Duo-mining with multi-agents: the next wave to control organised crime

C. Chibelushi[1] & I. Bentley[2]
[1]University of Wolverhampton, UK
[2]Staffordshire Police, UK

Abstract

Text and data mining are fast growing areas and are believed to have high commercial potential value in knowledge discovery and information filtering areas of application. Although text mining manages unstructured data, most of knowledge discovery and information filtering can be done using data mining. Despite that, both technologies do not actively predict and prevent problems, instead they leave the work to the experts to manually interpret the data, anticipate future events and make the final decision. This paper proposes the ASKARI approach outlined in this paper, which combines duo-mining (text and data mining) with multi-agent systems, the approach aims to predict and prevent crimes before they happen, and in fact, might become the next wave of knowledge discovery. Also, the paper highlights the benefits of combining duo-mining and multi-agents in prediction and preventing organised crimes.

Keywords: organised crime, duo-mining, multi-agent systems.

1 Introduction

Organised crime is certainly not a new phenomena but it has become, in recent years, increasingly more widespread and highly sophisticated, taking advantage of the advances in technology, particularly in the field of communication. Organised crime can be defined as a structured or not structured group of two or more people existing for a period of time and acting in concert with the aim of committing one or more serious crimes that are motivated by politics, religion, race or financial gain [1]. Organised crime can include terrorism, drug trafficking, fraud, gang robberies and other group-oriented criminal activities. A terrorist incident is perceived to be significant if it results in loss of life, serious

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080261

injury to persons, and/or major property damage. Instead of discussing about all forms of organised crime, this paper will use terrorism to demonstrate the problem, hoping that by understanding how to control one form, it may help to control other forms of organised crime.

Incidents such as September 11 dramatically demonstrated the need for the world to greatly improve its counterterrorism activities. However, terrorism appear in many forms, its elements seem to be everywhere and difficult to identify, and the more complex aspect of it is that the intentions and plans are even harder to uncover [2]. Most nations, and specifically in Europe and U.S.A, have significant number of institutions working in different departments of counterterrorism. These institutions store large amounts of data to support their efforts. However, they do not share the contents of their database, and are not set up to analyse or even to obtain data to cover full range of activities that take place during the period from the conceptualisation of a terrorist act to its execution. For example 17,000 shipping containers arrive in USA everyday, and only a small percentage of these are searched [2]. Some may contain materials or weapons that can be used to conduct a terrorist attack. Also there is no clear approved concept for an integrated approach to collect and analyse the data required to counter terrorism. This may be exacerbated with the laws which in some cases do not allow communication between different counterterrorism institutions.

Since organised crime depends on two main factors, communication and finance, counterterrorism institutions should use the same factors to control terrorism. However, this may be possible through analysing and linking various sources of data from criminal communications, bank reports, and Modus Operandi held by the Police or law enforcement officials.

Text and data mining are the main technologies that attempt to analyse data to discover interesting patterns such as clusters, associations, deviations, similarities, and differences in sets of text [3]. However, most of this data is in textual form and is unstructured. Hence, may need to be analysed using text mining technique.

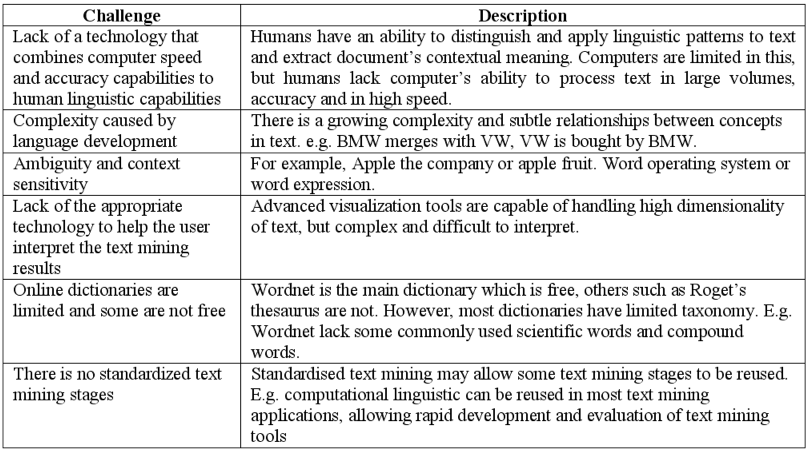
1.1 Text mining and applications

The growth of online scientific literature, coupled with the growing maturity of text processing technology, has increased the importance of text mining as a potential technology for discovery, decision making and problem solving. Text mining has been defined as "the discovery by the computer of new, previously unknown information, by automatically extracting information from different written resources" [4].

Since the early nineties, most of the research focused in the process of automating the knowledge discovery from databases that comprise of comprehendible patterns of structured data. This received credits as been crucial knowledge to support business management. But text mining which works with unstructured textual data has gained popularity and is been forced to expand rapidly due to many factors, among them include: corporate data flood [5–8], national security [9, 10], research and development support. To date most of its expansion is mainly in business management applications and specifically on marketing strategy [11, 12]. This involves applications such as customer relationship management, corporate knowledge, content management from company's intranets or portals, and information retrieval for marketing purposes [5].

Although national security stands as one of the main drivers to text mining expansion, its growth is slow and is mainly in academic domain as compared to business management which has recently been reported to have expanded into the realm of applied information technology [8, 13]. This may have resulted from the lack of adequate technology, few successful experiences reported and lack of adequate methodology to drive users in developing text mining applications [13]. Consequently, text mining is facing several challenges, among them are listed in Table 1. Such challenges limit text mining to an extent that most research development efforts have centred on Data mining efforts using structured data. Text mining is similar to data mining except that data mining tools are designed to handle structured data from databases or XML files, while text mining can work with unstructured or semi-structure data sets such as emails, full-text documents, HTML and other textual documents. Text mining is driven by its applications, and so it brings together techniques from data mining, linguistics, machine learning, information retrieval, pattern recognition, statistics, databases, extraction and visualization to help the user to narrow down a broad range of documents and explore related topics.

Table 1: Text mining challenges.



Challenge	Description
Lack of a technology that combines computer speed and accuracy capabilities to human linguistic capabilities	Humans have an ability to distinguish and apply linguistic patterns to text and extract document's contextual meaning. Computers are limited in this, but humans lack computer's ability to process text in large volumes, accuracy and in high speed.
Complexity caused by language development	There is a growing complexity and subtle relationships between concepts in text. e.g. BMW merges with VW, VW is bought by BMW.
Ambiguity and context sensitivity	For example, Apple the company or apple fruit. Word operating system or word expression.
Lack of the appropriate technology to help the user interpret the text mining results	Advanced visualization tools are capable of handling high dimensionality of text, but complex and difficult to interpret.
Online dictionaries are limited and some are not free	Wordnet is the main dictionary which is free, others such as Roget's thesaurus are not. However, most dictionaries have limited taxonomy. E.g. Wordnet lack some commonly used scientific words and compound words.
There is no standardized text mining stages	Standardised text mining may allow some text mining stages to be reused. E.g. computational linguistic can be reused in most text mining applications, allowing rapid development and evaluation of text mining tools

As a result of the above challenges, the majority of the data and text mining tools (as shown in [14]) which are used for security analysis are for structured data, use data mining, and mainly analyse historical organized crimes data. In other words, these systems support the analysts to deal with what we call '*already done crimes*' and not '*about to be committed crimes*'. Although it is also important to analyse 'the already committed organised crime', but this research main interest is on the '*about to be committed crimes'* as it is at the very interest of most nations to reduce the number of casualties by preventing organised crimes from happening.

However, with the challenges highlighted in Table 1, and the significance of the application presented in this paper, there is a need to find alternative methods that can support the text mining technique to go beyond simply providing patterns that are difficult to interpret. In addition, text mining for national security application is an issue that needs to be given extra effort because this area is trapped in a closed circle. A circle with organised crime needs to be analysed but requires much of data capture from a wide range of sources e.g. emails, telephone conversations, bank transactions, text messages, etc. for the analysts to understand the causes of organised crime to speed up the analysis process without information lag and at the same time be accurate, and effectively make decisions. This is illustrated using Figure 1.

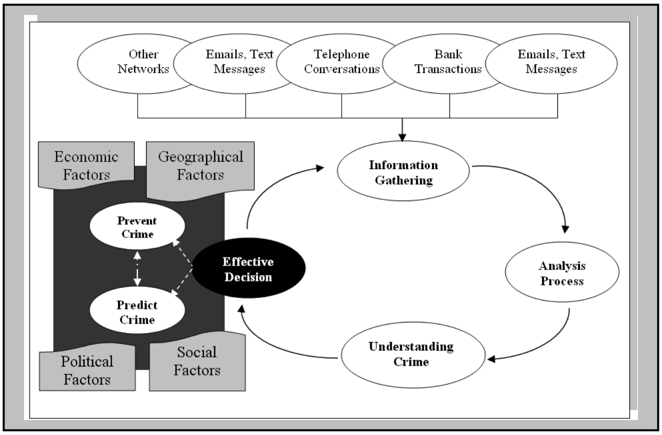


Figure 1: Security area in circle.

It is impossible to break this circle by using distinct items from the circle and ignore other items. Also, the multi-data sources are growing in complexity, variability, and data size ranging from the rate of four petabytes per month captured for crime analysis [14]. In addition, some degree of complexity in decision making may originate from geographical, social, economical and political factors. These factors are not the main focus of this paper, and so they will not be discussed.

Despite the immediate needs to prevent organised crime, most text and data mining systems do not incorporate the predictive modelling facilities in their applications [7, 14], and they are not combining patterns from both text and data mining. Also, the technology leaves the analyst with a considerable amount of work to interpret the output and predict crimes based on historical data (e.g. the work by [15] and [16]) and give little emphasis in developing an environment that can link data across various databases.

Linking data in this manner can help criminal analysts to have a wider perspective on what steps taken by the criminals to organise an unlawful activity, and prepare the law enforcement officials with better tactics to deal with future organised crimes. In addition, the criminal analysts can use the system to learn more about future preparation of criminal activity when they have little clue and speed up the process of making decisions.

An obvious need is to assist the national security analysts to predict and prevent crimes rather than emphasizing on assisting them on only one area; of analysing crimes which have already made damages.

This paper suggests the ASKARI (means policing in Kiswahili) approach that combines Duo-mining and multi-agent technology as a means to assist the national security analysts in crime prediction and prevention processes. Duo-mining integrates data and text mining in a single system to facilitate the analysis of both structured and unstructured criminal data. The multi-agent technology uses the Duo-mining results to predict and alert crime analysts on the 'about to be committed crimes'.

2 The ASKARI architecture

After several terrorist attacks, data and text mining became one of the dominant approaches in an increasing number of research projects associated with organised crime and in particular with anti-terrorist activities. A typical example is the homeland security programmes initiated by Defence Advanced Research Project Agency (DARPA) which combines data fusion, database searches, biometrics and pattern recognition technologies. This programme seeks to develop a network of technologies to help security officers predict and prevent terrorism activity but also limited as it depends on historical data.

The ASKARI approach is a multi-document (from various sources), and content based approach. The approach is based on DecM-Text Mining [17] which stands for Decision management using Text Mining approach. It is a text mining methodology for extracting the elements of decision making from transcripts generated from meetings; its aim is to reduce rework in software development projects. Based on the DecM-Text Mining, a simplified representation of the ASKARI approach is shown in Figure 2.

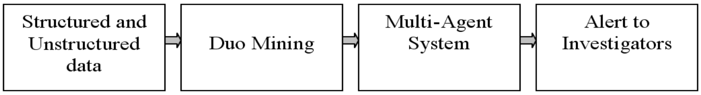


Figure 2: The ASKARI architecture.

2.1 Structured and unstructured data

The process of organising crime leave traces of criminal behaviour all over the place. These traces are beneficial and are crucial in developing patterns which are necessary to understand the type of criminality, the individuals involved, location and other important information as demonstrated in Figure 4. As a result, the ASKARI approach incorporates traces from emails, text messages, telephone conversation transcripts and other textual related sources as an input to be analysed.

2.2 Duo mining

Since text mining is a new field, there is no standardised methodology available in the literature to assist researchers involved in text mining. Also, the ASKARI approach is meant to analyse structured and unstructured data. Consequently, the ASKARI approach adapts the CRISP-DM [18] in its methodology. Figure 3 consists of the main Duo mining phases used by the ASKARI. This consist of data, pre-processing, text analysis, textual data patterns, link analysis, and the output phase which contains criminal patterns from both structured and unstructured data.

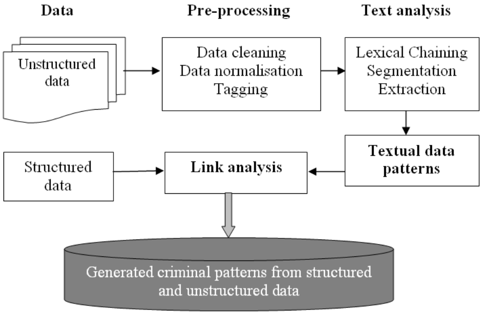


Figure 3: The ASKARI text mining process.

The first phase is discussed in Section 2.1, other phases will be discussed as follows.

2.2.1 Pre-processing

ASKARI pre-processing steps include data cleaning and data normalisation. Data cleaning and normalisation are done on unstructured texture data only. This is because, data such as that from the bank is well structured in the form of a table and contains less information which is not beneficial to the analysis. Data cleaning involves removing signature and non-textual features (such as * and <> - which may appear in transcribed text or SMS) which are not useful to the analysis.

Text normalisation: Text normalisation may engage four steps; tokenisation, case folding, stop words removal and lemmatisation [19]. In English, modifications done by lemmatisation do not normally change the words class but vary its tense or plurality [20]. Recently, the work of [17], [21], and [22] reported that lemmatisation provides no significant change to the processing of textual data, and hence are not used in this project. In most text mining techniques, the stop list comprises frequently occurring words such as 'is', 'or', 'a' and etc. They typically fall into syntactic categories such as 'determiners', 'prepositions', 'conjunctions' and 'auxiliaries'. Such words are usually considered of no interest, as for most tasks these words cannot be used in the differentiation of document classes, and also do little to illuminate the content of documents. Differently, with ASKARI approach, care is taken in eliminating these types of words. This is because; current SMS messages use a specific lexicon in which a word can be represented by one letter, a number or a character e.g. '*a u coming 2nite?*' instead of 'are you coming tonight?' or ' *R U ok? Am busy @ da mo*' instead of 'are you okay? I am busy at the moment'. In order to control this problem, text from SMS messaging is translated into a normal English sentence first and then passed to the pre-processing stage. This is done by comparing characters or letters with the manually developed SMS messaging dictionary. Ignoring the words such as 'r' or '2nite' from the data may result in an inaccurate analysis. The challenge lies mainly on the different versions of these words as they are used differently with different communities. SMS lexicon is advancing fast but it is not yet standardised.

2.2.2 Tagging

Tagging is done using a tool called Wmatrix [23] which is an online corpus analysis and comparison system that provides a variety of tools for NLP. Included among these tools is CLAWS part-of-speech tagging software, SEMTAG (word-sense tagger) and LEMMINGS (a lemmatiser) as well as statistical functions such as frequency lists and other facilities. The output from the tagger is XML formatted data.

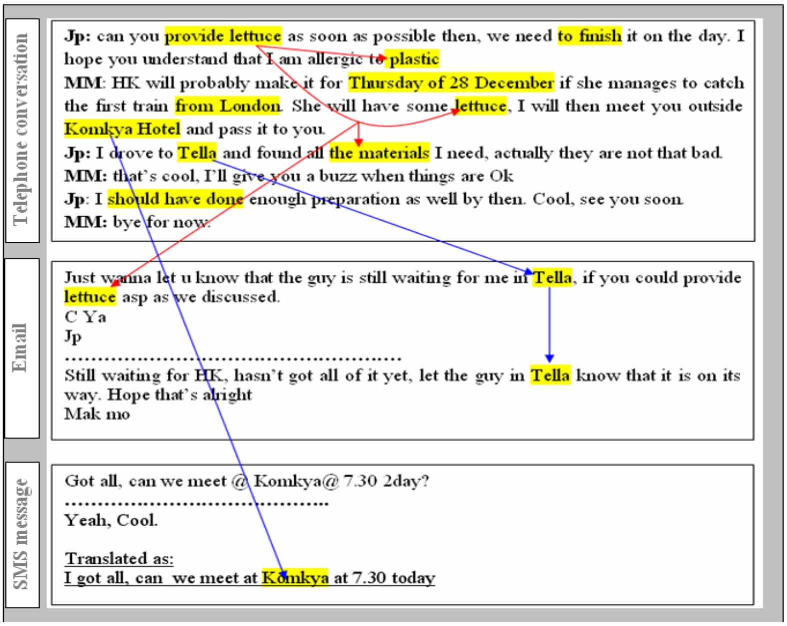
2.2.3 Text analysis

This phase involves lexical chaining construction, segmentation and extraction processes. Each process is expanded as follows.

Lexical chaining implements feature clustering, i.e. words are clustered depending on the semantic relationship (also known as senses) between them, and hence the analysis is contextual. The ASKARI opted to use lexical chaining technique because this technique has an advantage of being less complex and is context-based. Morris and Hirst [24] introduced the lexical chaining approach, since then many computational linguists have used lexical chains in a variety of tasks [17]. The ASKARI Lexical chaining approach use 5 senses from the Wordnet to construct the chains. These are repetition, synonym, hypernymy, meronymy and coordinate terms.

However, Wordnet taxonomy is limited as it does not contain compound words, Hirberno-English phrases (such as '*drugstore*' which is a sense of chemist, '*banjax*' – to break, '*jacks*' – toilet, '*wain*' – a child, and '*bold*' – badly behaved), and criminal argot. Criminal argot is commonly used to organise crime, hence it will further be discussed in this section. Criminal argot is the language used by criminals to communicate and understand themselves and not anybody else who is not within their community.

Nevertheless, research shows that when these criminals get used to this language, they tend to use it unconsciously and openly, leaving traces of their criminal behaviour in public [9]. This has helped the ASKARI approach to manually start to populate a database of words which belong to criminal argot. This may in future be developed to a criminal argot dictionary. Currently, the majority of the words in this database are related to drug(Criminal argot can be accessed from http://argot.com/) activities, but some will in future be related to different types of organised crime. Some examples of criminal argot commonly used in organising crimes [14], some of which are demonstrated in the fictitious data of Figure 4. The development and the use of criminal argot dictionary in the analysis of organised crime is crucial, as it will enhance the accuracy of criminal data analysis. For example the word ‘lettuce’ in Figure 4 represent cash, as it can be seen, the actor (denoted by ‘Jp’) insists on receiving cash by saying that he is ‘allergic to plastics’, meaning that he/she will not want credit cards to be used for payments.



Jp: can you provide lettuce as soon as possible then, we need to finish it on the day. I hope you understand that I am allergic to plastic
MM: HK will probably make it for Thursday of 28 December if she manages to catch the first train from London. She will have some lettuce, I will then meet you outside Komkya Hotel and pass it to you.
Jp: I drove to Tella and found all the materials I need, actually they are not that bad.
MM: that’s cool, I’ll give you a buzz when things are Ok
Jp: I should have done enough preparation as well by then. Cool, see you soon.
MM: bye for now.

Just wanna let u know that the guy is still waiting for me in Tella, if you could provide lettuce asp as we discussed.
C Ya
Jp
....
Still waiting for HK, hasn’t got all of it yet, let the guy in Tella know that it is on its way. Hope that’s alright
Mak mo

Got all, can we meet @ Komkya@ 7.30 2day?
....
Yeah, Cool.

Translated as:
I got all, can we meet at Komkya at 7.30 today

Figure 4: An example of the distribution of chain members.

Figure 4 shows an example of lexical chains developed from a multi-source textual data based on Wordnet senses. For example, the word ‘*lettuce*’ will be interpreted by the system as ‘cash’ and it appears 3 times, and is related to plastic. These chains are then used to segment textual data from different textual data source.

Segmentation: There are two tasks performed in ASKARI segmentation process; window identification and the application of lexical chains. A window is a region or a portion of data chosen to start the analysis. It is an important process which is used in most data analysis phases. The majority of lexical-cohesion-based segmentation techniques such as [22, 25] select this window arbitrarily. But, identifying the window arbitrarily in an application such as the one presented by ASKARI approach may jeopardise the end results. This is because criminal communication from different sources differs, for example telephone conversations can be longer, as most criminals may believe that it is difficult for people to tap telephone conversations, text messages (or SMS) can be very short; mainly because they tend to be used to communicate on the 'time' and 'geographical location' where criminals can meet. Emails may be short as well, because criminals believe that it may be easy to tap on their emails.

In order to identify the window, lexical repetition is used based on [26] who pointed out that cohesion can best be explained by focusing on how lexical repetition is manifested, in numerous ways, across pairs of sentences or utterances spoken by criminals through the telephone or other forms of communication. In order to identify the similarity *sim* between two utterances, U_i and U_j, cosine similarity measure is used in which the cosine of U_i and U_j frequency vectors should be close or equal to 1. The cosine similarity measure, denoted sim U_i, U_j, is defined as

$$\text{sim } U_i, U_j = \cos fi, fj = \frac{\sum_k f_{ik} \times f_{jk}}{\sqrt{(\sum_k f_{ik}^2) \times (\sum_k f_{jk}^2)}} \tag{1}$$

where $0 \leq \cos fi , fj \leq 1$.

$\sum_k fik \times fjk$ is the inner product of fi and fj, which measures how much the two vectors have in common. $\sqrt{(\sum_k f_{ik}^2) \times (\sum_k f_{jk}^2)}$ is a product of the two vector lengths which is used to normalise the vectors.

The similarity measure assumes that similar terms tend to occur in similar utterances. In such instances, the angle between them will be small, and so the cosine similarity measure will be close to 1. On the other hand, utterances with little in common will have dissimilar terms, the calculated angle between them will be close to π/2 and the similarity measure will be close to zero. As a result, similarity matrix chart is produced in which areas of related utterances/sentences are indicated. This also represents a temporary topical boundaries or segments. These segments are temporary as they are purely statistical and not based on the context within the textual document. On the other hand, these segments are useful as they can be a distinct window to start the analysis.

After obtaining the initial window for analysis, the lexical chains which were generated earlier are then applied on the chosen window. Similar procedures are applied as the ones in the *word frequency algorithm* presented by Reynar [19] who employs burstiness to determine the topic boundaries. However, instead of looking at the distribution of words in the document as done by Reynar [19], the distribution of the chain members in a transcript is identified. This helps ensuring that each segment is represented by the span of a lexical chain which represents a particular topic (which describes the main criminal activity) in the text. The highest frequency lexical chain within that window is considered as a main criminal activity, and is named as a *crime activity topic chain*, while the less frequent chains contain the sub-crime activity topics.

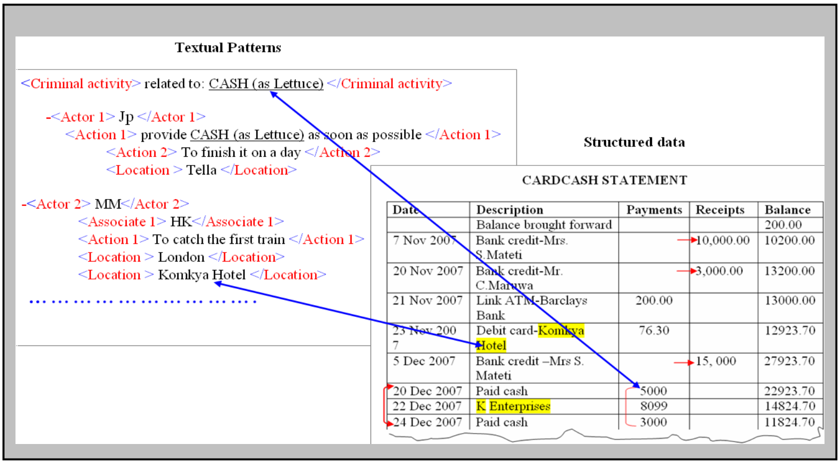
Starting from the first window, identified crime activity topic chain is used to extend the window or slide the window following the distribution of the topic chain members in the textual data. More specifically, the window slide following the appearance of the *crime activity topic chain* members. As the window expands, it will reach a stage whereby the appearance of any of the members from that particular lexical chain fades. This is the point where the *crime activity topic chain boundary* is identified.

Extraction: Information Extraction (IE) aims to extract *facts* from documents collection by means of NLP techniques [27]. In other words, IE is the process of identifying relevant information where the criteria for relevance are predefined by the user in the form of a template that is to be filled.

The ASKARI extraction task is concerned with identifying and extracting statements which belong to the most frequent lexical chain in that particular segment; this aims to extract the criminal activity to be performed, suspicious actions to be taken by the actors which appear in the segment, and the actors' identities. An action is a process of doing something to achieve the required objectives [28]. Linguistic pattern recognition method is used in the extraction process, a list of these patterns can be found in [17].

2.2.4 Textual patterns, Link analysis and structured data

Having extracted the statements indicating the criminal activity, the actions and actors involved, the resulting textual patterns will appear as shown in Figure 5. Each item in the pattern will be compared with items in structured data. Often, through relating various types of data, it is possible to find some clues that can alert the criminal investigators on possible 'organisation' of a criminal activity. For example the arrows in Figure 5 indicate some relationships between the



Textual Patterns

<Criminal activity> related to: CASH (as Lettuce) </Criminal activity>

-<Actor 1> Jp </Actor 1>
<Action 1> provide CASH (as Lettuce) as soon as possible </Action 1>
<Action 2> To finish it on a day </Action 2>
<Location > Tella </Location>

-<Actor 2> MM</Actor 2>
<Associate 1> HK</Associate 1>
<Action 1> To catch the first train </Action 1>
<Location > London </Location>
<Location > Komkya Hotel </Location>
……………………………….

Structured data

CARDCASH STATEMENT

Date	Description	Payments	Receipts	Balance
	Balance brought forward			200.00
7 Nov 2007	Bank credit-Mrs. S.Mateti		10,000.00	10200.00
20 Nov 2007	Bank credit-Mr. C.Maruwa		3,000.00	13200.00
21 Nov 2007	Link ATM-Barclays Bank	200.00		13000.00
23 Nov 2007	Debit card-Komkya Hotel	76.30		12923.70
5 Dec 2007	Bank credit –Mrs S. Mateti		15, 000	27923.70
20 Dec 2007	Paid cash	5000		22923.70
22 Dec 2007	K Enterprises	8099		14824.70
24 Dec 2007	Paid cash	3000		11824.70

Figure 5: The relationship between textual patterns and structured data.

Cardcash statement (structured data) and the textual data patterns. Such relationships and patterns are crucial in preventing an organized crime from happening.

2.3 Multi-agent system

Multi-agents can be defined as computing entities that perform user-delegated tasks autonomously. As agents are increasingly been used in a variety and many applications, it appears to be rare for an agent to act in isolation, and it is even uncommon for an agent to be useful on its own [29]. But it is common for an agent to be and to act within an environment that involves other agents, hence a multi-agent system (MAS). In addition, a collection of agents need ways to interact to each other in order to be useful.

Agent-based applications have been used in a variety of applications, namely in manufacturing and telecommunications systems, air traffic control, traffic and transportation management systems, information filtering and gathering, electronic commerce, as well as in entertainment and medical care domains [30]. One of the most compelling applications of agent technology is their ability to assist users in coping with the information overload problem. Agent systems are powerful medium to search, monitor and detect specific features across large corporate databases and texts on behalf of their users.

The ASKARI approach proposes to use MAS which includes three agents; Pattern scrutinizing agent, Standalone agent, and Communication agent.

- A *Pattern Scrutinizing Agent* (PASM): assessing the received patterns from the text mining stage and label them based on the type of crime, then rank them accordingly and develop a behaviour pattern which demonstrate the sequence of steps taken by the criminals to organise crimes. By understanding these sequences it can help the officials to find means to control the much early stages of the organised crime, which is its organisation processes.
- *Standalone Agent* (SA)**:** new patterns may appear which require knowledge on the new concepts. Standalone agent will incorporate expert-derived knowledge collected from the crime analysts and the rules defined by PASM. Further discussion on this agent will be published in the near future
- *Communication Agent* (CA)**:** This is an agent which allows information sharing or agent communication between agents within the MAS. Also it allows information sharing between the communication agent and the users crime analysts. This means that the agent will have particulars about specific crime analysts who are particularly working in different types of organised crime – say can have a list of all officers/analysts who deal with Paedophiles, terrorism, fraud etc. After PASM has labelled and ranked the patterns. The SA will check if there is an unknown pattern and use the given knowledge to rank and label the pattern. The CA will then use the ranks to identify the most threatening activity at that particular moment, and use the label to identify an appropriate crime analyst and send the crime alerts to them. The ASKARI approach aims to use KQML (Knowledge and Query Manipulation Language) because KQML semantics are described in terms of pre, post and completion

conditions, where's other languages such as FIPA ACL semantics are based on speech acts as rational actions [31–33].

PASM is the key agent that will facilitate the prediction process in the ASKARI systems. For this reason PASM features are further expanded in the next section.

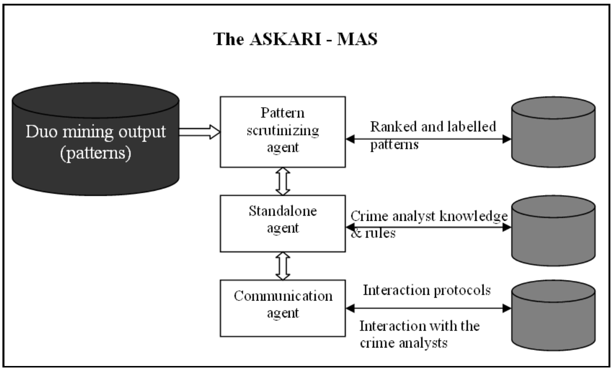

The ASKARI - MAS
Duo mining output (patterns)
Pattern scrutinizing agent
Ranked and labelled patterns
Standalone agent
Crime analyst knowledge & rules
Communication agent
Interaction protocols
Interaction with the crime analysts

Figure 6: The MAS in ASKARI.

2.3.1 PASM

In this section PASM is demonstrated based on the fictitious data discussed in Figure 4, which includes structured and unstructured data. One of the PASM objectives is to scrutinise and understand the criminal patterns received from the duo mining process and provide criminal behaviour pattern or a sequence of steps taken by the criminals to organise crimes. However, identifying criminal behaviour patterns is a complex problem, and the complexity is compounded when same criminals use different patterns for different crime activities, resulting into even larger criminal network with many more activities and locations related to the organisation of the crime. This problem may be exacerbated by human behaviour which exhibits both systematic regularities and inherent unpredictability.

In order to successfully draw on human behaviour in applications such as crime prediction and alerting, a reasoning mechanism about the inherently uncertain properties of crime entities must be introduced. Tasks such as information fusion for organised crime detection and deterring require both reasoning under uncertainty and logical reasoning about discrete entities. Such information requires some kind of decisions to be made in the presence of uncertainty. The main tool to deal with situations which characterized by the absence of certain knowledge is probability theory (as used in applications such as: [34, 35] and [36]). The basic element in probability theory is the random variable which can be thought of as describing a part of the world whose state is initially unknown. Probability is the most applied logic for computational scientific reasoning under uncertainty [37]. However, appropriate application of probability theory often requires logical reasoning about which variables to include and what the appropriate probabilities are, in order to maintain a semi or a full-real-time predicting and alerting system [38]. PASM incorporates a Bayesian network [39] which is based on the probability theory, in order to be able to describe the unknown state of some criminal entities or patterns.

For example, the mined patterns contain entities that can be used to produce Bayesian Network graphical structures such as the one shown in Figure 7.

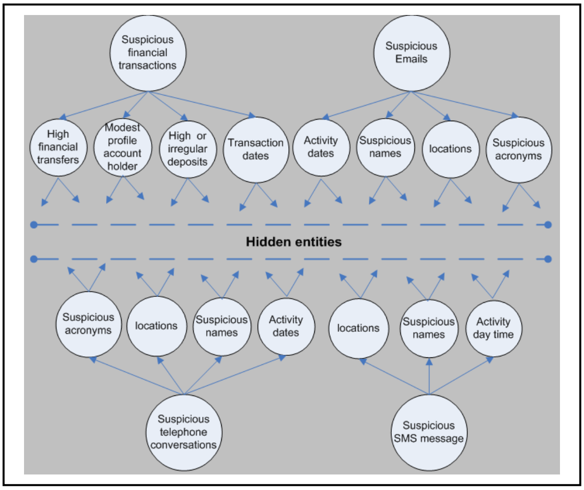
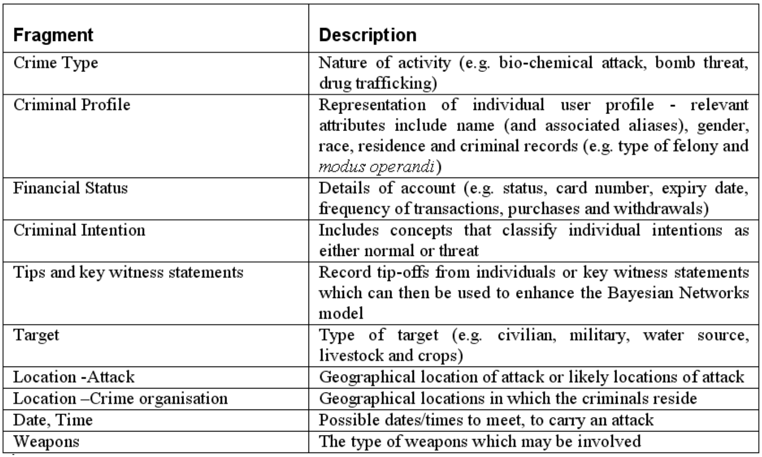


Figure 7: Bayesian networks entities of suspicious activities.

Some attribute-value representation of standard Bayesian networks can be insufficient to express the criminal behaviour problems. For example, a Bayesian network may apply to a single problem which is not related to a mesh of problems such as money laundering, weapons, drugs, unlawful trading, unlawful possession of wrong identities etc as those of organised crime. Figure 7 involves multiple sources of information, each of which owns and maintains multiple entities of different types, in which standard Bayesian network can provide no way of compactly representing the correlation between the entities in a multiple source. In other words, Bayesian Networks are limited in that they are able to represent only a single fixed set of random variables which has different evidence from problem to problem. A much more flexible representation capability is required to model human behaviour in different situations. As a consequent, PASM adapt Multi-Entity Bayesian Networks MEBN developed by [40], which has been used in different applications [40–42]. MEBN is a first order language for specifying probabilistic knowledge bases as parameterised fragments of Bayesian networks. MEBN fragments MFrags can be instantiated and combined to form graphical probability models. An MFrag represents probabilistic relationships among a conceptually meaningful group of uncertain hypothesis [40]. Each MFrag represents probability information about a group of related random variables.

The model, proposed in this paper, consists of ten fragment patterns which are used to distinguish normal patterns from criminal behaviour patterns that may pose a threat. The fragments identified by the ASKARI approach are listed in Table 5.

Table 2: Proposed MFrags for organised crime.



Fragment	Description
Crime Type	Nature of activity (e.g. bio-chemical attack, bomb threat, drug trafficking)
Criminal Profile	Representation of individual user profile - relevant attributes include name (and associated aliases), gender, race, residence and criminal records (e.g. type of felony and *modus operandi*)
Financial Status	Details of account (e.g. status, card number, expiry date, frequency of transactions, purchases and withdrawals)
Criminal Intention	Includes concepts that classify individual intentions as either normal or threat
Tips and key witness statements	Record tip-offs from individuals or key witness statements which can then be used to enhance the Bayesian Networks model
Target	Type of target (e.g. civilian, military, water source, livestock and crops)
Location -Attack	Geographical location of attack or likely locations of attack
Location –Crime organisation	Geographical locations in which the criminals reside
Date, Time	Possible dates/times to meet, to carry an attack
Weapons	The type of weapons which may be involved

Based on the MFrags shown in Table 5 PASM can reveal the various steps adopted for that particular criminal activity. Applying the well known spiral life cycle model [43] on the probability networks fragments (from Figure 10), a systematic way of understanding the sequence of steps which describes particular behavioural patterns of an organised crime will be provided as shown in Figure 11. These patterns function as a trigger to the firing of a potentially suspicious node of activity as interpreted by the PASM, and are sent to communication agent CA which may then issue a warning message to the appropriate analyst. As new evidence emerges, the agent increases the probability of the corresponding suspicious node and when a certain threshold of suspicion is reached the agent sends a strong alert to the analysts.

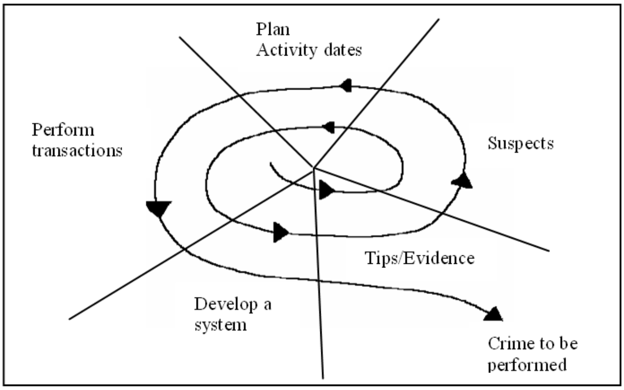


Figure 8: An example of spiral life cycle model.

3 Conclusions

This chapter has provided an overview of the potential of the duo mining and MAS technologies in assisting law enforcement agencies in their monitoring and anticipation of criminal activities. Duo mining is particularly suited to analyse large volumes of data sources and can provide analysts with a valuable tool to sift through huge amount of records and discover any useful patterns, any hidden clues, and any meaningful associations between individuals, organisations and crime incidents. The ASKARI project described combines duo mining with Multi-agent technology with the view to support analysts with new incoming data and also to provide a watchful eye on criminal activities. In addition, the systematic steps followed by different criminal that are provided by the PASM can be crucial for analysts to stop the organisation process of most crimes. Early stopping of crime organisation process can reduce cost, time and police and analysts efforts.

Acknowledgement

The authors would like to thank Professor Peter Musgrove for his suggestions which have considerably improved the paper.

References

[1] Organised Crime in South Africa, Monograph 28. The African Security Review, 1998.

[2] Scheiber, L. B., Hartka J. E., & Randall, S. M., Defender's Edge: Utilizing Intelligent Agent Technology to Anticipate Terrorist Acts, Institute for Defence Analyses Alexandria, VA ADA419005, 2003.

[3] Hidalgo, J. M. G., Tutorial on Text Mining and Internet Content Filtering, *Proceedings of the ECML/PKDD'02*, Helsinki, Finland, 2002.

[4] Hearst, M., Text Data Mining, in *The Oxford Handbook of Computational Linguistics*, ed. R. Mitkov: Oxford University Press, 2003.

[5] Zanasi, A., Text Mining: A New Paradigm?, in *Text Mining and its Applications to Intelligence, CRM and Knowledge Management,* ed. A. Zanasi, WIT Press: Southampton, UK, pp. xxvii, 2005.

[6] Meij, J. & Bosch, A., Text Mining Techniques, in *Dealing with the Data Flood*, ed. J. Meij, STT/Beweton: The Hague, Netherland, pp. 746–753, 2002.

[7] Tan, A. & Teo, C., Learning user profiles for personalized information dissemination. *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, Alaska, 1998.

[8] Sullivan, D., Application Integration in Applied Text Mining. *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, ed. A. Zanasi, WIT Press: UK, pp. 144–154, 2005.

[9] Mena, J., *Investigative Data Mining for Security and Criminal Detection*: Butterworth-Heinemann, 2003.

[10] Mena, J., *Homeland Security Techniques and Technologies*: Charles River Media, 2004.

[11] Berry, M., Survey of text mining: clustering, classification, and retrieval, *Proceedings of the 2nd SIAM International Conference on Data Mining*, Airlington Virginia, 2004.

[12] Berry, M. & Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support,* John Wiley & Sons, Inc, 1997.

[13] Silva, E. M., Prado, H. A. D. & Ferneda, E., Text Mining: crossing the chasm between academic and the industry, in *Data Mining III*, eds. A. Zanazi, C. A. Brebbia, N. F. F. E. Ebecken, and P. Melli, WIT Press: UK, pp. 351–361, 2002.

[14] Chibelushi, C. Sharp, B. and Shah, H., ASKARI: A Crime Text Mining Approach, *Digital Crime and Forensic Science in Cyberspace*, eds. P. Kanellis, E. Kiountouzis, N. Kolokotronis, and D. Martokos. Idea Group Inc: USA, pp. 155–174, 2006.

[15] Brown, D. E. & Gunderson, L. F., Using Clustering to Discover the Preferences of Computer Criminals, *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, **31**, pp. 311–318, 2001.

[16] Brown, D. E. and Hagen, S., Data Association Method with Applications to Law Enforcement, *Decision Support Systems*, **34**, pp. 369–78, 2002.

[17] Chibelushi, C., Text Mining For Meeting Transcript Analysis to Support Decision Management, *Faculty of Computing Engineering and Technology*. UK: Staffordshire University, 2008.

[18] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., CRISP-DM 1.0 Step-by-step data mining guide, in *SPSS*: CRISP-DM consortium, 2000.

[19] Reynar, J., Topic Segmentation: Algorithms and Applications, *Computer and Information Science*. College Park, Maryland: University of Pennsylvania, 1998.

[20] Trost, H., Morphology, in *The Oxford Handbook of Computational Linguistics*, ed. R. Mitkov, Oxford University Press: UK, pp. 25–47, 2003.

[21] Choi, F. Y. Y., Advances in Domain Independent Linear Text Segmentation, *Proceedings of NAACL00*, Seattle, 2000.

[22] Hearst, M., Multi-paragraph Segmentation of Expository Text, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994.

[23] Rayson, P., Wmatrix: a Web-based Corpus Processing Environment., *Proceedings of the ICAME 2001 Conference*, Université Catholique de Louvain, Belgium, 2001.

[24] Morris, J. and Hirst, G., Lexical Cohesion Computed, *Computational Linguistics*, **17**, pp. 21–48, 1991.

[25] Stokes, N., Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain, *Department of Computer Science*: University College Dublin, 2004.

[26] Boguraev, B. K. and Neff, M. S., Discourse Segmentation in Aid of Document Summarization, *Proceedings of 33rd Annual Hawaii*

International Conference on System Sciences, (HICSS), Maui, Hawaii, 2000.

[27] Gaizaukas, R. and Wilks, Y., Information Extraction: beyond document retrieval, *Journal of Documentation,* **54**, pp. 70–105, 1998.

[28] Soanes, C. Waite, M. and Hawker, S., (eds). *The Oxford Dictionary, Thesaurus, and Wordpower Guide*, Oxford, UK.

[29] Fasli, M., *Agent Technology for E-commerce*. West Sussex: John Wiley & Sons, 2007.

[30] Jennings, N. R. and Wooldridge, M., Applying agent technology, *Applied Artificial Intelligence*, **94**, pp. 351–361, 1995.

[31] Cohen, P. R. and Levesque, H. J., Rational Interaction as the Basis for Communication, *Intentions in Communication*, eds. P. R. Cohen, J. Morgan, and M. E. Pollack, Cambridge, MA, 1990, pp. 221–256.

[32] Bretier, P. and Sadek, D., A Rational Agent as the Kernel of a Cooperative Spoken Dialogue System: Implementing a Logical Theory of Interaction in *Intelligent Agents III: Agents Theories Architectures, and Languages*, **1193**, eds. J. Muller, M. Wooldridge, and N. Jennings, Springer: Berlin, pp. 189–203, 1997.

[33] Sadek, M. D., A Study in the Logic of Intention, *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, Cambridge, MA, 1992.

[34] Howson, C. and Urbach, P., *Scientific Reasoning: The Bayesian Approach*. Chicago, IL: Open Court, 1993.

[35] Finetti, B., *Theory of Probability: A Critical Introductory Treatment*, Wiley: New York, 1934/1990.

[36] Jaynes, E. T., *Probability Theory: The Logic of Science*. Cambridge University Press: Cambridge, UK, 2003.

[37] Fung, F., Predicate Logic-based Assembly of Situation-specific, **2005**: IET, 2004.

[38] Sargunar, V., An Introduction to Bayesian Networks for Multi-Agent Systems, *Intelligent Systems Laboratory (ISLAB) Workshop*, 2003.

[39] Ramoni, M. and Sebastiani, P., Bayesian Methods, in *Intelligent Data Analysis*, eds. M. Berthold and D. J. Hand, 2 edn. Springer: Germany, pp. 130–168, 2003.

[40] Laskey, K. MEBN: A Language for First-order Bayesian Knowledge Bases, *Artificial Intelligence*, **172**, pp. 140–178, 2008.

[41] Hudson, L. D. Ware, B. S., Mahoney, S. and Laskey, K., Antiterrorism Risk Management for Military Planners (with) George Mason University Homeland Security and Military Transformation Laboratory, George Mason University Homeland Security and Military Transformation Laboratory, 2005.

[42] AlGhamdi, G., Wright, E., Barbara, D. and Chang, K., Modelling Insider Behaviour Using Multi-Entity Bayesian Networks, *Proceedings of the 10th Annual Command and Control Research and Technology Symposium*, 2005.

[43] R. Pressman, *Software Engineering: a practitioner's approach*, 2nd edn. McGraw-Hill, 1987.