Prediction and Visual Intelligence for Security Information: The PREVISION H2020 Project

Konstantinos Demestichas
Institute of Communication and Computer Systems, Athens, Greece
cdemest@cn.ntua.gr

Thi Bich Ngoc Hoang
IRIT UMR5505 CNRS, Toulouse, France; Danang Univ. of Economics, Vietnam
Thi-Bich-Ngoc.Hoang@irit.fr

Josiane Mothe
IRIT UMR5505 CNRS, INSPE, Univ. de Toulouse
Josiane.Mothe@irit.fr

Olivier Teste
IRIT UMR5505 CNRS, Univ. de Toulouse, France. Olivier.Teste@irit.fr

Md Zia Ullah
IRIT UMR5505 CNRS, Toulouse, France. mdzia.ullah@irit.fr

ABSTRACT

This paper presents the on going work within PREVISION H2020 project. The mission of PREVISION is to empower the analysts and investigators of agencies with tools and solutions not commercially available today, to handle and capitalize on the massive heterogeneous data streams that must be processed during complex crime investigations and threat risk assessments.

CCS CONCEPTS

• **Information systems** → *Information integration*; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Data stream management, Data heterogeneity, Cybercrime, Social media analysis, Data fusion, Linguistic analysis, Machine Learning

1 INTRODUCTION

http: The emerging threats caused by terrorism, organized crime and cybercrime as interlinked cross-border challenges are showing how important a joint European answer to these threats is. Especially, the protection of so-called soft targets is a challenge for LEAs (Law Enforcement Agencies). In these complex cases, the investigators are also more and more confronted with huge amounts of data, which have to be analysed in a short time. The heterogeneous nature of these data streams forces the LEAs to link together and priorities in order to be able to understand and analyse them. PREVISION intends to improve LEAs operational capacities and capabilities by providing a unique and innovative platform. This platform is built by 28 consortium partners (IT companies, universities, research centers...) from 13 different European countries. Moreover, some results inherited from other PREVISION partners' projects.

2 PREVISION OBJECTIVES

Therefore, PREVISION has seven specific and measurable objectives (SO), which are described in Figure 1 and are as follows:

- SO-1: Deliver an open, scalable and customizable toolset that provides support for extreme-scale data streams analytics,
- SO-2: Semantically integrate heterogeneous data streams delivering powerful knowledge graphs combined with advanced reasoning and machine learning engines,
- SO-3: Configure and tailor situation awareness enabling techniques and applications to meet specific operational needs of LEAs and address human factors,
- SO-4: Integrate and deploy the developed functions and capabilities into a common platform architecture, making it available to end-users for thorough validation,
- SO-5: Demonstrate and evaluate the developed technologies in realistic cases, organize relevant training activities and create a framework for the transfer of knowledge in the use of PREVISION tools from one LEA to another,
- SO-6: Ensure compliance with the legal, ethical, privacy, societal and court-acceptance guidelines and EU best practices,
- SO-7: Ensure the high multi-dimensional impact, continuity and business perspective of project results and allow for incremental investments.

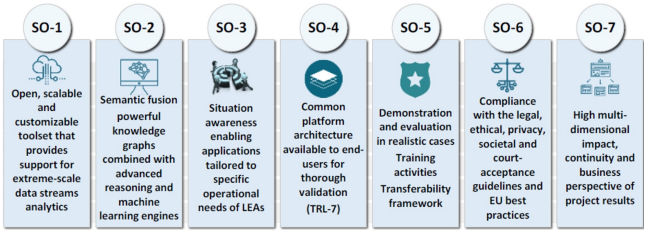

Figure 1: Overview of special objectives of PREVISION.

3 USE CASES AND DATA

Five use-cases have been developed. In each of them, the LEAs described a typical case, according to their interests. These use-cases will be the basis for any testing done within the framework of the project. When defining initial use-cases the LEAs also describe there currently implemented procedures and structures. This will be the basis for identifying problems and weaknesses of the current procedures. The reported timeline of the events will be useful to provide an overview of the particular data sources selected for the specific investigation that the PREVISION platform has to be able to analyse. LEAs also identified all end-user requirements for the

rt

development of the PREVISION platform. By distinguishing between the different requirement categories "Security", "Functional", "Operational" and "Communication", a complex picture of all needs will be compiled. To ensure interoperability of the requirements they have been prioritized using the MoSCoW methodology (Must have, Should have, Could have, and Won't have) [5]. The initial uses cases are briefly detailed in Table 1.

Table 1: Topics of the use cases defined by LEA partners.

	Topic
UC1	Soft targets protection –Attempted terrorist attack at stadium
UC2	Radicalization detection and terrorist threat prevention –Terrorist threats at EU summit
UC3	Financial crime investigation –Detection of fraudulent companies
UC4	Fighting cyber-enabled crime –CNP fraud as terrorist act facilitator
UC5	Illicit markets investigation –trafficking of cultural goods

The necessity of analyzing big data coming from diverse sources such as camera devices, deep web, dark web, etc. has become a big challenge in the field of security and are indeed necessary to develop the five use cases the project targets. These data sources are of three main types: Video surveillance cameras, Deep/Dark/Shallow web, Social Networks data. Indeed, the collection of the data that will be used by PREVISION's platform includes datasets crawled from deep/dark/shallow web. These data sets are textual-based pseudonymized data sets. PREVISION also considers visual content generated by CCTVs or video files as well as social network data.

4 HANDLING HETEROGENEOUS DATA

These heterogeneous data sources need to be managed and to be analyzed in a short time for the vast amount of data. NoSQL data stores are well-tailored to efficiently load and manage massive collections of heterogeneous data without any structural validation (shemaless principle). This flexibility becomes a serious challenge when querying data; i.e. users have both to build queries taking into account multi-structured datasets and reformulate existing queries whenever new structures are introduced. This also implies to set up modules for homogenizing the data search and analysis. Among them, we will develop a component following the approach developed by Ben Hamadou *et al.* [1] for building schema-independent queries, which is designed to query multi-structured datasets into NoSQL document stores such as MongoDB. This component automates the process of query reformulation via a set of rules that reformulate most document store operators (select, project, unwind, aggregate and lookup). The component then produces queries across multi-structured documents, which are compatible with the native query engine (MongoDB) of the underlying document store. The schema of this component is presented in the Figure 2.

Community detection and key actor identification framework is one of the tool of the PREVISION platform; preliminary result has already been proposed [2, 4]. PREVISION linguistic analysis is based on multiple entities and multiple languages. Social analytics services could be able to consider proposed linguistic features, as the outcome of the deep linguistic analysis.

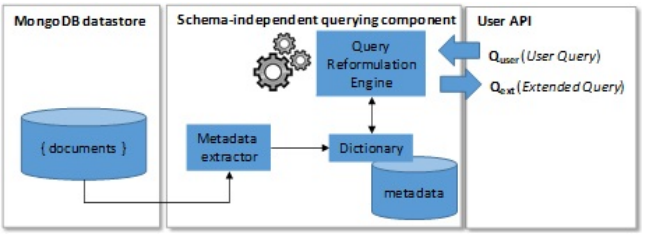

Figure 2: Schema independent querying component flow

5 CONCLUSIONS

PREVISION is a two year long project that gathers both LEAs and top level technical partners including IT companies and universities/labs. This helps to strengthen collaborations of experts from several disciplines. However, differences among viewpoints, or difficulty in seamless integration among partner's modules might happen during the project implementation.

The results of the project will be an open and future-proof platform which handle and capitalize on the massive heterogeneous data streams processed during complex crime investigations and threat risk assessments. This platform will provide cutting-edge practical support to LEAs in their fight against terrorism, organized crime and cybercrime. Results will be made publicly available for those that can be but also will serve the LEAs in their daily work. A workshop was organized earlier this year on related topics [3].

Ethical issues. in order to achieve its purpose PREVISION will process big amounts of heterogeneous data including personal data and carry out research with humans (interviews, surveys, workshops etc.). This arises key ethical issues. Moreover, there is a risk of misuse of research results for unethical purposes. These issues are carefully taken into account in the project.

Acknowlegments. This work has been performed in the context of the PREVISION project, which has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 833115. The paper reflects the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Hamdi Ben Hamadou, Faiza Ghozzi, André Péninou, and Olivier Teste. 2019. Schema-independent querying for heterogeneous collections in NoSQL document stores. Inf. Syst. 85 (2019), 48–67. https://doi.org/10.1016/j.is.2019.04.005
[2] Thi Bich Ngoc Hoang. 2020. Topical Community Detection: an Embedding User and Content Similarity Method. In [3]. 1–7.
[3] Thi Bich Ngoc Hoang, Pascal Marchand, Béatrice Milard, and Josiane Mothe. 2020. Workshop on Machine Learning for Trend and Weak Signal Detection in Social Networks and Social Media, Toulouse, France, Feb. 27-28, 2020, Proceedings.
[4] George Kalpakis and et al. 2019. Identifying Terrorism-Related Key Actors in Multidimensional Social Networks. In MultiMedia Modeling. 93–105.
[5] Eduardo Miranda. 2011. Time boxing planning: buffered moscow rules. ACM SIGSOFT Software Engineering Notes 36 (11 2011), 1–5.