Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights

Rick Adderley, Patrick Seidler
A E Solutions (BI) Ltd.
Badsey, UK
rickadderley@a-esolutions.com
patrickseidler@a-esolutions.com

Atta Badii, Marco Tiemann
University of Reading
Reading, UK
atta.badii@reading.ac.uk
m.tiemann@reading.ac.uk

Federico Neri, Matteo Raffaelli
Synthema srl
Pisa, Italy
federico.neri@synthema.it
matteo.raffaelli@synthema.it

Abstract—**This paper describes the implementation of a Data Mining (DM) system under the EU FP7 Security Research Project Multi-Modal Situation Assessment & Analytics Platform (MOSAIC). The system aims to enable the part-automatic detection and recognition of crime threats in uncertain environments. It facilitates the automatic retrieval of intelligence data providing deep semantic information access and dynamic classification features for distributed data sources, such as Policing legacy databases, Police text documents and free text database fields. A specific pipeline of linguistic processors that share a common knowledge base on crime patterns has been created to retrieve entities and events from text documents and websites. Structured and unstructured data retrieved from the individual data sources are integrated in a semantically query-able unified data representation using specific ontological models. A domain specific entity resolution module ensures the resolution of conflicting and misleading identities to enable data retrieval and fusion from disparate data sets. As criminal network analysis depicts a major part of the intelligence process, specific measures and algorithms have been developed to support analysts in retrieving, analysing, and disrupting criminal networks.**

Keywords-Data mining; text mining; entity recognition and resolution; social and criminal network analysis; semantic interoperability.

I. INTRODUCTION

Despite huge progress in Data Mining (DM) in the last decade, a gap remains between DM technologies and the actions that are taken upon knowledge creation based on them [1]. The most labour intensive and at the same time most expensive parts of mining projects are generally concerned with data pre-processing, i.e., with preparing data in such a way to be able to further examine data for meaningful information [2]. The fact that data pre-processing is often embedded in a large amount of domain knowledge might explain the slow progress in the area which is to be retrieved repeatedly for each project and is then often encoded in low level system parts such as in Structured Query Language (SQL) statements.

Therefore, the efficiency of any institution still relies heavily on the human factor to close this gap [3], limiting the DM process and the applicability of DM itself. DM can, for example, reveal all the data to create an offender profile, but the existing systems are often not able to sufficiently link known profiles with unsolved crimes, i.e., other forensic evidence such as the method of offending [4]. This lack of sufficiently enriched data in some parts of the DM process often creates a knowledge gap that hinders effective and targeted intervention, but leaves analysts with labour-intensive bottlenecks [5].

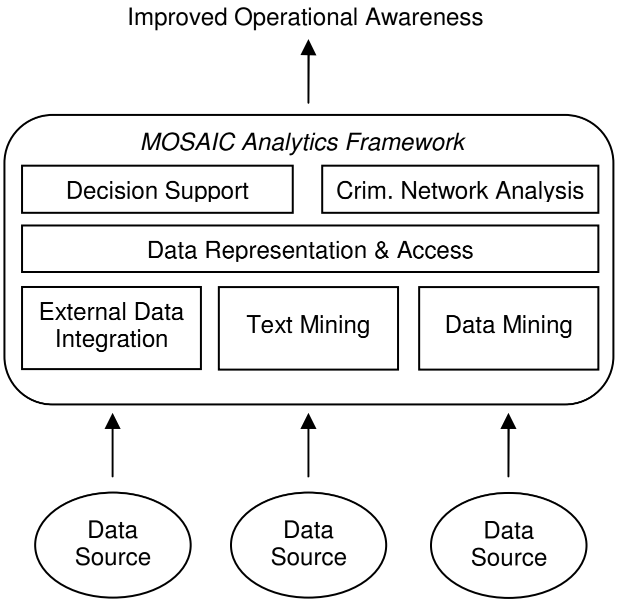
Improved Operational Awareness
MOSAIC Analytics Framework
Decision Support
Crim. Network Analysis
Data Representation & Access
External Data Integration
Text Mining
Data Mining
Data Source
Data Source
Data Source

Figure 1. Overview diagram of the MOSAIC Analytics Framework.

The primary objective of the Multi-Modal Situation Assessment & Analytics Platform (MOSAIC) (see Fig. 1) is to improve the targeted surveillance of and intervention into complex systems of criminal behaviour by combining intelligence to provide a decision support system for the relevant authorities. The system facilitates the correlation of data from disparate sources into a semantically operational system to form contextual and valuable information – the information whole being greater than the sum of its parts, and thus to enable targeted surveillance. The system uses a loosely coupled system architecture where sensing and analysis components communicate through Web Services and exchange data through a central system.

Mining and analysis of different kinds of data including data taken from legacy databases and heterogeneous sources of text from sanitised Police reports, from free text database fields, and from WWW public sources allows the user to integrate those data within a unified framework in order to be able to conduct social and criminal network analysis. The framework has been designed to be compatible with existing procedures, tools and legacy systems used by Police forces within the European Union.

In this paper, we begin by describing the MOSAIC data sources. We then outline the system architecture, including the analysis components and the semantic interoperability approach. Finally, we report on a preliminary case study and user experiments.

II. Data Sources

Police repositories hold millions of entries on crimes, offenders, and other intelligence. For the purpose of the project and the mining of those data by the DM Component, a representative MOSAIC legacy database has been created and sanitised using original Police data, representing data on Nominals, Crimes, Intelligence, Automatic Number Plate Recognition (ANPR), and Stop & Search.

The Text Mining (TM) Component takes as input two types of data, whereas anonymised entities within the documents have been reproduced in the MOSAIC legacy databases. The first type is two case studies provided by Police partners spanning a wide variety of Police disciplines. For example, one of the case studies is an 85 pages case document of a missing person investigation, which took place over three months. The second type constitutes long text fields mainly in the Intelligence table. This free text cannot be analysed with standard DM algorithms, but will be passed to the TM Component to retrieve additional information on entities and their relations.

III. Data Preparation

A. Entity Resolution Component

Poor data quality is a constant issue in Policing systems. Many errors are introduced when data is entered into the systems. Moreover, we cannot expect that data are always easily identifiable through global unique identifiers. If an organisation or institution is not able to identify unique objects, suboptimal decisions will be the result.

Using the Apache Lucene framework as a backend indexing tool, an entity resolution module is being developed in order to resolve those issues. A decision engine uses scenarios that contain predefined probability levels for matches on specific database field types which are used to calculate the final probability that a match has been found. Matching fields are thereby compared using the Bayes function using as input the pre-determined probability for each field. Finally, fuzzy string matching is to be introduced providing Metaphone [6] and Soundex [7] string matching options.

Preliminary results on the algorithm's performance and its accuracy in correctly matching entities show a minimum accuracy of 65% for 90% of test runs, whereas several combinations show an accuracy of up to 88% when compared to the Gold Standard. In comparison, compilation of the Gold Standard took the analyst 1 ½ days, involving handcrafting 1002 offender records into sets containing the same individual.

B. Data mining workbench

During their work, analysts iterate through a set of unspecified tasks in no particular order and as needed. The widely used National Intelligence Model (NIM [8]) does not provide a structured approach to those tasks.

We formalise analysis tasks using the Cross Industry Standard Process for Data Mining (CRIPS-DM [9]) model in conjunction with the intelligence cycle. DM algorithms for the preparation and analysis of data are being implemented into an integrated MOSAIC DM workbench to assist analysts in manipulating data without the hassle of having to access disparate systems. Using the workbench, analysts will be able to use data search and linking, exploration, modelling and visualisation capabilities through a process of interconnected nodes. The approach taken accommodates for the various possible working environments and data requirements in which the final system could be applied. The resulting DM processes will be reusable and can be re-run any time taking into account data that have newly arrived in the system.

IV. Analysis Component

Data analysis inside law enforcement has remained a time-consuming process, with technical support restricted to a large number of unconnected systems tools lacking support in the provision of actionable intelligence. It has further been argued that an operational gap exists between intelligence analysis and operational policing, with advanced technology often used to manage offenders rather than providing insights on criminal behaviour and possible interventions [5].

A. Data analysis algorithms

To fulfil a request for information the analyst performs a query through disparate systems, researching, e.g., names, addresses and telephone numbers. Intelligence logs are individually read by the analyst and recorded into a spreadsheet and/or directly into one of the existing link visualisation tools. This largely manual acquisition and preparation of data is time-consuming and prone to error as the amount of data to search exceeds the humanly comprehensible limit [10].

The MOSAIC system offers DM support that has been tailored to analysts' needs regarding their work tasks, processes, and their needs for actionable operational intelligence. The main focus is put on the creation of such results that are immediately and easily applicable inside the intelligence cycle.

- Offender mining and automatic assignment of priorities to offenders: Track prioritised criminal behaviour and enable law enforcement to allocate responsive actions in order to meet Force priorities.
- Identification of crime series and mapping of known offenders to unsolved crimes: Application of self-organising maps to link spatial, temporal and modus operandi (MO) and overlay of offender data onto clusters of similar crimes thereby suggesting possible involvement.
- Identification of criminal roles: Create offender profiles, group offenders by their profile, and apply a K-means clustering algorithm to determine the prominent group for all offenders.

B. *Text Mining Component*

Data which have been retrieved from free text database fields, from the document repository and from the World Wide Web (WWW) will be converted into Clean TXT format, processed by the Natural Language Processing (NLP) engine and indexed.

In MOSAIC, TM and entity extraction are going to be applied through a pipeline of linguistic and semantic processors that share a common knowledge and a common ontology. A crime ontology and a domain specific knowledge base with crime patterns, abbreviations, technical terms and terms relationships mainly extracted from the sanitised Police reports are created. This shared ontology and knowledge base guarantees a uniform interpretation layer for the diverse information from different sources.

The TM process is implemented by the following steps:

- Morpho-syntactic or Part-of-Speech (POS) tagging
- Multiword tagger (MWT)
- Word-sense disambiguation (WSD)
- Named-entity recognition (NER)
- Semantic role labelling (SRL)
- Entity Relationship extraction

At the heart of the morpho-syntactic analysis module, which aims at identifying the part-of-speech (POS tagging), is McCord's theory of Slot Grammar [11][12]. The module will analyse each sentence, cycling through all its possible constructions and trying to assign the context-appropriate meaning – the sense – to each word by establishing its context. The parser – a bottom-up chart parser – employs a parse evaluation scheme used for pruning away unlikely analyses during parsing, as well as ranking the final analyses. It will build the syntactical tree incrementally. Multi-word combinations are then identified and ambiguous terms disambiguated depending on the syntactic and semantic context, by considering super-subordinate related concepts.

These two modules are closely related to named-entity recognition. Extensive effort is being spent on the identification of pre/suffixes, specific linguistic patterns and specific data formats for the English language in order to recognise the following entities in texts: dates, addresses, person names, locations, license plate numbers, brands, web entities (web addresses, Internet Protocol (IP) addresses, email addresses, etc.), bank accounts and phone numbers. Entities are reduced to their semantic roles (agent, predicate, theme, recipient, time and location; in simpler terms: *who* does *what* to *whom*, *how*, *when* and *where*), identified as a result of the dependency parsing. The NLP engine will then be able to extract entity relationships from a text. Heuristic algorithms are being implemented in order to extract all kinds of relationships between the entities mentioned above.

C. *Social and Criminal Network Analysis and Visualisation Component*

Empirical research has shown that people who have a propensity to commit crime rarely work in isolation, but in a group of associates who have differing skills and interests to complement the activities of individuals or sub groups within their criminal network [13][14][15]. As security and law enforcement resources are not unlimited, prioritisation decisions have to be made for policing and investigative effort. It is, therefore, highly desirable to be able to identify, characterise and rank the networks which are operating within an area so as to identify, and prioritise for further investigation, those networks and individuals within them that are most significant in terms of who are causing the most harm.

The aim of the MOSAIC criminal network analysis and visualisation component is to support law enforcement in continuously grasping a full picture of current criminal activity and close the gap towards previsional systems by evaluating beforehand the impact of decisions. Results shall enable agencies to create improved intelligence products on effective ways for effectively disrupting criminal activity.

To create networks from structured data, we use the approach outlined by Adderley et al. [16]. The algorithm identifies all of the criminal networks that are present in a dataset and prioritises those that are causing most harm to the community based on a crime scoring mechanism. We further provide algorithms that combine network topological measures with domain based weighting scores, and enable the identification of criminal roles, sub group and network themes, and the running of network robustness simulation tests against target law enforcement interventions. Visualisation will be achieved by presenting the network structure in a 3D environment with textual statistics and data overlays.

V. Semantic Interoperability

When extracting and analysing data from multiple and quite distinct data sources, integrating the gathered and extracted data and information from these sources becomes a significant issue: in current practice, police intelligence analysts need to gather the available information from a multitude of completely separate systems with different output formats and to then manually create unified representations based on the data gathered - clearly not an efficient procedure. To reap the benefits of automated data analysis on a large scale, data must be made accessible through a single system. And to be able to combine different sources of information in order to find previously “unfindable” connections, the data to be integrated must “speak the same language”. The available information must be made semantically interoperable. In MOSAIC, semantic interoperability involves three main aspects: the definition of a semantic domain model which can represent the available information while preserving its meaning; the development of a system that organises the available information using the developed model that makes it accessible; the connection to the individual data sources and to any further “consumers” of the data.

The world model for MOSAIC is being defined as an Ontology Web Language (OWL)-Lite model [17]. This model represents actors, objects, actions and other relevant information types as subject – predicate – object triples that establish object types, their properties and their relations to other object types. Data gathered is added to this model as instances of the defined types with the relevant properties and relations to other instances, thus populating the data model. The data model has been developed in collaboration with police partners using real-world scenarios and refined given the available data in order to retain conceptualisable and groundable concepts only [18].

A semantic data store will be used to manage the processes of creating, reading, updating and deleting instance data within the semantic representation model. MOSAIC uses a data store implementation that stores data triples – a triple store. The MOSAIC data store is based on the Apache Jena project and uses the core Apache Jena components for data storage and access as well as the Apache Fuseki Web Service front end. Data in the MOSAIC data model can be queried and updated using the Simple Protocol and Resource Description Framework (RDF) Query Language (SPARQL) [19], which provides an equivalent to SQL for accessing and updating data that is stored in the form of triples. The data store implementation has been extended with additional MOSAIC-specific features such as the ability to subscribe with queries in order to receive notifications when new relevant data are added to the data store.

Semantic interoperability can only be achieved when the data of interest is adequately integrated into the MOSAIC data store; to this end, data importers have been integrated as data store plugins. These importers provide Web Service endpoints to which source data can be sent in their native formats. The data are then analysed for consistency, converted in terms of terminology and representation via mediator components and added to the MOSAIC data store.

The MOSAIC data representation and data store system allows analysts and operators to query a single data representation for information across information provided by all of the data sources described. The ontology used in MOSAIC extends this by allowing users to make use of the knowledge encoded in the ontology while querying it – a trivial example for this is the ability to query for persons involved in violent crimes without having to enumerate the individual identifiers for violent crimes as might be necessary in a conventional SQL database.

The semantic representation of data can also be used to reason using the world model and to define complex events that may be hard to spot by human operators but that can be defined as sets or sequences of events that taken together either lead to new information or should trigger a specific (re-)action [20]. A reasoning and rule engine that is suitable to work with the triple data representation and can describe groups and sequences of observations entered to be matched and actions to be taken with the help of the MOSAIC system is currently being integrated.

VI. Evaluation

A preliminary case study has been conducted. The goal for users was to automatically identify the network(s) with the highest police force priority, the most prolific offenders inside the network, as well as appropriate interventions.

The analyst extracts data with the DM workbench and creates a problem profile that will be enhanced as more of the data is understood. To increase data quality, offender identities from the joined data set entries are resolved before starting a DM process. The output contains 995 unique identities compared to 1505 unique ids in the original data set. A DM process was then developed and applied which retrieves police force priority scores, crime roles and travel distances to develop a criminal profile for each offender.

Applying the network generation, 568 networks were identified from the dataset for networks with two degrees of freedom in 3.5 seconds. Respective generation of networks with 3 degrees of freedom took 69 seconds to run, and 165 seconds, while as a rule analysts will use two degrees of freedom in most circumstances as those cover the most common crimes and criminal networks for most Police areas.

Utilising offenders’ criminal profiles, the highest ranked network containing 57 unique offenders was identified and further analysed. Topological measures are added to each offender’s criminal profile and we retrieve a final prioritised list of offenders (see Table I) which facilitates decision making in targeting the appropriate person(s).

TABLE I. Top 3 Criminal Profiles In Data Set

Id	Role	Harm	Distance	Inform. Control	Access	Activity	Score
1	Burglary	600	Compact	Controller	Best	Active	30.49
2	Burglary	600	Compact	Some	Best	None	27.49
3	Violence	360	Compact	None	Average	None	17.91

We further evaluate effectiveness of interventions on the network. Based on degree centrality and domain scores, in each step the vertex with the highest overall rank amongst all vertices is selected for sequential removal, compared with a random removal approach. Testing the network for its robustness based on the largest remaining component [21], results show that by removing only the top two offenders, we are able to disrupt this specific network by 70% (see Fig. 2).

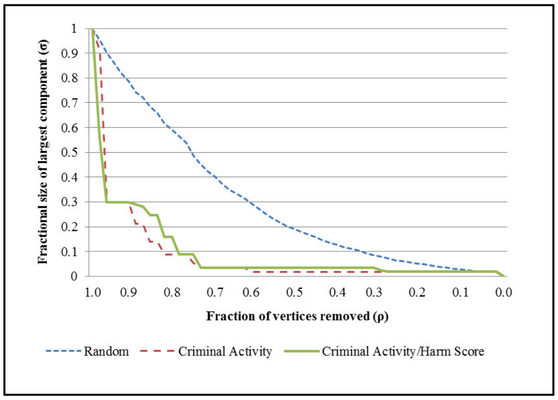

Figure 2. Robustness of criminal network under sequential attacks.

To provide explicit prototype evaluations, additional experiments were conducted involving seven domain experts who were asked to perform several sub tasks under two experimental conditions: 1) automated visualisation and analysis; 2) automated visualisation and manual analysis. The average total score provided by the experts was 13.86 from a maximum of 20, resulting in a 69.3% satisfaction level. Comments regarding how the prototypes could be improved were also provided.

VII. Conclusion and Future Work

This contribution described interim results of the MOSAIC project. It is in particular concerned with showing how data analysis and information mining techniques are applied in order to extract useful information from large amounts of noisy data, and how the extracted data can be represented and made accessible using a semantic integration system.

Future work in the project will involve the effective presentation of extracted information and reasoning over the extracted data with in order to aid in decision making processes based on the information extraction and analysis processes outlined in this contribution.

Acknowledgment

This work is supported by the European Commission in the 7th Framework Programme, within the Security Research Theme, under Grant 261776 with the acronym MOSAIC.

References

[1] P. Domingos, “Toward knowledge-rich data mining,” Data Mining and Knowledge Discovery, vol. 15, no. 1, 2007, pp. 21–28.

[2] L. Pipino and D. Kopcso, “Data Mining, Dirty Data, and Costs,” in Proceedings of the Ninth International Conference on Information Quality, MIT, 2004, pp. 164–169.

[3] Z. T. Kardkovács, “Business Intelligence and Data Mining,” in Research and Development in E-Business through Service-Oriented Solutions, K. Tarnay, S. Imre, and L. Xu, Eds. IGI Global, 2013, pp. 57–70.

[4] R. Adderley, “Exploring the Differences Between the Cross Industry Process for Data Mining and the National Intelligence Model Using a Self Organising Map Case study,” in Business Intelligence and Performance Management, P. Rausch, A. F. Sheta, and A. Ayesh, Eds. Springer London, 2013, pp. 91–105.

[5] P. Seidler and R. Adderley, “Criminal network analysis inside law enforcement agencies – a data mining system approach under the National Intelligence Model,” IJPSM, vol. 15, no. 4, 2013, pp. 323–337.

[6] L. Philips, “Hanging on the Metaphone,” Computer Language, vol. 7, no. 12, 1990, pp. 39-43.

[7] M. K. Odell, "The profit in records management," Systems, vol 20, no. 20, 1956.

[8] National Criminal Intelligence Service, “The National Intelligence Model.” National Criminal Intelligence Service, 2000.

[9] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” Journal of Data Warehousing, vol. 5, no. 4, 2000, pp. 13–22.

[10] Y.-W. Si, S.-H. Cheong, S. Fong, R.P. Biuk-Aghai, and T.-M. Cheong, “A layered approach to link analysis and visualization of event data,” in Seventh International Conference on Digital Information Management, 2012, pp. 181–185.

[11] M. C. McCord, “Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars,” in Proceedings of the International Symposium on Natural Language and Logic, London, UK, 1990, pp. 118–145.

[12] M. C. McCord, “Slot Grammars,” Comput. Linguist., vol. 6, no. 1, 1980, pp. 31–43.

[13] E. Patacchini and Y. Zenou, “Juvenile delinquency and conformism,” Journal of Law, Economic, and Organization, vol. 28, 2012, pp. 1–31.

[14] M. Warr, Companions in Crime. Cambridge Univ Pr, 2002.

[15] D. L. Haynie, “Delinquent peers revisited: does network structure matter,” American Journal of Sociology, vol. 106, 2001, pp. 1013–1057.

[16] R. Adderley, A. Badii, and C. Wu, “The Automatic Identification and Prioritisation of Criminal Networks from Police Crime Data,” in Intelligence and Security Informatics, vol. 5376, D. Ortiz-Arroyo, H. Larsen, D. Zeng, D. Hicks, and G. Wagner, Eds. Springer Heidelberg, 2008, pp. 5–14.

[17] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, “From SHIQ and RDF to OWL: the making of a Web Ontology Language,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 1, no. 1, pp. 7–26, 2003.

[18] A. Jakulin and D. Mladenic, “Ontology Grounding”, in Proc. 8th Intl. Multi-Conf. Information Society, 2005, pp. 170-173.

[19] J. Pérez, M. Arenas, and C. Gutierrez, “Semantics and Complexity of SPARQL,” in The Semantic Web - ISWC 2006, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds. Springer Berlin Heidelberg, 2006, pp. 30–43.

[20] J. Z. Pan, “A Flexible Ontology Reasoning Architecture for the Semantic Web,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, 2007, pp. 246–260.

[21] B. Keegan, M. A. Ahmed, D. Williams, J. Srivastava, and N. Contractor, “Dark Gold: Statistical Properties of Clandestine Networks in Massively-Multiplayer Online Games,” presented at the 2010 IEEE Second International Conference on Social Computing, Los Alamitos, CA, USA, 2010.