Characterizing Crimes from Web

Márcio V. C. da Silveira[1], **Wladmir C. Brandão**[1]

[1]Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Brasil

marcio.campos@outlook.com, wladmir@pucminas.br

Abstract. *Crime prevention requires the effective use of police resources, which demands the access of criminal information for planning security actions. The number of crime occurrences is higher than the official reported numbers. Many victims do not report crimes directly to the security agencies. Instead, they prefer to anonymously report using different channels, such as the Web. In this article, we introduce our approach to characterize crimes reported in the Web. Particularly, we collect criminal data from popular websites that store crime occurrences, and we use clustering analysis to discover crime patterns on the collected data. Applying our approach to a popular Brazilian crime report website, we observe that more than 41% of the crimes were not reported to the security agencies, and most of them are thefts and robberies occurring at night and dawn. In addition, minor offenses present different patterns of serious crimes. Moreover, crime patterns are different in rich and poor neighborhood.*

1. Introduction

Security is a worldwide concern. According to the article 3 of the Universal Declaration of Human Rights, everyone has the right to life, liberty and personal security. Particularly in Brazil, the constitution guarantees that public safety is a duty of the State [United Nations General Assembly 1948]. Crime prevention and control require the effective use of police resources. Thus, the knowledge on crime patterns is paramount for planning actions to avoid crimes.

The Bulletin of Occurrence (BO) is the official document used by the Brazilian security agencies to record a crime occurrence [Azevedo 2014]. However, many victims do not used it. According to the Brazilian national victimization survey [Secretaria Nacional de Segurança Pública - SENASP 2013], only 19.9% of crime occurrences are reported. In addition, the lack of trust in the police and the idea that the police could not do anything about the crime are the main reasons to not report crime occurrences. In many cases, the victims prefer to report the crime occurrence in websites or social networks. Thus, the number of crime occurrences is much higher than the reported in official statistics.

Onde Fui Roubado[1] is a website which manages a large criminal database with information on the most diverse types of crimes. If a person has been a victim or witnessed a crime, one can report the crime through the site itself. As the user must mark the location where the crime occurred, most of the registered cases have the latitude and longitude of the occurrence, as well as several relevant data about the nature of the crime, such as the value of the loss and the date and time the crime occurred.

[1]http://www.ondefuiroubado.com.br

In this article, we introduce our crime characterization approach to characterize crimes reported in the Web. In particular, our approach collects crime occurrences from *Onde Fui Roubado*, and uses clustering analysis to discover crime patterns on the collected crime occurrences. Analytical results on a sample of crime occurrences show that we have eight groups of crime patterns, considering crimes occurred in the Brazilian city of Belo Horizonte, from January 1, 2012 to August 31, 2016. Additionally, we observe that two in each five crimes were not reported to the security agencies. Moreover, the crimes range from minor offenses, such as housebreaking and vehicle break-ins, to serious crimes such as kidnappings, but most of them are thefts and robberies occurring at night and dawn. The type and severity of the crimes depends mostly on the neighborhood and period of time, and victims report more serious crimes with great danger and financial losses and less minor offenses.

The remainder of this article is organized as follows: Section 2 describes the theoretical underpinnings of this article. Section 3 discusses the related work. Section 4 presents our approach used to capture and analyze criminal data from the Web. Section 5 presents the characterization of the crimes. Finally, Section 6 presents the conclusion and shows directions for future research.

2. Background

In this section we present concepts related to our approach for crime characterization, including Web crawling, classification and clustering techniques.

Web Crawling evolves collecting Web documents as quickly as possible to build a comprehensive body of documents that will be used later for indexing and searching [Levene and Poulovassilis 2004]. Web crawlers are information retrieval system components that must request and store documents from web servers, extract links from documents, and schedule the next crawling step using the extracted links [Myllymaki 2002].

Classification is the task of automatically assign natural language texts to predefined categories based on their content [Hayes and Weinstein 1990]. Similarly, text categorization consists on identifying the main textual documents and associate them with one or more predefined categories. For this, we should determine in which categories, already classified previously, a certain attribute in question presents more similarity and can be considered of that class [da Silva Filho et al. 2010]. Categorized texts are represented as a class, in a more organized way, thus allowing a certain content to be accessed easily and without much effort.

Clustering is the process of class discovery, where objects are sorted into groups and classes are previously unknown [Malathi and Baboo 2011]. Clustering techniques are basically used to obtain data patterns while classification techniques are used to obtain the classes for future prediction [Sharma and Kumar 2013]. Clustering has as its basic principle the gathering of records that have similarities in a database, partitioning them into subsets, called clusters [Goldschmidt and Passos 2005]. In particular, a cluster is a collection of objects that are similar to each other within the same group and unequal to objects

in other groups [Zubi and Mahmmud 2013]. In this way, the registers belonging to the same cluster have similarities between them and, at the same time, the objects belonging to different clusters have a high dissimilarity. Clustering only identifies similar data groups and does not have the pretension of classifying, estimating, or predicting the value of a variable [Camilo and da Silva 2009]. One of the most common clustering technique used by researchers is k-means, which consists on partitioning the objects into k clusters based on their similarities measured by a distance function [Han et al. 2011]. Although efficient, k-means has the limitation of working only with numeric values. The x-means extends k-means, with the advantage of estimating the best number of k groups from the dataset to be evaluated [Hartigan and Hartigan 1975]. The x-means algorithm overcomes the limitation of k-means and does not require to previously setup the number of clusters [Hamerly and Elkan 2004].

3. Related Work

The investigation on crime patterns and analysis of criminal data have been the focus of much research reported in the literature. They differ in terms of the data sources and the techniques used to analyze the data. [Singh et al. 2016] describe a system that analyzes crime records, increasing the accuracy of crime prediction. They consider different crime occurrences to accurately predict whether a similar crime pattern is observed. In their work, they analyze the geographic patterns of crimes and predict a particular type of crime that could occur in the near future. The data used in analysis were collected by a crawler from news feeds, blogs and articles on the Web. The prediction of the crime category is possible if one detect a similar pattern for an area where the occurrence of crimes is highly probable. The authors deliver the results in the form of a criminal category of a crime by which an effective measure can be deployed by police forces to protect the neighborhood. Also, when the criminal records were rich, they are able to predict the timeline for crimes.

The ARCA approach uses association rules to discover crime patterns from an official crime database provided by the Department of Social Defense of Minas Gerais, in Brazil [Laporais and Brandão 2014]. The dataset contains two years of Brazilian crime occurrences., and the approach recognizes mutual implications between crime occurrences, retrieving relevant information about criminal behavior. It processes and loads the data into a Data Warehouse (DW), filters, and extracts relevant data samples. Lastly, they use the *Apriori* algorithm to discover association rules on DW. The association rules discovered by ARCA were used to characterize criminal behavior, which pointed to patterns of nontrivial criminality that motivate further investigation.

[Zubi and Mahmmud 2013] analyzes the Libyan crime occurrences using the k-means algorithm for data clustering and the *Apriori* algorithm for association rules. Their work is aimed to help the Libyan government making strategic decisions to address the increasing of criminal activities. Data were collected from the Libyan Police Department. These data were preprocessed to obtain clean and accurate records using preprocessing techniques, and used to uncover different crimes, tendencies and criminal behaviors that were grouped according to their attributes. For their work, more than 350 records were used. They have gained overall statistical knowledge of the criminal age in relation to the type of crime. According to the authors, the model aims to help the Libyan Security Committee to identify criminal behavior by specifying the types of offenses and relating them to criminal groups in Libya.

4. Crime Characterization Approach

In this article, we introduce our approach to characterize crimes from criminal records extracted from the Web. Figure 1 shows the components of our approach. First, the crawler component extracts raw criminal records from data sources in the Web, such as *Onde Fui Roubado*. Next, the filter component selects relevant criminal records, eliminating ambiguities and preparing the criminal features for clustering. Finally, the clustering component generates groups of crime patterns.

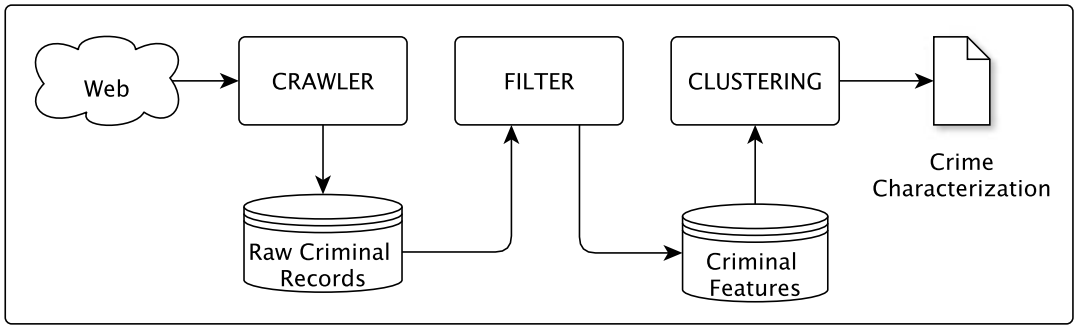

Figure 1. The workflow of our crime characterization approach

4.1. Crawling Criminal Records from Web

We extract criminal data from Web by crawling crime occurrences from data sources, such as *Onde Fui Roubado*. Particularly, we created a dataset of raw criminal records composed by 33,845 crime occurrences in big Brazilian cities, such as São Paulo, Rio de Janeiro, Belo Horizonte, Salvador, Porto Alegre, Fortaleza, and Curitiba. All the criminal records are from January 1, 2012 to August 31, 2016. The dataset contains different crime attributes. Table 1 shows these attributes, its description and data type.

Table 1. Dataset Attributes

Attribute	Description	Type
id	Internal ID of the crime record	int
description	Textual description of the crime	text
type_of_crime	Type of the crime	enum
city_id	City ID where the crime occurred	int
region	City region where the crime occurred	int
address	Address where the crime occurred	varchar
latitude	Latitude where the crime occurred	float
longitude	Longitude where the crime occurred	float
date_time_record	Date and time of the crime	datetime
gender	Gender of the victim	bit
loss_amount	Total loss of the victim	double
registered_bo	A flag informing if the victim reports the crime to security agencies	bit

From Table 1, we observe the *type_of_crime* attribute, which describes the type of the crime. Particularly, there are 11 types of crimes reported by the victims: assault, assault by groups, theft, robbery, housebreaking, store break-in, vehicle break-in, vehicle theft, "saidinha"[2], "arrastão"[3], and "lightning kidnapping"[4].

4.2. Filtering Criminal Features

The filter component prepares data for clustering. In particular, we select a sample of 5,268 criminal records of crimes in the Belo Horizonte city. We also select relevant features for the clustering procedures. Since our clustering algorithm only works with numeric data, we transform non-numerical features to numeric. No transformation was required for the *gender*, *registered_bo*, and *type_of_crime* attributes, since they are numerical. However, we have to define city regions based on the latitude and longitude attributes. According to [DETRAN-MG 2016], Belo Horizonte city is divided into nine regions. Additionally, each region is divided into neighborhoods. We use the Google Maps Geocoding API[5] to map latitude and longitude to city regions. Table 2 shows the Belo Horizonte city regions, and the number and percentage of criminal records in each region.

Table 2. Number of crime records by region

City region	Number	Percentage
Barreiro	140	2.73%
Centro-Sul	2,421	47.27%
Leste	199	3.88%
Nordeste	435	8.50%
Noroeste	464	9.06%
Norte	144	2.81%
Oeste	559	10.91%
Pampulha	622	12.15%
Venda Nova	138	2.69%
Total	5,122	100.00%

From Table 2, we observe that some regions concentrate more criminal records than others. For example, the "Centro-Sul" region concentrates almost half of all the crimes reported in Belo Horizonte city, while the "Venda Nova" region concentrates less than 3% of the crimes. This behavior represents a socioeconomic bias, since the richer regions have more records than the poor ones.

Another required transformation was performed in the *loss_amount* attribute, which provides the estimated value, in Brazilian currency (R$), of the victims' losses. Based on the value of the minimum wage in Brazil, we define 10 level of losses. The level 1 consists of losses until one minimum wage (R$ 880.00), the level 2 consists of losses between one minimum wage (R$ 880.00) and two minimum wages (R$ 1,760.00),

[2]A type of attack in which the thief stalks the victim waiting for him to leave the bank with cash withdraw from the ATM to carry out the robbery.

[3]A type of attack in which groups of thieves assault groups of people in waves.

[4]A type of attack in which the victim is forced to visit banks and withdraw money for the kidnappers.

[5]http://developers.google.com/maps/documentation/geocoding.

and so on until the level 10, which consists of losses greater than ten minimum wages (R$ 8,880.00).

We also performed transformations in the *date_time_record* attribute. In particular, we divided the day into 4 periods: morning, afternoon, night and dawn. Table 3 shows the schedule we used to determine the periods of the day.

Table 3. Periods of time

Period	Schedule
Morning	06:00:00 to 11:59:59
Afternoon	12:00:00 to 17:59:59
Night	18:00:00 to 23:59:59
Dawn	00:00:00 to 05:59:59

4.3. Clustering Crime Patterns

The clustering component uses the x-mean clustering algorithm implemented in the Weka package [WEKA 2016], which is very popular between researchers. It contains tools for data processing, regression, clustering, association rules, and data visualization. Particularly, we run and test the performance of the clustering algorithm over our dataset, using six features: region, level of loss, type of crime, gender, recorded BO and the period of the day of crimes.

Preliminary experiments show that without using a predefined number of clusters, the algorithm tends to focus on only two features: gender and recorded BO. We have decided to remove these two features for the creation of the clusters. The features were maintained for analysis, but were not considered by the clustering algorithm. Moreover, one of the variables that can be defined in the execution of x-means is the minimum and maximum number of clusters that it will generate. In order to obtain a larger number of clusters we defined the minimum number of clusters to ten. However, the algorithm itself returned us eight clusters, demonstrating that this is the number of relevant relationships that it was able to find.

5. Crime Characterization

In this section we present our throughout characterization of crimes based on our proposed approach. Particularly, we present and discuss the clusters generated by the clustering component we used to analyze the criminal features we crawled from Web. Table 4 shows the number and percentage of criminal records in each generated cluster. From Table 4 we observe that the clustering component generated eight clusters with different number of criminal records. In the following, we present the properties of each generated cluster:

Cluster 1 - Afternoons thefts and robberies: Mostly composed by crimes occurred in the Barreiro region, but there are also crimes occurred in the Leste and Nordeste regions. More than 92% of the crimes caused victims' losses of up to 3 minimum wages, more than 57% of the victims are men, and more than 47% of victims did not reported the crimes to security agencies. All the crimes are thefts or robberies occurred in the afternoon.

Table 4. Number of criminal records by cluster

Cluster	# records	Percentage
1	699	13.65%
2	462	9.02%
3	432	8.43%
4	600	11.71%
5	361	7.05%
6	862	16.83%
7	1,226	23.94%
8	480	9.37%
Total	5,122	100.00%

Cluster 2 - Night and dawn vehicle break-ins and assaults: Mostly composed by crimes occurred in the Centro-Sul region. More than 76% of the crimes are vehicular break-ins, and more than 51% of all assaults occurrences are concentrated in this cluster. More than 80% of the victims are men, and more than 47% of victims did not reported the crimes to security agencies. The crimes occurred exclusively during the night and dawn.

Cluster 3 - Mornings assaults by groups: The crimes occurred in the Barreiro, Centro-Sul, Leste and Nordeste regions. The crimes caused victims' losses from 1 to 7 minimum wages, more than 50% of the crimes occurred in the morning, and more than 30% of the crimes are assaults made by groups.

Cluster 4 - Daylight poor neighborhood crimes: Mostly composed by crimes occurred in the Noroeste, Norte, Oeste, Pampulha and Venda Nova regions. The crimes caused victims' losses of up to 4 minimum wages. The crimes are diverse, some serious as "lightning kidnapping". More than 56% of the victims are men, and more than 43% of victims did not reported the crimes to security agencies. All the crimes occurred in the morning and afternoon.

Cluster 5 - Afternoons and nights *saidinha* and *arrastão*: There are no predominant regions. The crimes caused victims' losses of up to 6 minimum wages. More than 94% of "saidinha" occurrences, and more than 86% of "arrastão" occurrences are concentrated in this cluster. More than 63% of the victims are men, and more than 47% of victims did not reported the crimes to security agencies. More than 77% of the crimes occurred in the afternoon and night.

Cluster 6 - Night and dawn crimes in Pampulha: Mostly composed by crimes occurred in Pampulha region. Crimes varies from thefts, robberies, and store break-ins, and caused victims' losses of up to 6 minimum wages. More than 57% of the victims are men, and more than 42% of victims did not reported the crimes to security agencies. All the crimes occurred at night and dawn.

Cluster 7 - Moonlight rich neighborhood crimes: Mostly composed by crimes occurred in the Centro-Sul, Barreiro, and Nordeste regions. Crimes varies from thefts, robberies, and assaults, and caused victims' losses of up to 7 minimum wages. More than 62% of the victims are men, and more than 47% of victims did not reported the crimes to security agencies. All the crimes occurred at night and dawn.

Cluster 8 - Great danger and financial loss crimes: More than 48% of the crimes occurred in Centro-Sul and Pampulha regions. The crimes caused severe danger and victims' losses greater than 8 minimum wages, which is justified by the fact that more than 44% of all vehicle thefts and "lightning kidnappings" is here. More than 75% of the victims are men, and only 11% of victims did not reported the crimes to security agencies.

6. Conclusions

In this article we introduced our crime characterization approach to characterize crimes reported in the Web. Particularly, our approach extracts raw criminal records from crime data sources in the Web, filters criminal features and performs clustering analysis to discover crime patterns.

We used our approach to characterize crimes in Belo Horizonte, a big city in Brazil, from 2012 to 2016. From analytical results, we observe that crimes range from minor offenses to serious crimes and the type and severity of them depends mostly on the neighborhood and period of time. In addition, we observed that victims tend to report more serious crimes with great danger and financial losses, and less minor offenses. Moreover, serious crimes occur at night and down in rich neighborhoods and at morning and afternoon in poor neighborhoods.

Our analysis show that the proposed approach can be effectively used by security forces to plan actions for crime prevention. Thus, they can address the problem of lack of trust in the police, creating a relationship of collaboration and effective participation of society in the fight against crime.

As future work we intent to expand our analysis to other cities and countries. In addition, we intent to evaluate the performance of other clustering techniques, such as hierarchical clustering. Moreover, we intent to integrate data from other crime data sources from Web, such as WikiCrimes and social networks. With these changes, we believe that it will be possible to incorporate crime prediction modeling into our approach.

Acknowledgments

The authors are thankful for the partial support given by the Pontifical Catholic University of Minas Gerais (Grant PUCMINAS-FIP 2016/11086-S2), MCTI/CNPq (Grant 444156/2014-3), FAPEMIG/PRONEX (Grant APQ-01400-14), and the authors' individual grants and scholarships from CNPq.

References

Azevedo, S. N. d. (2014). *O protesto de títulos e outros documentos de dívida*. EdiPUCRS, 2 edition.

Camilo, C. O. and da Silva, J. C. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Technical report, Universidade Federal de Goiás (UFC).

da Silva Filho, L. A., Favero, E. L., Dias, M. M., and de Mendonca, C. K. L. (2010). Mining association rules in data and text - an application in public security. In *Proceedings of the 7th International Conference on Information Systems and Technology Management*, pages 3644–3672.

DETRAN-MG (2016). Lista de regiões e bairros de Belo Horizonte. Available in: `https://www.detran.mg.gov.br/habilitacao/cnh-e-permissao-para-dirigir/solicitar-renovacao-da-cnh/consulta-lista-de-regioes-e-bairros-de-belo-horizonte`. Access in: Oct 7, 2016.

Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações*. Campus, 1 edition.

Hamerly, G. and Elkan, C. (2004). Learning the k in k-means. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, pages 281–288.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier, 3 edition.

Hartigan, J. A. and Hartigan, J. (1975). *Clustering algorithms*. Wiley New York.

Hayes, P. J. and Weinstein, S. P. (1990). CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In *Proceedings of the 2nd Annual Conference on Innovative Applications of Artificial Intelligence*, pages 49–64.

Laporais, B. and Brandão, W. C. (2014). ARCA: Mining crime patterns using association rules. In *Proceedings of the IADIS International Conference Applied Computing*, pages 159–166.

Levene, M. and Poulovassilis, A. (2004). *Web dynamics: Adapting to change in content, size, topology and use*. Springer Science & Business Media, 3 edition.

Malathi, A. and Baboo (2011). Algorithmic crime prediction model based on the analysis of crime clusters. *Global Journal of Computer Science and Technology*, 11(11):139–145.

Myllymaki, J. (2002). Effective Web data extraction with standard XML technologies. *Computer Networks*, 39(5):635–644.

Secretaria Nacional de Segurança Pública - SENASP (2013). Pesquisa nacional de vitimização. Available in: `http://www.crisp.ufmg.br/wp-content/uploads/2013/10/Relat%C3%B3rio-PNV-Senasp_final.pdf`. Access in: May 7, 2016.

Sharma, A. and Kumar, R. (2013). The obligatory of an algorithm for matching and predicting crime - using data mining techniques. *International Journal of Scientific and Engineering Research*, 4(2):289–292.

Singh, A. K., Prasad, N., Narkhede, N., and Mehta, S. (2016). Crime: Classification and pattern prediction. *International Advanced Research Journal in Science, Engineering and Technology*, 3(2):41–43.

United Nations General Assembly (1948). Universal Declaration of Human Rights. Available in: http://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf. Access in: May 7, 2016.

WEKA (2016). The university of waikato. weka 3: Data mining software in java. Available in: http://www.cs.waikato.ac.nz/ml/weka. Access in: May 9, 2016.

Zubi, Z. S. and Mahmmud, A. A. (2013). Using data mining techniques to analyze crime patterns in the libyan national crime data. In *Proceedings of the 1st WSEAS International Conference on Image Processing and Pattern Recognition*, page 79–85.