Drug-Related Crime Information Extraction and Analysis

Khmael Rakm Rahem
Center for AI Technology, FTSM
University Kebangsaan Malaysia,UKM
43000 Bangi Selangor, Malaysia
khm2006_rakem@yahoo.com

Nazlia Omar
Center for AI Technology, FTSM
University Kebangsaan Malaysia,UKM
43000 Bangi Selangor, Malaysia
nazlia@ukm.edu.my

***Abstract*—Although valuable crime information is available in human-readable form in online newspapers and electronic archives, software systems that can extract and present relevant information are limited and are of significant interest to researchers in the field of information extraction. This work aims to extract available drug crime information from online newspaper articles. This work has the following subtasks: assess where and how drug traffickers hide drugs, identify the nationalities of drug dealers, identify the types (names) of drugs, and assess the quantity and prices of drugs in the local market. This paper presents a rule-based approach to extract information on the basis of a set of drug crime gazetteers and on a set of grammatical and heuristic rules. This work is validated through experiments. Results show that the technique developed here are promising.**

Keywords—Drug crimes; Information extraction; Rule-based Approach; Grammatical and heuristic rules

I. INTRODUCTION

Information extraction (IE) aims to extract valuable information from unstructured data. For example, the names of people, locations, organizations, and miscellaneous data (date, time, percentage, and monetary expressions, etc.) can be extracted from documents without a "deep understanding" of the text. IE techniques have been used for many different purposes such as for extracting gene or protein names and determining their relationships. With the advent of the Internet, huge volumes of data have become available online. Electronic newspapers have been increasingly read by users from anywhere at any time. Newspapers are a source of (mostly) authentic and timely information. A large amount of information is available in newspaper articles.

Valuable crime information is available in human-readable form in online newspapers and electronic archives. However, software systems that can extract and present relevant information are scarce and have been of significant interest to researchers in the field of IE. Although search engines can be used to search for crime information, a manual that could read through the results is needed to extract valuable data. This process is tedious and prone to errors. In drug crimes, drug abuse is defined as the consumption of a particular substance that may provide particular effects to a person. The chronic use of such substance may cause physical and psychological dependence. Drug crimes have been increasing every year. Hence, drug crimes are one of the significant challenges faced by a community. Drug trafficking has been an increasing concern for international security studies.

We believe that a research on crime domains that specifically focuses on drug crimes would be a success. This work is mainly motivated by the need to develop a successful and useful IE system for crimes. Developing a tool that helps the community and enforcement agencies to extract drug-related crime information is significantly needed. This tool should filter the vast amount of knowledge available in the web, use indexed reliable sources such as online newspapers and blogs, and extract drug-related information. Thus, this study primarily aims to extract and analyze available drug crime information in online newspaper articles. This work contributes the following: (1) develops several drug-dependent specific linguistic resources such as drug-crime-specific lexicons and gazetteers; (2) develops a rule-based extraction tool to assess where and how drug traffickers hide drugs, identify the nationalities of drug dealers and the types (names) of drugs, and determine the quantity and prices of drugs in the local market.

The remainder of this paper is organized as follows. Section II reviews the related work in the area of crime domain. Section III describes the methodology and different key techniques and approaches. Section IV presents the experiment setup and discusses the experimental results. Finally, Section V concludes our work and discusses the future research directions.

II. RELATED WORK

This section highlights recent work that extract crime information from unstructured documents. Coscia and Rios [2] develop a tool that uses web content to obtain quantitative information on the mobility and modus operandi of criminal groups. Unambiguous query terms and the Google search engine are used to identify the areas of operation of criminal organizations and extract information about the particularities of their mobility patterns. This method is applied to Mexican criminal organizations to identify their market strategies, preferred areas of operation, and evolution over the last two decades. The results provide evidence that criminal organizations are more strategic and operate in more differentiated methods than those suggested by current academic literature.

978-1-4799-5423-0/14/$31.00 ©2014 IEEE

Alkaff and Mohd [3] evaluate the direct and indirect extraction of nationality from crime news. Additional references are used to identify the nationalities of suspects, victims, and witnesses. Their model is based on gazetteers and rule-based extraction, as well as co-reference resolution to link the references. The proposed approach is evaluated and compared with a manual extraction system. Alsharef, Omar, and Albared [4] present a named entity recognition (NER) system from Arabic crime news. Several named entities including names of persons, organizations, and locations are used. Their model is based on several gazetteers and grammatical rule-based extraction.

Ku, Iriberri, and Leroy [5] develop a crime interviewing system to collect information from victims and witnesses. Natural language processing (NLP) is used to extract crime information from police reports, newspaper articles, and narrative crime reports of victims and witnesses. Data are collected to develop lexicons (wide range of needed terms such as weapons, vehicles, scenes, clothes, shoes, and physical features) from several sources. The first source is the official crime information from Uniform Crime Reports. The second source is the encyclopedia of information from resources, including Wikipedia and MSN Encarta, which provide many main and sub-categories of vehicle and weapon information. The third source comprises general web sites and blogs. The system leverages the GATE components to extract crime-related information from police and witness narratives. Ito [6] develops LonMaps, which is an information system for crime and accident mapping based on news articles, to extract specific information from news articles. Four information items such as incidents, places, occurrence dates, and persons are obtained from a news article to establish the system. These items are captured by using thesaurus and patterns. Dates, places, and personal names are captured on the basis of sentence patterns. A thesaurus consisting of two types of terms, namely, daily and legal terms, is used to capture incidents. Daily terms are used in daily life, whereas legal terms are used in legal situations. Places, dates, and personal names are extracted on the basis of typical news report patterns. Finally, experiments are performed by using LonMaps to evaluate its effectiveness on processing queries and extracting information. De Bruin et al. [7] apply clustering techniques to analyze criminal careers. Four important factors (crime nature, frequency, duration, and severity) are identified on the basis of their analysis. These factors have been used to create criminal paroles, compare each criminal with other criminals by using a new distance measure, and cluster similar criminals. Data are obtained from the Dutch National Criminal Record database.

Jaiswal et al. [8] develop an online crime report and managing system that is accessible to the public, police department, and administrative department. The system is intended for use in a community to help residents easily interact with each other and encourage the reporting of suspicious behavior. This system registers online complaints from people and will help police departments in capturing criminals. In this system, any person can file a complaint any time.

Arulanandam, Savarimuthu, and Purvis [9] recently propose an IE model that focuses on extracting information for one type of crime, namely, theft, and then extracts the crime location. The theft-related information is extracted from newspaper articles from three different countries, namely, New Zealand, Australia, and India. The model uses NER to determine whether a sentence contains the crime location. The model focuses on extracting crime location from newspaper articles. The approach is conditional random field, which is a machine learning method for checking whether a sentence shows the crime location. However, the main drawback of this model is considering the crime location only.

For monitoring crimes, the Wiki Crime projects [10, 11] allow individuals to report crime details online; hence, other users can use this information to make decisions. However, a limitation of this approach is the difficulty of verifying the authenticity of the posted crimes.

Literature shows that the common entities extracted in crime domains include the names of victims, weapons, and drugs, which could be used to achieve effective results. This work proposes a new model that uses rule-based approaches in drug crimes to extract the type, amount, and price of drugs, as well as the hiding patterns and nationalities of suspects, from unstructured crime news documents. The model is tested and evaluated and is expected to achieve relatively high evaluation metrics to analyze applications involving drug-related crimes.

III. METHODOLOGY

This section provides the general methodology for extracting and analyzing drug-related crime information from newspaper articles. Several techniques are adopted in each phase to extract related entities. As shown in Fig. 1, the proposed methodology consists of the following main tasks:

- Language resource planning and compiling
- Preprocessing
- Extraction of drug crime entities
- Evaluation

A. Language Resource Planning and Compiling

In this step, the dataset is collected from the Malaysian National News Agency(BERNAMA). The news articles contain drug crime documents. To our knowledge, no drug crime ontology or vocabulary is readily available to researchers.

We used different strategies to develop several lexicons. The drug information from drug crime documents is analyzed, and a set of resources is developed:

- Drug list (DL): This list contains the names of known drugs such as heroin, ketamine, and syabu.

The indicator words (IW) lists play a central role in the development of the rules. The text is first used as an input to the system that will perform the preprocessing process. IW and part-of-speech (POS) can be used to design the rules to identify drug entities.

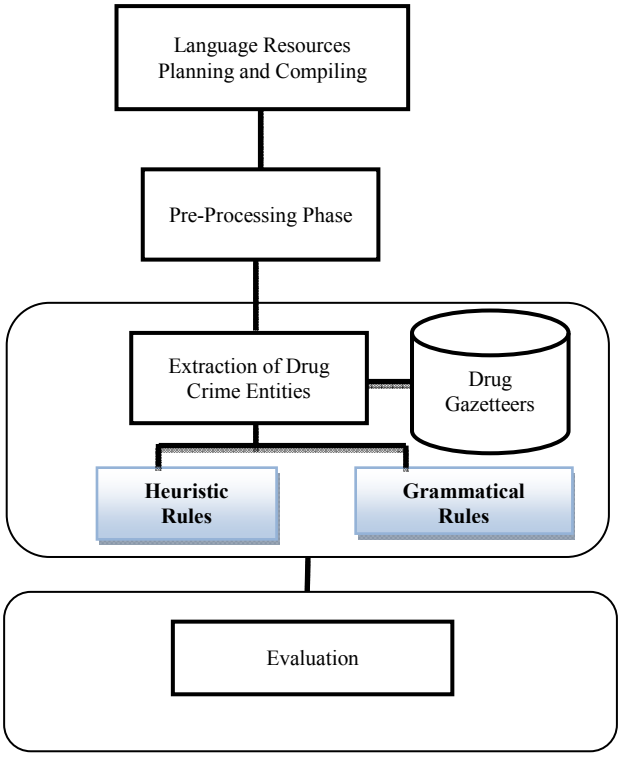

Fig. 1. Drug-related crime IE and analysis

Various IW lists are described as follows:

- Drug IW (DIW): The DIW list contains special words that are identified as introducing drug types. This list includes words such as arrested for, charged with, and possessing.
- Price IW (PIW): The PIW list contains special words that introduce drug prices. This list includes words such as worth RM, worth some RM, and worth about RM.
- Quantity IW (QIW): The QIW list contains special words that introduce drug quantities. This list includes words such as gram, kilogram, and packets.
- Hidden position IW (HPIW): The HPIW list contains special words that introduce the hidden position of drugs. This list includes words such as beneath the carpet, black plastic bags, and biscuit tins.

B. *Preprocessing*

In this phase, the data collected are pre-processed by using several NLP preprocessing tools. The Stanford NLP Tool is used to build our IE modules. The Stanford NLP Tool is an open-source library that comprises several modules. The modules adopted include tokenizer, sentence splitter, POS tagger, and parser.

The preprocessing procedure consists of the following tasks:

- Sentence splitter and tokenization: The text is split into several sentences. Subsequently, the tokenizer splits the text into tokens such as punctuation, words, numbers, and symbols.
- POS tagging: Each token is annotated with its POS tag, which is a grammatical tag (e.g., verb, noun, or adjective). Fig. 2 shows the output obtained from the Stanford tagger.
- Noun phrase chunking: The sentences are parsed with a custom chunk parser that identifies prepositions, nouns, and verbs. The so called core phrases in the drug crime text will then be extracted.

```
Asmabi/NNP said/VBD to/TO find/VB Zuhaila/NNP
and/CC Rayes/NNP guilty/JJ the/DT court/NN
had/VBD to/TO be/VB convinced/VBN that/IN
the/DT drug/NN was/VBD in/IN their/PRP $
possession/NN and/CC control/NN at/IN the/DT
time/NN of/IN the/DT raid/NN ./.
```

Fig. 2. The Output obtained from stanford POS tagging

- NER: All person and location names mentioned in the given drug crime text are obtained. Fig. 3 shows the output obtained from Stanford NER.

PERSON	Asmabi
PERSON	Zuhaila
PERSON	Rayes

Fig. 3. The output obtained from stanford NER

C. *Extraction of Drug Crime Entities*

In this phase, a set of heuristic and syntactic/grammatical rules and patterns are designed to extract drug-related information from online news articles. The following illustrations demonstrate some examples of the heuristic algorithms used.

Algorithm 1 Indirect Drug Extraction

```
Require: file ≠0
  Token <= file
  DL<= Drugs list
DIKL<= Drug indicator keywords list
        while file ≠0 do
        previousToken(s) <= Token[-1…-3]
        nextToken(s) <= Token[+1…3]

        For DIKLcounetr=0 to length [ DIKL ]-1 do
                If ( previousToken(s) == DIKL [ DIKLcounter])
                or ( nextToken(s) == DIKL [ DIKLcounter]) then
                        For DLcounetr=0 to length [ DL ]-1 do
                            If ( Token == DL [ DLcounter])
                            then
                                Drug extracted <= token
                            End if
                        End for
                End if
        End for

        End while
```

```
Algorithm 2 Drugs Quantity Extraction
Require: file ≠0
  Token <= file
  DQIKL<= Drugs Quantity indicator keywords list
          while file ≠0 do
          previousToken <= Token[-1]
                    For DQIKLcounetr=0 to length [ DQIKL ]-1 do
                        If ( previousToken== number and Token==
                                DQIKL[ DQIKLcounetr])
                                Drugs Quantity extracted <=
                                previousToken + token
                         End if
                    End for
          End while
```

```
Algorithm 3 Nationality extraction
Require: file ≠0
 Token <= file
 NL<= Nationality list
 ENKL<= Extra Nationality keywords list
SIKL<= smugglers indicator keywords list
ExtractedNationalities <=empty list
while file ≠0 do
          previousToken <= Token[-1]
           For NLcounetr=0 to length [NL]-1 do
                    If ( token= =NL[NLcounter]) or
                    ( previousToken+ token= =NL[NLcounter] )) do
                    Nationality extracted <= token
                    ExtractedNationalities.add( Nationality extracted)
                    End if
           End for
             For ENKL counetr=0 to length [ENKL]-1 do
                 If ( token= = ENKL [ENKL counter])
               or ( previousToken+ token= = ENKL [ENKL counter] )) do
                Nationality extracted <= token
                 ExtractedNationalities.add( Nationality extracted)
                   End if
             End for
End while
 For ENcounetr=0 to length [ExtractedNationalities]-1 do
 For SIKLcounetr=0 to length [SIKL]-1 do
   If ( token= = SIKL [SIKLcounter])
    Distance=compute-distance (token, ExtractedNationalities[ENcounetr] )
    Distancelist[ENcounetr]= Distancelist[ENcounetr]+ Distance;
   End if
 Endfor
 Endfor
    Return Extracted Nationality with smallest distance
```

In addition to the heuristic rules developed on the basis of the developed gazetteers, POS and NER information are also used to develop grammatical patterns for extracting drug crime entities. Several rules or patterns are developed. An example of these patterns is presented as follows:

Quantity <= CD + { NN |NNS} + [1]* + { NN |NNS}

Example of extracted items

- 201/CD grammes/NNS of/IN heroin/NN
- 105/ CD g/NN heroin/NN
- 46/CD syabu /NN pills/NNS
- 2.36/CD gm/NN nimetazepam/NN

PRICE <= IN+{ DET |IN}*+{ NNP}+CD

Example of extracted items

- worth/IN RM/NNP 3000/CD
- worth/IN some/DT RM/NNP 100000/CD
- worth/IN about/IN RM/NNP 35000/CD

IV. EXPRIMENTAL SETTING

To evaluate our models, we determine the performance measures for evaluating the entity extraction system; such measures include precision, recall, and weighted mean Fβ = 1− score. Precision is the percentage of the correct entities obtained by the system. Recall is the percentage of the entities present in the corpus that are found by the system. The definitions of the performance measures used are summarized as follows. Similar performance measures are used to evaluate the results of the baseline experiments.

$$\text{Recall} = \frac{\#\text{ of correctly classified drugs NEs}}{\#\text{ of drugs in the corpus}} = \frac{tp}{tp+fn} \quad (1)$$

$$\text{Precision} = \frac{\#\text{ of correctly classified drugs NEs}}{\#\text{ of drugs in the corpus}} = \frac{tp}{tp+fp} \quad (2)$$

$$F_\beta = \frac{(1+\beta 2)*(\text{precision}*\text{recall})}{(\beta 2*\text{precision}+ \text{recall})} \quad (3)$$

V. EXPERIMENTAL RESULTS

In this experiment, we examine the overall performance of the rule-based extraction tool for drug crime entities. Table I shows the summary of the experimental results for extracting drug name, nationality, crime location, drug quantity, drug price, and drug hiding way.

TABLE I. PERFORMANCE OF THE RULE-BASED DRUG CRIME ENTITY EXTRACTION TOOL ON EACH CRIME ENTITIES.

	Precision	Recall	F-Measure
Drugs names	0.96	0.97	0.96
Nationality	0.92	0.87	0.89
Crime location	0.83	0.79	0.81
Drugs quantity	0.87	0.86	0.86
Drugs price	0.88	0.93	0.90
Hiding way	0.85	0.73	0.79

Optimal results are achieved for extracting drug names. To our knowledge, drug names are limited; thus, the high result obtained is anticipated. The worst result is achieved for extracting the methods of hiding drugs. Extracting the methods of hiding drugs is difficult compared with extracting common entities because the former is not written in common patterns.

The experiments generally show promising results. The experimental results clearly demonstrate that a rule-based algorithm is appropriate for extracting drug-related crime entities. The promising result motivates scholars to comprehensively and comparatively extract and analyze drug-related crime information.

VI. CONCLUSION AND FUTURE WORK

This work developed a rule-based IE system for collecting relevant drug crime information to help the community and police investigators. This study revealed several drug-dependent specific linguistic resources such as drug-crime-

specific lexicon and gazetteers. Moreover, a rule-based extraction tool is established to assess where and how drug traffickers hide drugs, identify the nationalities of drug dealers and the types (names) of drugs, and evaluate the quantity and prices of drugs in the local market. Finally, this experiment has shown promising results and clearly demonstrated that a rule-based algorithm is suitable for extracting drug-related crime entities.

Our future works will focus on developing a general crime IE and analysis tool that encompasses all crime types.

REFERENCES

[1] C. Development Research Center of the State Council, "Innovating to Overcome Scarcity: China's Experiences and Future Development, " presented at the Aix-en-Provence Economic Forum on "A World of Scarce Resources", 2006.

[2] M. Coscia, and V. Rios, "How and where do Criminals Operate? Using Google to Track Mexican Drug Trafficking Organizations," 2012.

[3] A. Alkaf, and M. Mohd, "Extraction Of Nationality From Crime News," Journal of Theoretical and Applied Information Technology, vol. 54, 2013.

[4] M. Asharef, N. Omar, and M. Albared, "Arabic Named Entity Recognition In Crime Documents, " International Journal of Theoretical and Applied Information Technology, 2012.

[5] Ku, Iriberri, and Leroy, "Crime Information Extraction from Police and Witness Narrative Reports, " in IEEE International Conference on Technologies for Homeland Security, Boston, 2008.

[6] H. Ito, "LonMaps: An Architecture of a Crime and Accident Mapping System based on News Articles," in ICONS 2014 : The Ninth International Conference on Systems, 2014.

[7] D. Bruin, Cocx, Kosters, Laros, and Kok, "Data Mining Approaches to Criminal Career Analysis," presented at the Sixth International Conference on Data Mining 2006 (ICDM'06), 2006.

[8] Jaiswal, Gunjala, Londhe, Singh, and Solanki, "Crime Automation & Reporting System," International Journal of Science and Modern Engineering (IJISME), vol. 1, 2013.

[9] Arulanandam, Savarimuthu, and Purvis, "Extracting Crime Information from Online Newspaper Articles," in Proceedings of the Second Australasian Web Conference (AWC 2014), Auckland, New Zealand, 2014.

[10] W. C. (2013), " Mapping Crimes Collaboratively, http://www.wikicrimes.org.."

[11] Furtado, Ayres, d. Oliveira, Vasconcelos, Caminha, D'Orleans, et al., "Collective Intelligence in Law Enforcement : The Wikicrimes System," Information Sciences, 2010.