An Automated System for Criminal Police Reports Analysis

Gonçalo Carnaz[1,2], Vitor B. Nogueira[1,2], Mário Antunes[3,6], and Nuno Ferreira[5,6]

[1] Informatics Departament, University of Évora, Portugal
[2] LISP, Portugal
d34707@alunos.uevora.pt and vbn@di.uevora.pt
[3] School of Technology and Management, Polytechnic Institute of Leiria, Portugal
mario.antunes@ipleiria.pt
[4] Institute of Engineering of Coimbra, Polytechnic Institute of Coimbra, Portugal
nunomig@isec.pt
[5] GECAD, Institute of Engineering, Polytechnic Institute of Porto, Portugal
[6] INESC TEC, Portugal

Abstract. Information retrieval and fusion are complex fields and have been applied in several domains to deal with heterogeneous data sources. Criminal polices are challenged in forensics activities with the extraction, processing and interpretation of several documents from different types and with distinct formats (templates), such as narrative criminal reports, police databases and the result of OSINT activities, just to mention a few. Such challenges implies, among others, to cope with and manually connect some hard to interpret meanings such as license plates, addresses, names, slang and some emotions. In this paper we aim to deal with forensic information retrieval and fusion, for that, we present a system that automatically extract, transform, clean, load and connect police reports that arrived from different sources. The same system aims to help police investigators to identify and correlate interesting extracted entities.

Keywords: Information fusion, forensics, ETL, criminal police reports.

1 Introduction and Motivation

Criminal police departments deal with forensic information from heterogeneous data sources during a crime investigation, such as narrative police reports, crime scene reports, spreadsheets, police databases or Open Source Intelligence (OSINT) [1], that generate a deluge of data. In daily activities police investigators manually create criminal police reports that describe the crimes investigated with relevant forensics assets to be analyzed. Therefore, structured, semi-structured and unstructured data produced could benefit from an information fusion approach to deal with heterogeneous police data sources, and to support a forensic

[1] Data collected from publicly available sources to be used in an intelligence context.

decision making system. All issues regarding information fusion must be analyzed and solved in order to be applied appropriated computational methods to extract relevant information.

The Portuguese Internal Security Report [1] compiles all crimes investigated and reported by the institutions that compose the Internal Security System. According to the 2016 edition of such report, in that year there were 330,872 investigated crimes. Considering that for each crime the police investigators must not only produce several reports and forensic evidences but also analyze and process such elements, this leads to an extremely time-consuming task for the police forces. Therefore, our motivation is to resort to a computational approach to help in analyzing and processing the vast amount of police reports.

We focus our study in one Police Institution and their criminal police reports, with different file types and formats (templates). Therefore, our contribution in this paper is to present a system that automatically extract, transform, clean, load and connect police reports that arrived from different sources, and additionally aims to help police investigators to identify and correlate interesting extracted entities.

The rest of the paper is organized as follows: in section 2 we describe the related works useful to support our work; the section 3 details the system by which we proposed to extract, transform, clean and load the narrative police reports into a common format and finally in section 4, we expose our conclusions and future work.

2 Literature Review

The motivation for our work is two-fold: *"how to extract, transform and load relevant information?"* and *"how to identify forensic information from police reports?"*. Therefore, we have investigate literature with related works for: Extract,Transform and Load (ETL) approaches and the named-entity recogniton (NER) systems for Portuguese language, that could help to create a information fusion environment.

ETL concept is defined as *"The ETL process extracts the data from source systems, transforms the data according to business rules, and loads the results into the target data warehouse."* [1]. There are several proposed approaches, e.g. from a real-time data ETL framework [2] that processes historical data and real-time data separately; or a technique for data streams handling that synchronizes data streams from incoming data sources [3]; or the proposal of a Integration Based Scheduling Approach (IBSA) that deals with data syncronization issue [4]; using a ontology-based methodology approach to resolve homogeneity regarding data sources and the integration of data by its meaning [5]; [6]using a domain ontology (stored inside a data warehouse as metadata) and all findings in data sources are semantically analyzed, and finally [7] proposed an ontology-based approach to support the conceptual design of the ETL processes, based on a

[1] www.ansr.pt

graph-based representation, to support structured and semi-structured data extraction. There are several ETL tools, from commercial tools, e.g. IBM Infosphere, Oracle Warehouse Builder or Informatica Powercenter and Open Source tools, e.g. Talend OpenStudio, Pentaho Kettle or CloverETL that enables the development of ETL tasks.

Regarding NER systems, Nadeau et al. [8] defined NER as *"...is a sub problem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies..."*.

In 2014, Dozier et al. [9] proposed a NER system to extract entities from legal documents, e.g. judges, companies, courts or others; Arulanandam et al. [10] proposed a system to extract crime information from online newspapers is proposed, focused in the "hidden" information related to the theft crime.

Shabat et al. [11] proposed in 2015 a system for crime information extraction from the Web, with a NER task, using classification algorithms, like Naive Bayes, Support Vector Machine and K-Nearest Neighbor. An indexing module was added to crime type identification, using the same classification algorithms. Yang et al. [12] proposed an approach based on raw text to extract semi-structured information using text mining techniques.

Bsoul et al. [13] proposed in 2016, a system to extract verbs and their use, using two datasets: a real datasets from crime and a industrial datasets with benchmarks. Additional, the Porter stemming algorithm for word stemm was used, for verbs identification.

In 2017, Schraagen et al. [14], proposed a NER system, named as *Frog*, using an manually annotated corpus, created from 250 criminal complaints reports, where domain experts identified: entities, like location, person, organization, event, product and others. Al-Zaidy et al. [15] proposed several methods to be applied on discover criminal communities, analyzing their relations, and extract useful information from criminal text data.

3 Proposed Approach

Figure 1 illustrates a high-level design of our system proposal for information fusion related with criminal police reports analysis to achieve a path for reducing police investigators time-consuming and accuracy on forensic activities.

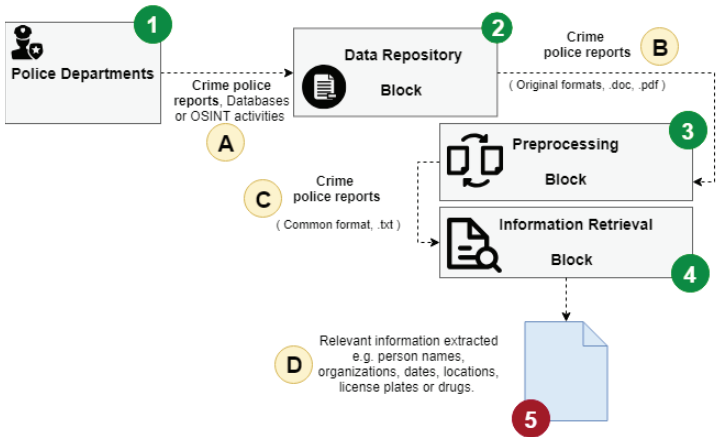

Fig. 1. Criminal police reports analysis - a system for information fusions.

Our approach is divided into four main blocks: (1) *Police Departments*, represents the police investigations heterogeneous data sources, which produces all criminal police reports that will be processed by the following blocks; in (2) *Data Repository Block*, we centralize all criminal police reports in their different formats and types, and originated from each police department; the (3) *Preprocessing Block*, responsible for data extract, parse, clean, filter and integrate into a common format; (4) *Information Retrieval Block*, aim to retrieve relevant information related to crime investigations from criminal police reports, e.g. as license plates, addresses, names, slang and emotions. Finally, the (5) represents the file which the relevant information for police decision making.

3.1 Police Departments and Data Repository Block

The (A) and (B) represents the data flow between blocks. In this case, the criminal police reports in their original format, e.g. Microsoft™ Word file type. The table 1 shows: the number of pages, words, characters, paragraphs and lines for each four criminal police reports analyzed.

Police Reports	Pages	Words	Characters	Paragraphs	Lines
Police01	14	4866	24883	174	458
Police02	3	516	2763	66	120
Police03	32	11811	71864	310	864
Police04	47	16678	85099	484	1307

Table 1. Criminal police reports summary.

The (2) *Data Repository Block* works as a staging area for non-processed documents. We need this block, because our sources are from different police departments (narcotics, homicides or economic crime) and this is a confluent point to archive the police criminal reports.

3.2 Preprocessing Block

Figure 2 depicts the *Preprocessing* block that aims to propose a system that automatically extract, transform, clean, load and connect criminal police reports that arrived from different sources into a common format. To that purpose we have defined a group of ETL tasks that could decrease the "dirty" data and facilitate the retrieval of police information assets, such as crime types, criminals name, locations, vehicles, and so on. Therefore, the ETL proposed are divided into: the (1) *Extract/Parse* task aims to extract and parse the criminal police reports from it's original format into a common format (.txt); in (2) *Data Preparation* task we proposed to normalize the extracted text, following (A) Regex (Regular Expressions) rules based on domain lexicon for text normalization, and cleaning rules; in (3) *Data Loading* task aims to load the normalized file into a target (4) Data Staging Area.

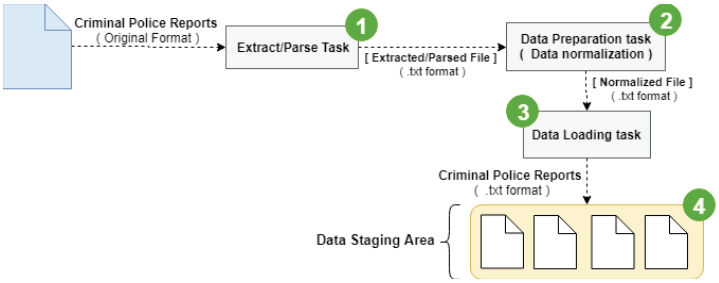

Fig. 2. Criminal police reports analysis - our ETL proposal.

To developed the proposed system, we used Java [1] programming language and other toolkits, namely Apache™ Tika [2] used to extract and parse documents from different formats, e.g. word and pdf files formats.

Extract and Parse We extract all data from criminal police reports, not any subset of interest, assuming that the relevant data identification will be done at later in time. To execute this task the following features have been added: (a) File formats identification; (b) Extract data, fetching data with appropriated connector to the external sources; (c) Parse data, the parsing task was performed

[1] https://docs.oracle.com/javase/7/docs/technotes/guides/language/
[2] https://tika.apache.org/

by reading the data stream and building an in-memory model to facilitate the data transformation. In our case, and during the parse task, we add two different tags: <PLAIN TEXT> *Unstructured data*</PLAIN TEXT>, that identifies unstructured data, and <TABLES> *Semi-structured data*</TABLES>, that identifies semi-structured data.

Police Reports	Words	Characters	Lines
Police01	6187	36314	254
Police02	1110	6113	293
Police03	11901	71906	326
Police04	18083	108719	572

Table 2. Criminal police reports after extraction/parsing task summary.

Table 2 shows the results obtained after the extract and parse tasks, each criminal police report were transformed into a common file format, like text document. We choose this format, because: reduced formatting, easy to read in every text editor and manipulate in any programming language, and also the file size is reduced.

Data Preparation We build a *Data Preparation* task supported by different sub-tasks: cleaning, remove duplicates, filters and transformation. We developed our tasks with a set of regular expressions, that could be updated along process for a more accurate data preparation and normalization. Therefore, text analysis needs some tasks to increase the regularity, e.g. each sentence contains end-mark, commas are followed by a space, cleaning double spaces, replace nulls or adding structure, e.g. splitting extracted text into different sections that have a special meaning, such as witnesses or suspects. The filter and transformation tasks, reads the extracted text seeking for two tags: <PLAIN TEXT> *Unstructured data*</PLAIN TEXT> and <TABLES> *Semi- structured data*</TABLES>. The desired outcome is to detect and separate the unstructured data (plain text) from the semi-structured data (tables). Figure 3 shows how the narrative police report will be after *Data Preparation* task.

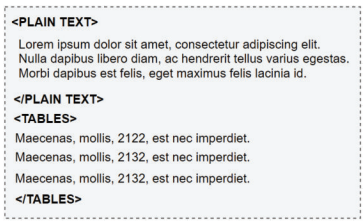

Fig. 3. Portuguese narrative police reports after data preparation.

Data Loading and Data Staging Area We proposed in this phase to populate a target location, named by *Data Staging Area*, with a common file format and information from police investigations. from a set of information sources and after sequential grouped tasks, e.g. Extract/Parse and Data Preparation, controlled by a set of rules. Therefore, we have the *Data Staging Area* were all documents are stored to be processed by the natural language processing module, after all preprocessing tasks.

3.3 Information Retrieval Block

The *Information Retrieval* block looks for relevant information from criminal police reports. Then the information obtained integrating these forensic information by identifying in each criminal police report relevant information. Figure 4 represents our approach based on a named-entity recognition (NER) system that recognize relevant named-entities in our criminal police reports, e.g. Persons, Locations, Organization and Time/Date.

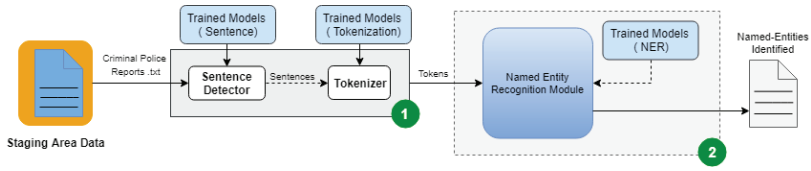

Fig. 4. Our named-entity recognition (NER) system proposal.

To retrieve named-entities from our dataset, we have implemented the following modules: (1) a module that detects sentences and tokens (words), called Sentence Detector, that aim to detect sentences in each text, using a trained model for Portuguese language. The chosen model use a machine learning algorithm, called Maxent [16] already trained and downloaded [1]. The Tokenization module perform tokens (words) detection, and also uses a trained model retrieved from the same website, also using the Maxent algorithm (the pre-defined algorithm in both cases).

The (2) Named-Entity Recognition module detects named-entities, such as persons, places, organizations and dates. To detect named-entities. we have trained a corpus, called Amazonia Corpus [2], that have 4.6 millions of words (about thousand sentences) retrieved from Overmundo [3] website, written in Portuguese-Brazilian language, and automatic annotated by PALAVRAS [17].

To train the Amazonia corpus, we realize the following steps:

[1] http://opennlp.sourceforge.net/models-1.5/
[2] http://www.linguateca.pt/floresta/ficheiros/gz/amazonia.ad.gz
[3] http://www.overmundo.com.br/

- Download the Apache OpenNLP [1] toolkit;
- Use the TokenNameFinderConverter tool to convert the Amazonia corpus into OpenNLP format;
- To train the model, we used the TokenNameFinderTrainer tool, with the predefined Maxent algorithm to generate the model, named by ner-amazonia.bin, that will be used as the trained model in our NER system.

3.4 Obtained Results

To perform the experiments, we obtained a dataset with four documents that are police reports from a real process, described in table 1. We have select two other systems for dataset evaluation based on the following criteria: (1) multilingual support; (2) focused on Portuguese language; (3) Open source tools. The systems selected were:

- Linguakit: a multilingual toolkit for natural language processing with a NER system incorporated, created by the ProLNat@GE Group [2] (CITIUS, University of Santiago de Compostela);
- RAPPort - A Portuguese Question-Answering System [18]: that uses a NLP pipeline with a NER system;

Table 3 shows the results obtained. We have used information retrieval metrics [19] to measure NER systems performance, namely P : Precision, R : Recall and $F - Measure$. *Precision* is defined by the ratio of correct answers (True Positives) among the total answers produced (Positives),

$$P(Precision) = \frac{TP}{TP + FP}$$

where TP - *True Positive*, a predicted value was positive and the actual value was positive and FP - *False Positive*, predicted value was positive and the actual value was negative [20].

R - Recall is defined as a ratio of correct answers (True Positives) among the total possible correct answers (True Positives and False Negatives),

$$R(Recall) = \frac{TP}{TP + FN}$$

where FN - *False Negative*, a predicted value was negative and the actual value was positive [20].

$F - Measure$ - is a harmonic mean of precision and recall,

$$F\text{-}Measure = \frac{2 * precision * recall}{precision + recall}$$

[1] https://opennlp.apache.org/
[2] https://gramatica.usc.es/pln/

	Person			Organization			Place			Date/Time		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Our proposal	0.98	0.93	0.95	0.73	0.27	0.39	0.98	0.97	0.96	0.95	0.91	0.93
RAPPort	0.91	0.96	0.94	0.33	0.33	0.33	0.92	0.94	0.93	0.86	0.98	0.92
Linguakit	0.67	0.28	0.39	0.12	0.82	0.21	0.50	0.78	0.60	0.90	0.90	0.90

Table 3. Test evaluation results.

Globally, our system have reached the highest F-measure result for each entities for the detected entities, having the best trade-off regarding both measures (precision and recall). The results obtained were quite better, because we have cleaned and transformed the used dataset, before been processed by the NLP pipeline, which removes some existent entropy from dataset that are passed to the NER systems.

4 Conclusion and Future Work

The work developed in this paper tries to achieve a forensic information fusion for criminal police reports retrieved from heterogeneous data sources. we have focused on the obtained results, mainly in the *Information Retrieval* block as it allow us to prove the necessity of a process to integrate relevant information from police heterogeneous sources and associated information security. Therefore, we achieve relevant results about information extraction regarding named-entities, e.g. person, place, organization and date, for criminal analysis and to support a decision making system for police investigations. For future work, we define the following goals:

- To integrate information from other Police Institutions, where criminal police reports follow different document templates;
- To apply a machine learning algorithm to recognize "dirty" data without human intervention and correct data errors or anomalies;
- To improve our natural language processing system in Portuguese, with a named-entity recognition task, such as adding relation extraction, adding new named-entities, e.g. vehicles, license plates or drugs;

References

1. Kamal Kakish and Theresa A Kraft. Etl evolution for real-time data warehousing. In *Proceedings of the Conference on Information Systems Applied Research ISSN*, volume 2167, page 1508, 2012.
2. Xiaofang Li. Real-Time Data ETL Framework for Big Real-Time Data Analysis. In *IEEE Int. Conf. Inf. Autom.*, number August, pages 1289–1294, 2015.
3. F. Majeed, Muhammad Sohaib Mahmood, and M. Iqbal. Efficient data streams processing in the real time data warehouse. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 57–61, July 2010.

4. J. Song, Y. Bao, and J. Shi. A triggering and scheduling approach for etl in a real-time data warehouse. In *2010 10th IEEE International Conference on Computer and Information Technology*, pages 91–98, June 2010.
5. J. Villanueva Chávez and X. Li. Ontology based etl process for creation of ontological data warehouse. In *2011 8th International Conference on Electrical Engineering, Computing Science and Automatic Control*, pages 1–6, Oct 2011.
6. L. Jiang, H. Cai, and B. Xu. A domain ontology approach in the etl process of data warehousing. In *2010 IEEE 7th International Conference on E-Business Engineering*, pages 30–35, Nov 2010.
7. Dimitrios Skoutas and Alkis Simitsis. Ontology-based conceptual design of etl processes for both structured and semi-structured data. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(4):1–24, 2007.
8. David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
9. Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Semantic processing of legal texts. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts*, chapter Named Entity Recognition and Resolution in Legal Text, pages 27–43. Springer-Verlag, Berlin, Heidelberg, 2010.
10. Rexy Arulanandam, Bastin Tony Roy Savarimuthu, and Maryam A. Purvis. Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference - Volume 155*, AWC '14, pages 31–38, Darlinghurst, Australia, Australia, 2014. Australian Computer Society, Inc.
11. Hafedh Ali Shabat and Nazlia Omar. Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East J. Sci. Res.*, 23(6):1215–1221, 2015.
12. Y. Yang, M. Manoharan, and K. S. Barber. Modelling and analysis of identity threat behaviors through text mining of identity theft stories. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 184–191, Sept 2014.
13. Qusay Bsoul, Juhana Salim, and Lailatul Qadri Zakaria. Effect Verb Extraction on Crime Traditional Cluster. *World Appl. Sci. J.*, 34(9):1183–1189, 2016.
14. Marijn Schraagen. Evaluation of Named Entity Recognition in Dutch online criminal complaints. *Comput. Linguist. Netherlands J.*, 7:3–15, 2017.
15. Rabeah Al-Zaidy, Benjamin C. M. Fung, and Amr M. Youssef. Towards discovering criminal communities from textual data. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 172–177, New York, NY, USA, 2011. ACM.
16. Rahul Sharnagat. Named entity recognition: A literature survey. *Center For Indian Language Technology*, 2014.
17. Eckhard BICK. *The parsing system "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 412 f.* PhD thesis, Thesis (PhD)-Aarhus University, Denmark University Press, 2000.
18. Ricardo Rodrigues and Paulo Gomes. Rapport—a portuguese question-answering system. In *Portuguese Conference on Artificial Intelligence*, pages 771–782. Springer, 2015.
19. Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. Named Entity Recognition Approaches. *Journal of Computer Science*, 8(2):339–344, 2008.
20. Ing Michal Konkol. *Named Entity Recognition*. PhD thesis, University of West Bohemia, 2015.