CAPER: Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime

Carlo Aliprandi[1], Juan Arraiza Irujo[2], Montse Cuadros[2], Sebastian Maier[3], Felipe Melero[4], Matteo Raffaelli[1]

[1] Synthema srl, Pisa, Italy
{carlo.aliprandi,matteo.raffaelli}@synthema.it
[2] Vicomtech, San Sebastián, Spain
{jarraiza,mcuadros}@vicomtech.org
[3] Fraunhofer IGD, Darmstadt, Germany
sebastian.maier@igd.fraunhofer.de
[4] S21sec, Pamplona, Spain
fmelero@s21sec.com

Abstract. Law Enforcement Agencies (LEAs) are increasingly more reliant on information and communication technologies and affected by a society shaped by the Internet and Social Media. The richness and quantity of information available from open sources, if properly gathered and processed, can provide valuable intelligence and help drawing inference from existing closed source intelligence. This paper presents CAPER, a state-of-the-art platform for the prevention of organised crime, created in cooperation with European LEAs. CAPER supports information sharing and multi-modal analysis of open and closed information sources, mainly based on Natural Language Processing (NLP) and Visual Analytics (VA) technologies.

Keywords: Open Source Intelligence (OSINT), Focused Crawling, Social Web, Natural Language Processing (NLP), Named-Entity Recognition (NER), Semantics, Visual Analytics (VA).

1 Introduction

The revolution in information technology is making open sources more accessible, ubiquitous, and valuable. LEAs have seen open sources grow increasingly in recent years and most valuable intelligence information is often hidden in files which are neither structured nor classified. The process of accessing all these raw data, heterogeneous in terms of source, format and language, and transforming them into information is therefore strongly linked to multi-modal and multi-lingual data analysis and VA technologies with powerful Human Computer Interfaces.

CAPER is an open source intelligence (OSINT) platform that supports collaborative multilingual analysis of unstructured text and audiovisual contents (video, audio, speech and images). CAPER is not focused on the development of new technology, but on the fusion and real validation of existing state-of-the-art to solve current bottlenecks faced by LEAs.

Traditionally, text and data mining systems can be seen as specialised systems that convert raw data into a structured database, allowing people to find information. For some domains, text mining applications are well-advanced, for example in the domains of medicine, military and intelligence, and aeronautics [1]. In addition to domain-specific miners, general technology has been developed to detect named-entities [2], co-reference relations, geographical data [3], and time points [4].

Current baseline information systems are either large-scale, robust but shallow (standard information retrieval (IR) systems), or they are small-scale, deep, but ad hoc and maintained by experts in language technologies, not by people in the field (Semantic-Web and ontology-based systems). The table below gives a comparison across different state-of-the-art information systems (ad hoc Semantic Web solutions, WordNet based information systems and traditional information retrieval are compared with CAPER).

Table 1. Comparison of semantic information systems

Key features	*Semantic Web*	*WordNet based*	*IR*	*CAPER*
Large scale and multiple domains	NO	YES	YES	YES
Deep semantics	YES	NO	NO	YES
Automatic acquisition and indexing	NO	YES/NO	YES	YES
Multi-lingual	NO	YES	YES	YES
Cross-lingual	NO	YES	NO	YES

2 Interoperability and Central Management Application

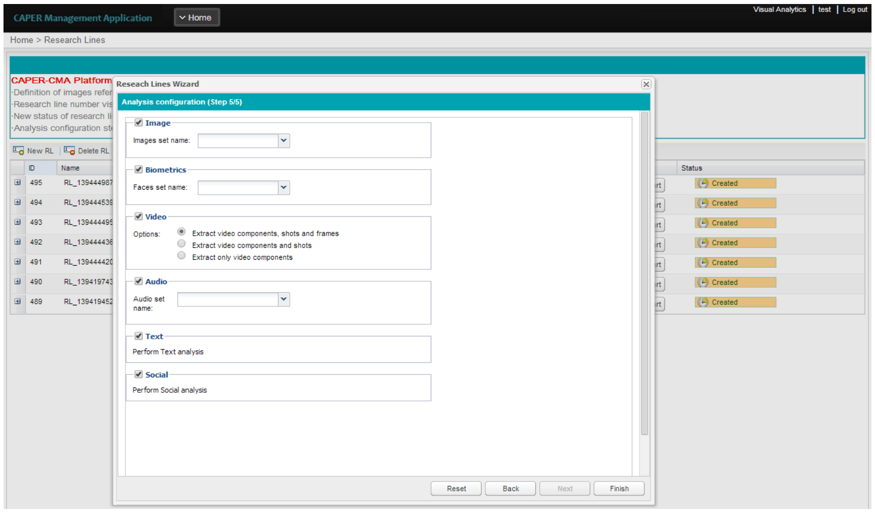
CAPER is a broad platform that aims at providing LEAs with integrated, advanced modules for capturing, analysing, storing and intelligently displaying large data sets collected from open sources. Both semantic and operational interoperability received specific attention throughout the design and development phases, as CAPER comprises multiple modules, systems and applications, developed globally by numerous partners, varying in expertise and focus.

Semantic interoperability and standardisation of information processing are guaranteed through the adoption of KAF (Knowledge Annotation Format) [5]. There have been numerous attempts to standardise different aspects of natural language processing [6][7][8][9]. KAF is a multi-layered XML format for the semantic annotation of unstructured text documents that has been proven to be suitable as data representation standard and has been extended within CAPER in order to be able to represent also multimedia audiovisual contents.

Operational interoperability is guaranteed through a service-oriented architecture (SOA). All system modules are called by an orchestrator, which executes the data collection and analysis workflow that the end-user has configured on the CAPER Central Management Application (CMA). The CMA is the end users' workbench, built on a web based collaborative platform, and is one of the two modules developed within CAPER with graphical user interface (the other being the VA application, described in §5). It allows LEA analysts to:

- Configure a Research Line (RL) by setting up a web crawling process.
- Configure the analysis modules. These may require parameterisation depending on the gathered content.
- Control the overall system by monitoring servers, services and processes and applying corrective actions.
- Manage system security (access control, authentication and authorisation).
- Obtain system reports: check the actions performed within the platform.
- Configure alerts among other features.

Figure 1. CMA - Configuration of the analysis modules


CAPER Management Application
Home
Visual Analytics | test | Log out
Home > Research Lines
CAPER-CMA Platform
·Definition of images refer
·Research line number vis
·New status of research li
·Analysis configuration st
New RL
Delete RL
Reseach Lines Wizard

ID	Name	Status
495	RL_139444987	Created
494	RL_139444539	Created
493	RL_139444495	Created
492	RL_139444436	Created
491	RL_139444420	Created
490	RL_139419743	Created
489	RL_139419452	Created

Analysis configuration (Step 5/5)

☑ Image
Images set name:

☑ Biometrics
Faces set name:

☑ Video
Options:
● Extract video components, shots and frames
○ Extract video components and shots
○ Extract only video components

☑ Audio
Audio set name:

☑ Text
Perform Text analysis

☑ Social
Perform Social analysis

3 Data Acquisition

The CAPER crawler is a multimedia content gathering and storing system, whose main goal is to manage huge collections of data coming from heterogeneous and distributed information sources. Text, audio and video content is retrieved via crawling of the worldwide web in three ways:

- Looking for documents in a given URL until a parametric depth of levels. A focused crawler has been developed. Users can specify key-words when setting up the crawling process. The crawler follows all the links of the web page and rejects the pages that don't contain the specified key-word.

- Looking for a parametric number of documents on the web with a key-words search. Users can specify queries that are redirected to the principal search engines in order to retrieve their results.
- Looking for a parametric number of pages on Facebook with a key-words search. The crawler is able to capture collaboratively created content and to retrieve specific information from Facebook, like users names and IDs, users networks based on "likes", friends networks.

4 Information Analysis

Whilst data are raw the interpretation of those data in a given context produces information. In OSINT solutions data are crawled from publicly available sources and then stored in normalised formats that are ready to be processed by the system. At this stage data are ready to be analysed. The analysis of the data can take many different forms depending on the purpose. The same data can be analysed with different approaches and therefore different information can be obtained out those same data. The results produced by a contextualised (focused) analysis of a set of data produces focused knowledge. In the CAPER project an OSINT platform has been built with the aim of producing valuable knowledge for the prevention of organised crime from publicly available data. For the multimedia data collected CAPER develops automated networks of entities and their relationships, one of the most (if not the most) important objectives of OSINT solutions when fighting crime [10].

When a RL is created on the CMA, end users can select the set of languages they want for the analysis of the text and audio files and they can upload a set of images and/or audio reference files as well. Then every time a new normalised text is sent for analysis the text analysis module will identify the language of that text and redirect it to the appropriate linguistic processor. Additionally, every time a new image and/or audio is crawled and normalised it will be compared against those reference files that the user might have configured when creating the RL by the corresponding analysis modules.

CAPER includes the following six analysis modules: (1) Image analysis, which compares crawled images and video frames with a set of reference images and/or classes of objects (images) and provides a similarity score; (2) Multilingual text analysis, which covers 13 different languages including Arabic, Basque, Catalan, Chinese, English, French, German, Italian, Japanese, Portuguese, Romanian, Russian, and Spanish; and which uses natural language processing techniques to identify entities and relationships among them; (3) Multilingual analysis of audio content so that it can be reduced to its base components for deeper analysis (i.e. text transcripts of voice, speaker recognition and tracking, gender and age identification); (4) Analysis of videos so that they can be reduced to their base components for deeper analysis (i.e. scenes, frames, audio); (5) Integration of semantic-Web technologies and data to improve and relate analysis results (e.g. in the Named Entity Recognition process) and analysis of data coming from Social Media; and (6) Biometric analysis, which includes face recognition and speaker identification.

5 Visual Analytics Application

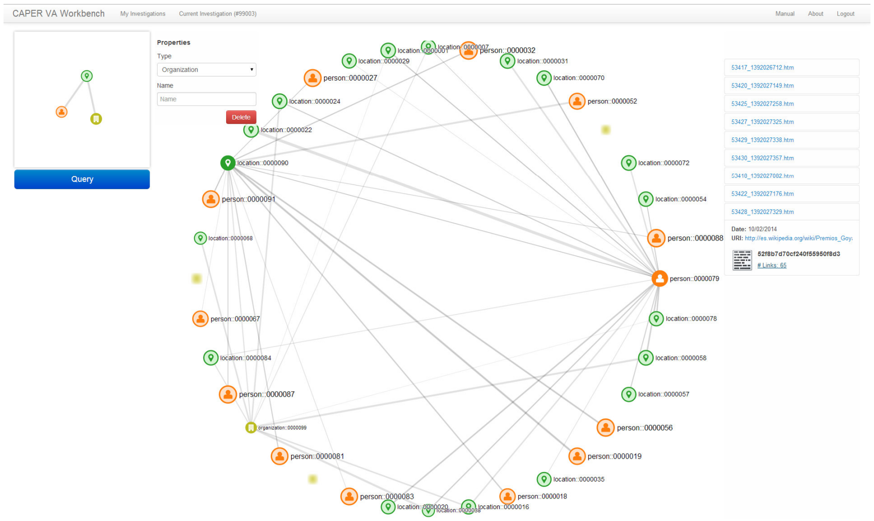
Large information spaces like the one created by the CAPER Information Analysis module call for a suitable way to explore, drill-down and analyse this information [11]. The CAPER VA Application provides several ways to access the information, including an overview of the different cases currently assigned to the analyst. It also provides an overview of the documents which have been collected for a single case. Here the analyst can access the contents of a single document, whether it is a text, audio, video or image document with recognised entities being highlighted.

Before the data are presented to the human analyst we create an integrated model of all entities and their relationships as they are in the analysed documents. Relations between entities are either provided by the Information Analysis Module or are created based on co-occurrence by the VA Application. Each relationship is given a certain degree of credibility. For text documents this degree of credibility is currently based on the distance of the entities in the text but could also be based on the actual syntactic or semantic relationship as expressed within the text.

The social graph can be explored using the VA Browser and Editor to manually select those parts of the graph which are of interest to the investigator. Here it is also possible to annotate the graph with custom relations and entities. This is used to enhance the automatic model with knowledge coming from other sources like an interrogation.

Another way to access the information space is the search for specific patterns in the data. For this use-case we provide a visual query interface to define graph patterns which can then be searched within our database. Search results are then visualised using a circular graph layout, showing all entities being part of the specified pattern.

Figure 2. VA Application – Visualisation of entity relationships



Properties

Type: Organization

Name: Name

53417_1392026712.htm
53420_1392027149.htm
53425_1392027258.htm
53427_1392027325.htm
53429_1392027338.htm
53430_1392027357.htm
53410_1392027082.htm
53422_1392027176.htm
53428_1392027329.htm

Date: 10/02/2014
URI: http://es.wikipedia.org/wiki/Premios_Goy

52f8b7d70cf240f55950f8d3
Links: 65

6 Conclusions

This paper has presented CAPER, a state-of-the-art OSINT platform for the prevention of organised crime, created in cooperation with European LEAs. LEAs have special intelligence analysis units to support extended investigations against organised crime. In these units analysts might work for months or even years on specific investigations. Today the intelligence cycle is characterised by manual collection and integration of data. CAPER supports the automatic collection and analysis of unstructured text and audiovisual contents (video, audio, speech and images) and develops automated networks of entities and their relationships. These networks are automatically integrated into LEAs' systems, thus drastically reducing data integration efforts for intelligence analysts.

Acknowledgments. This work is partially funded by the European Commission (CAPER Project, 261712 FP7-SECURITY SEC-2010.1.2-1).

References

[1] Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics, I, pp. 466–471. Kopenhagen (1996).

[2] Hearst, M. A.: Untangling Text Data Mining. In: Proceedings of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics, pp. 123-129. College Park (1999).

[3] Miller, H. J., Han, J. (eds.): Geographic Data Mining and Knowledge Discovery, 2nd edition. London (2009).

[4] Wei, L., Keogh, E.: Semi-Supervised Time Series Classification. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 748-753. New York (2006).

[5] Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Aliprandi, C., Monachini, M.: KAF: a generic semantic annotation format. In: 5th International Conference on Generative Approaches to the Lexicon, pp. 157-164. Pisa (2009).

[6] Clément, L., Villemonte de La Clergerie, É.: Maf: a morphosyntactic annotation framework. In: Proceedings of the 2nd Language & Technology Conference, p. 90-94. Poznań (2005).

[7] Declerck, Th.: Synaf: Towards a standard for syntactic annotation. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) Proceedings of the 5th Conference on International Language Resources and Evaluation, pp. 229-233. Genova (2006).

[8] Ide, N., Romary, L.: Outline of the international standard linguistic annotation framework. In: Proceedings of ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right, p. 1-5. Sapporo (2003).

[9] http://semantic-annotation.uvt.nl

[10] Army, U. S. (2012). Open-Source Intelligence ATP 2-22.9 (p. 91). Retrieved from http://www.fas.org/irp/doddir/army/atp2-22-9.pdf

[11] Keim, D., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in Visual Data Analysis. In: 4th International Conference on Information Visualisation, pp. 9-16. Washington, DC (2006).