ph.D. degree in computer science from the Universidad de Granada, in 2005 and 2010, respectively. She is a Lecturer at the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain, since 2020. She has participated in more than ten projects, including the EU FP7 Projects ePOOLICE and Energy IN TIME, and the COPKIT H2020 project. Her research interests include data mining, information retrieval, energy efficiency, big data, correlation statistical measures, sentence quantification, and fuzzy sets theory. She has organized several special sessions about Data Mining in international conferences and was part of the organization committee of the FQAS'2013, SUM'2017 and FQAS'2023 conferences. She belongs to the Approximate Reasoning and Artificial Intelligence Research Group and the Cybersecurity Lab, Universidad de Granada. She is the Principal Investigator of several projects about federated mining and desinformation detection.

Maria J. Martin-Bautista received the Ph.D. degree. She has been a Full Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, Spain, since 2018. She has supervised several Ph.D. thesis and published more than 100 papers in high impact international journals and conferences. She has participated in more than 20 research and development projects and has supervised several research technology transfers with companies. Her current research interests include recommender systems, intelligent information systems, big data analytics in data, text and web mining, knowledge representation, and uncertainty. She is a member of the Intelligent Data Bases and Information Systems (IDBIS) research group. Furthermore, she has served as a program committee member for several international conferences.

How to cite this article: Fernandez-Basso, C., Gutiérrez-Batista, K., Gómez-Romero, J., Ruiz, M. D., & Martin-Bautista, M. J. (2024). An AI knowledge-based system for police assistance in crime investigation. *Expert Systems*, e13524. https://doi.org/10.1111/exsy.13524