Discovering Geographical Patterns of Crime Localization in Mexico City

Roberto Zagal-Flores[1], Félix Mata-Rivera[2],
Instituto Politécnico Nacional,
UPIITA-IPN[1, 2]
Ciudad de México, México
e-mail: rzagalf@ipn.mx[1], mmatar@ipn.mx[2]

Christophe Claramunt[3], Edgar Catalan-Salgado[4]
Naval Academy Research Institute[3]
Brest Naval, 29240, France
Instituto Politécnico Nacional, ESCOM-IPN[4]
Ciudad de México, México
e-mail: claramunt@ecole-navale.fr[3], ecatalans@ipn.mx[4]

***Abstract*. The search for a better understanding of crime patterns in large urban areas is still a crucial issue that deserves novel research methods and approaches. In particular, combination of institutional databases and novel information medias such as social networks appear as a promising trend that might favor development of more efficient criminal information management and crime prevention systems. However, most existing systems do not take into account to the best of our knowledge the geographical dimension although this might provide a better representation of how crimes spread over space and time. The research presented in this paper develops a knowledge discovery approach based on a close integration of official, social and geographical data sources. The result is a modeling approach that provides a-priori knowledge of safe and unsafe places and the ones that are even candidates to become unsecured places. The aim is to not only give an overall geographical representation of crime patterns that might be useful for decision-makers, but also web-based resources to the citizens. The whole approach has been applied to Mexico City.**

Keywords-GKD; Crossing-data; Social Web mining; Geo-social Web Analytics; Web ontology.

I. INTRODUCTION

The research presented in this paper is grounded in the assumption that social networks offer novel resources to reflect population's opinion on insecurity problems occurring in urban zones. For instance, In Mexico City, people often denounce complaints, events, and unsafe places. Many Facebook pages and Twitter accounts also regularly report crimes that citizens suffer thus providing rich descriptions often located in space and timely stamped. Therefore, it clearly appears that if such crime data can be geographically analyzed, then it will be possible to build services to citizens and decision-aided systems to municipal authorities [20].

Over the past few years, many novel research approaches have been designed and developed based on social web mining and Geographic Knowledge Discovery (GKD) [5]) in order to solve novel challenges related to the integration of unstructured social data and structured data sources [23]. In fact not only these novel unstructured data sources require development of novel data integration and representation mechanisms, but also sound approaches to evaluate quality and veracity of the incoming data. It has been recently shown that quality and certainty of opinions and denunciations expressed on social networks can be increased when linked and related to additional data sources [5].

Crossing data would allow validating, confirming, or discovering non-intuitive knowledge at first sight (e.g., location and profile of the place or victim of a given crime). Therefore, external social data sources such as news websites, as these regularly report and discuss situations, events, or facts from different perspectives, can be combined with other web data sources to discover and search for geographical and temporal patterns.

On the other hand, news media might are likely to influence social perception of the insecurity problem as they occur in a given city. It has been particularly observed that for example social perception of insecurity in Mexico City has increased up to 69% [7], where 64.2% of the population reported that in the last three months, they modified their habits for fear of being a victim of a crime.

The research presented in this paper introduces a preliminary GKD social framework applied to crime data. The main contribution relies on the crossing of different datasets to discover some geographic and temporal crime patterns as reflected by social media, institutional data sources and the final public perception. The remainder of the paper is organized as follows. Section 2 outlines related work. Section 3 introduces the GKD social framework. Section 4 develops the data extraction while Section 5 presents the preprocessing process. Section 6 discusses the lessons learned from the analysis of crime data. The data crossing approach is presented in Section 7. Section 8 presents the preliminary results obtained and some evaluations. Finally, Section 9 draws the conclusions and outlines future work.

II. RELATED WORK

Studies on crime data have been conducted by many research domains such as spatial data mining, geographical analysis, Big Data analysis and crime prediction. For example, the work in [1] presents a collaborative community alert application that enables citizens to publish a complaint at the time a crime occurs, and to inform other users close to that area. In Mexico City, a web system regularly and interactively informs the state of insecurity in different neighborhoods [2]. A survey of crime prevention systems has been recently published and includes an analysis of the main web and mobile crime information systems, but so far

 ISBN: 978-1-61208-557-9

there are no approaches that integrate unstructured social data and official data sources [19].

A series of spatial data mining approaches have been applied to the metropolitan area of Washington DC to discover and extract patterns from crime datasets [8] [9] [10}. A spatiotemporal-textual search engine and pattern discovery approach identify crime categories from collections of crime information [8]. The work developed in [9] provides a smart device-based Internet application to enable real-time location-based and search for crime incidents and reports. This system has the advantage of leveraging crowdsourced data to provide safe paths and crime analytics. The integrated framework proposed in [10] is also based on natural language processing whose objective is to extract, rank and score some crime events in space and time. Colocation analysis has been explored using geographical analysis and machine learning algorithms to identify and categorize regular patterns of crimes in some given locations [11]. For instance, the relationships between drink locations and habits (e.g., bars, alcohol, respectively) and crimes have been studies and show that crimes are often related to such behaviors. Next, the authors developed a neighborhood graph-based approach that searches for spatio-temporal patterns. A Geographical Information Systems (GIS) approach searches for criminal hot spots and drug cases in China [21]. Understanding the reasons behind crime patterns in space and time is also a domain of study that has been considered by machine learning algorithms, this is a crucial issue for crime prediction studies [18]. A crime analytics system based on temporal prediction analysis has been suggested in [22]. The authors developed a predictive system based on natural language processing, and whose objective is to identify the intensity of some crime events, using a combination with social-based data, public and private data.

Overall, and despite the novelty and interest of the above mentioned approach, most of them do not completely take advantage of the potential offered by a combination of crime, social and spatio-temporal data. We believe that such integration should provide novel means for discovering new geographical and temporal patterns, as well as an evaluation of the degree of certainty associated to such trends. The research presented in this paper proposes a hybrid semantic approach that considers such challenges, and combines and crosses social and official crime data using ontology exploration and machine learning techniques.

III. THE GKD SOCIAL FRAMEWORK

The GKD social framework is fed from official and unofficial sources, data from the national public security system [16], verified user publications, Facebook and Twitter communities, as well as news web pages.

The information collected refers to descriptions of criminal events and complaints, which will be analyzed and displayed on a crime map. The goal is to characterize crime from a geographical perspective. What characteristics describe a crime place? What is the geographic and spatial profile of a crime place? How often does a crime occur in a particular place? What places can become a potential place where a crime can happen? The geographic information is provided by collaborative maps and official cartography from the local government.

Our approach adapts knowledge discovery in databases (KDD) methodology, which refers to the non-trivial process of discovering knowledge and potentially useful information within the data contained in some information repository [2]. Our framework uses four phases: 1) retrieval and extraction of data related to crime, events, and social complaints by geographical areas from Facebook and Twitter, news web pages, and official data sources; 2) semantic analysis of data using crime web ontology; 3) classification, which considers supervised machine learning techniques to separate the crime activity and geographic patterns; and 4) crossing geographic patterns and official data to acquire new knowledge and evaluate the certainty of social data. Finally, the main principles of our approach are illustrated in Figure 1 and are part of a previous work [5].

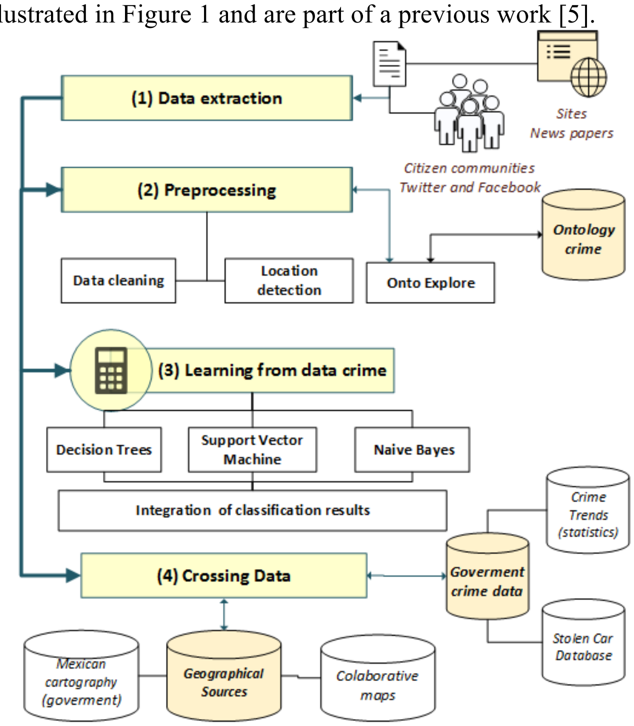


Figure 1. A GKD Social Framework for Crime Data

Figure 1 shows the main phases of our approach; the GKD is the combination of the semantic analysis and machine learning layers. The first layer uses the ontology to pre-classify all the unstructured data obtained from the web and social networks. The second layer makes the classifications for official and social data (it means structured and unstructured data). The following subsections will explain each one of these phases:

IV. DATA EXTRACTION

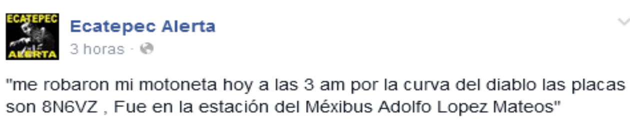
The extraction process consists of retrieving the reports of complaints from official sources (a fragment of the information can be seen in Figure 2) and unofficial sources (see a sample of publication in Figure 3).

	ENTIDA	MUNICIPIO	MODALIDAD	TIPO	SUBTIPO	ENERO	FEBRERO
2011	MEXICO	TECAMAC	DELITOS PATRIMC	ABUSO DE CONFI/	ABUSO DE CONFIANZA	9	5
2011	MEXICO	TECAMAC	DELITOS PATRIMC	DAÑO EN PROPIE	DAÑO EN PROPIEDAD AJENA	34	32
2011	MEXICO	TECAMAC	DELITOS PATRIMC	EXTORSION	EXTORSION		
2011	MEXICO	TECAMAC	DELITOS PATRIMC	FRAUDE	FRAUDE	2	1
2011	MEXICO	TECAMAC	DELITOS PATRIMC	DESPOJO	CON VIOLENCIA		
2011	MEXICO	TECAMAC	DELITOS PATRIMC	DESPOJO	SIN VIOLENCIA		
2011	MEXICO	TECAMAC	DELITOS PATRIMC	DESPOJO	SIN DATOS	9	7

Figure 2. Data from National Public Security System (data in Spanish)

In the case of Figure 2, complaint data were obtained from [2], and it contains fields, such as municipality (location of the crime), type of crime, subtype of crime, and date. Figure 3 shows a post (in Spanish) retrieved from a Facebook community specializing in citizens' reports of crimes called "Ecatepec on alert" (text in Spanish « Ecatepec en Alerta»). Ecatepec is a municipality (in the country of Mexico) with high crime rates, and it is located near Mexico City.

Regarding the number of crime complaints dataset in official and unofficial sources, a Facebook page named "Ecatepec denunciation" (in Spanish « Denuncia Ecatepec ») extracted 6771 posts of complaints. In opposite, the news web page named "A Fondo Estado de Mexico" (site of news crime) only extracted 100 datasets from complaints.

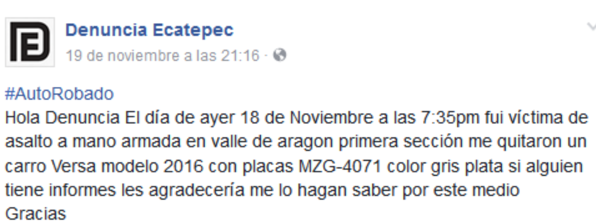


Translation: Ecatepec on alert, "My motorcycle was stolen today at 3 am at the Devil's Curve. The numberplate is 8N6VZ. It was at the Méxibus in Adolfo Lopez Mateos station."

Figure 3. Nonofficial information from Facebook (post in Spanish)

V. PREPROCESSING

The data extracted is processed with a computational linguistic process whereby the text is cleaned and converted to lower case, words are lemmatized, and punctuation marks and emoticons are eliminated. An example in Figure 4 shows an original post from Facebook about a complaint, then below shows the result processed linguistically.



Lemmatization: `autorobado denunciar dia ayer 18 noviembre 7 35 pm victima asalta mano armada valle de aragon primero seccion quitar carro versa modelo 2016 placa mzg 4071 color grises plata informa agradecer hacer medio`

Translation: Complaint Ecatepec, #StolenAuto "Hello Complaint Ecatepec, yesterday November 18 at 7:35 p.m. I was the victim of an armed robbery in the Valley of Aragon first section. A car Nissan Versa model 2016 was stolen. The numberplate is MZG-4071, and the color is gray silver. If anyone has reports, I would be grateful."

Output: "StolenAuto report day yesterday 18 november 7 35 pm victim assault armed hand valley of Aragon first section car Nissan versa model 2016 plate mzg 4071 color gray silver informs thank to do medium"

Figure 4. Lemmatization of a Facebook post

A. Location analysis

Once the data is cleaned and processed, the following phase consists of determining the location where the event occurred. It is limited only to data containing terms, words, or names of places. The process uses GeoNames [17], gazetteers, and a terms dictionary. This phase comprises three steps:

1. The first step consists of identifying all place names within the data and it is achieved using a prebuilt specialized terms dictionary. It ensures that the tweet, post, or data from the web page describes a geographical place (the data that cannot be identified are discarded.)
2. The second step disambiguates the possible names for the same geographic location but with different granularity (municipality, colony, street, and avenue). Then, the GeoNames web service [17] is used to determine the granularity of the place.
3. In the third step, a gazetteer is used to identify place names and synonyms within a particular area. Figure 5 shows an example of this analysis.

```
Ecatepec -> a geographical location
Ecatepec of Morelos -> a geographical location
Ecatepec of Morelos -> a synonym: Ecatepec
Ecatepec -> a part of the State of Mexico
```

Figure 5. Location Analysis

The location analysis is an experimental phase. Hence, it can be enhanced by machine learning methods [4].

B. Semantic analysis

This process focuses on social network data and web pages. The analysis consists of pre-classifying the data to determine to which category the security domain belong to (e.g., security, theft, complaint) by using an ontology prebuilt based on concepts derived from security domain. In particular, the ontology helps to determine if the data describes a crime, complaint, or they are part of another context. The algorithm OntoClassifier is shown as follows.

OntoClassifier Algorithm

```
1. Begin
2. Let q[i] elements of user query
3. N = 0
4.  while n < i
5.   parsing and identification (q[i])
6.   node.start()
7.     while node != null
8.     j++, i++
9.     if similarity(concept_name)
         conVec[j]neighborhood_relations(node)
         node.next()
10.      geoVector[k] = geographic_search(conVec[j])
11.   thematic_and_social_search(conVec[j])
12.   temporal_search(conVec[j])
13.  j++, k++, n++
14. End
```

The algorithm works as follows: the input is the data extracted from social media and web pages; then, the ontology is explored using the algorithm OntoClassifier (adapted from [3]). Each term identified by the social media source is correlated with the Ontology. If a match occurs, the context of the concept (all neighborhood concepts) is extracted, and they are stored into a vector (Algorithm 1, lines from 4 to 10). The vector is used to determine which domain and category the data belongs; it is known by using the hyperonym relation of the ontology. This relation indicates the domain and category. In that way, the output obtained is a pre-classification of data showing what type of crime is described or mentioned in the data by using three parameters: time, thematic, and location components. The classification is based on the scale of categories used in [2]. An example of the output appears in Figure 5 where a console screenshot is shown after the OntoClassifier program is executed and the matching is listed.

```
Robar motoneta → Delito tipo Robo de auto
3 am → Hora
Placa → Atributo con valor → 8n6vz
```

Translation:"
Steal motorcycle -> Auto theft -> type of crime,
3 am -> Time,
Plate -> Attribute with value -> 8n6vz"

Figure 6. Post pre-classification process

As Figure 6 shows a pre-classified post, in the first line (in Spanish) the term "Steal scooter" is pre-classified as a crime; "Theft." In the second line, the text "3 am" is classified as a time component. In the third line, the term "vehicle registration plate" is classified as a value attribute, and finally, the numeric value "8n6vz" is extracted.

The ontology was designed based on the methodology suggested in [6] and the concepts that are used more often in social networks. A brief extract of the ontology is shown in Figure 7. The ontology describes the nature of the crime context in Mexico City.

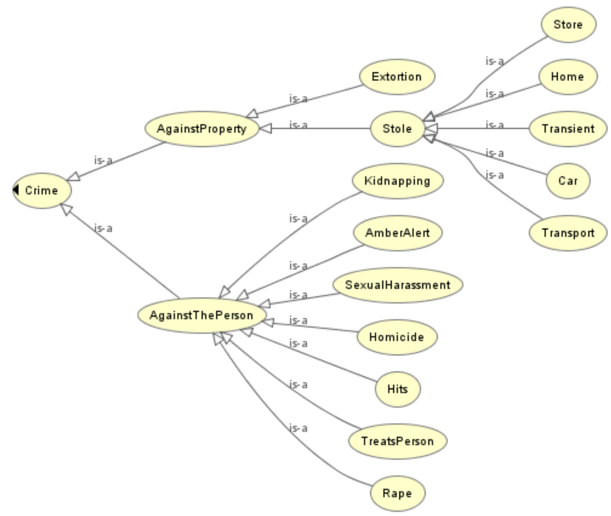


Figure 7. Crime Ontology (fragment)

In addition to this process, a data dictionary is based on a taxonomy that was built using the data from Statistical Classification of Crimes (CED, by the acronym in Spanish). The CED offers classification in three categories: crimes against people, crimes against society, and crimes against the state (they are available at [24]). The taxonomy is applied as a refinement process in the classification of crime type using the OntoClassifier on the data.

VI. Learning from Data Crime

In this stage, characterization is processed and the data is classified using machine learning methods. The main goal of the classification is to get the separation of the crime activity and geographical profile. The characterization is performed by using three classification algorithms: *C5.0 Decision Tree (DT), Naive Bayes (NB), and Support Vector Machine (SVM)*. These algorithms have been applied in previous approaches like [4][12] and represent a layer of supervised learning. In this stage, when a pre-processed crime document (from social networks or web sources) is assigned to a crime class (e.g., carjacking category), two of the three classifiers has to match in the same class. In the case that the three algorithms assign the same object to different classes, the *SVM* result is chosen because this classifier has a better performance according to our tests.

The classification process is executed twice (each execution uses *TD*, *NB and SVM*):

Classification by crime profile: This is the categorization of a new document into a specific class of crime, such as carjacking, assault on passersby, and kidnapping. To improve the visualization of discovered results, a sub-process detects and highlights relevant words inside a classified document, for example, gun, knife, car, violence, thieves (persons, guys), schedule (night, day, morning, time stamp). This sub-process uses a bag of crime words (based on a weighted selection of concepts from our ontology) and considers the lexical and grammatical information of the classified document [13].

Classification by geographical area: in the previous pre-processing phase, a location analysis was developed to obtain a general location (e.g., state or municipality). However, it is necessary to classify the specific location; this allows a correlation between crimes, crowdsourcing complaints, and localities. The classes are known places in municipalities or states: popular zones, tourist sites, buildings, and others.

Finally, the cross-validation is used to estimate how accurate a predictive algorithm will be in practice. In K-fold cross-validation, the sample data are divided into K subsets. One of the subsets is used as test data, and the rest (K-1) as training data. This ensures that the classifications are independent of the partition between training and test data. The preliminary results of K-fold cross-validation show that the SVM algorithm maintains an average accuracy of 87 percent. The DT and SVM algorithms have a performance of around 78.77 percent. However, for small data, NB develops a higher performance than the other two classifiers [4].

Finally, the WEKA software [25] was used to test the performance of the classifiers and make an analysis of the output obtained.

VII. Crossing Data

After identifying the "type of crime," it is possible to correlate this information with other data sources to discover unexpected knowledge. For example, a post extracted from Facebook that belongs to the "carjacking" class and whose text contains the plate number, the framework can use the official database of stolen vehicles (vehicle information system of the Mexican government) [14] and search the number of the "vehicle registration plate." If that number exists, one can conclude whether the car was stolen or not and discover the level of certainty of the information extracted.

Besides using official [15] and collaborative maps, one can discover additional geographical relations around the crime locations. For example, the post with the text "A Honda Civic car was stolen violently near to 'Parque Lindavista' mall by three armed men wearing sports clothes at 9 p.m" (by crime profile, the post is classified as "carjacking" class). "Parque Lindavista" is a known mall of the Delegation Gustavo A. Madero in Mexico City (geographically, the post was classified in the "popular zone" class) and there are hospitals around the mall (crossing with official maps). We discover that carjacking occurs in desolate urban areas. Another observation is that this crime happens in Ecatepec municipality (in the State of Mexico, Mexico) in similar conditions to Delegation Gustavo A. Madero: at night, by two or three armed men, and in desolate zones.

Other patterns regarding carjacking are the number of individuals who are committing this crime (in this case, two or three persons), the use of cars without a number registration plate, and the co-occurrence of crimes like drug trafficking and carjacking. According to the official data, carjacking happens regularly in Mexico City and in the country of Mexico.

VIII. Preliminary Results and Data Visualization

This section outlines and shows the preliminary results of the experiments made in the city of Mexico. The input data includes around 105,036 users, 1,396,408 datasets from Twitter, 25,360 posts from Facebook, and 150,000 from news web pages. The following attributes were considered in the dataset: type of crime, year, month and number of complaints, places where it occurred, the number of users and sources used.

In Figure 8, the areas with a similar unsafe geographical profile are represented in red.

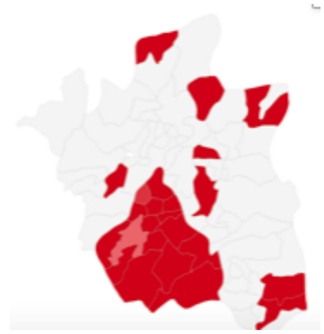


Figure 8. Zones with similar unsafe geographical profile (Mexico City)

While Figure 9 shows a prototype user interface for data visualization and patterns discovery, crimes are also clustered in categories and key terms.

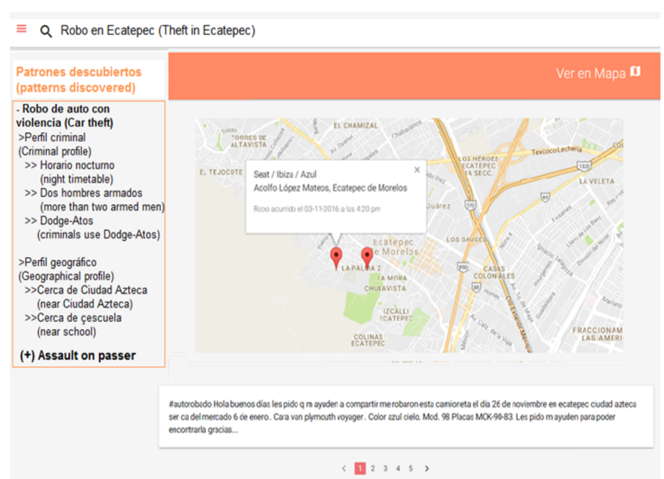


Figure 9. User interface for visualization of geographic patterns

Figure 10 shows how the carjacking crime behaves in the last seven years in the first six months and different geographic places.

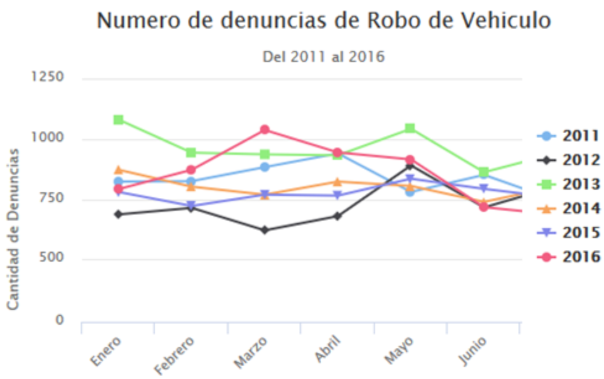


Figure 10. Behavior of crimes in unsafe geographic profiles

Figure 9 and Figure 10 show the integration of official and social media data.

IX. Conclusions

The preliminary research presented in this paper introduces a GKD approach that combines Semantic and Machine Learning techniques, data correlations applied to discover patterns and the emerging levels of data certainty. The preliminary results identify and show the geographical profiles of unsafe places, patterns, and crime behaviors in different geographical areas.

Future work will consider additional classification experiments and depped learning methods using different crime data sources. At the interface level, we plan to develop some user-oriented interfaces designed with usability principles, advanced data visualization, experiments using Stanford API, and mobile applications development.

Acknowledgment

The authors of this paper thank God, CONACYT project number 1051, the Laboratorio de Cómputo Móvil-UPIITA, COFAA-IPN, SIP-IPN project 20171086, and Instituto Politécnico Nacional (IPN) for their support.

References

[1] Semaforo.com.mx. (2017). NGO Crime and Safety Report. [Online] Available at: http://www.semaforo.com.mx/ [Accessed 1 Apr. 2017].

[2] Webmining.cl. (2017). KDD Knowledge Discovery in Databases | WebMining. [online] Available at: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/ [Accessed 1 Apr. 2017].

[3] F. Mata, and C. Claramunt, GeoST: geographic, thematic and temporal information retrieval from heterogeneous web data sources. In Web and Wireless Geographical Information Systems (pp. 5-20). Springer Berlin Heidelberg, 2011.

[4] A. Sampson, Comparing classification algorithms in data mining, Master of Science, Central Connecticut State University, 2016.

[5] R. Zagal-Flores, M. Mata, and C. Claramunt, Geographical Knowledge Discovery applied to the Social Perception of Pollution in the City of Mexico. 9th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. San Francisco, California, USA.

[6] M. Corcho, et al., Methodologies tools and languages for building ontologies. Where is their meeting point? Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Madrid 28660, España, 2001.

[7] C. Jasso López, “Perception of Insecurity in Mexico”, Mexican Journal of Public Opinion, Volume 15, July–December 2013, Pages 12-29, ISSN 1870-7300.

[8] X. Liu, C. Jian, and C. Tien Lu, A spatio-temporal-textual crime search engine. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10). ACM, New York (2010), NY, USA, 528-529.

[9] S. Shah, F. Bao, C. Lu, I. Chen. CROWDSAFE: Crowd Sourcing of Crime Incidents and Safe Routing on Mobile Devices (Demo Paper). ACM SIGSPATIAL GIS’11, November 1-4, 2011. Chicago, IL, USA. ACM, ISBN 978-1-4503-1031-4/11/11.

[10] B. Wang, et al.., An integrated framework for spatio-temporal-textual search and mining. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems (2012), New York, NY, USA, 570-573.

[11] P. Mohan, et al., A neighborhood graph based approach to regional co-location pattern discovery: a summary of results. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '11)

[12] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, Jun 9, 2011 - 744 pages.

[13] G. Sidorov, “Should Syntactic N-grams Contain Names of Syntactic Relations?”. International Journal of Computational Linguistics and Applications, Vol. 5, No. 1, 2014, pp. 139–158.

[14] www2.repuve.gob.mx. (2017). *Consulta Ciudadana*. [online] Available at: http://www2.repuve.gob.mx:8080/ciudadania/ [Accessed 1 Apr. 2017].

[15] inegi.org.mx. (2017). [online] Available at: http://www.inegi.org.mx/geo/contenidos/mapadigital/ [Accessed 1 Apr. 2017].

[16] Secretariadoejecutivo.gob.mx. (2017). Secretariado Ejecutivo :: Inicio. [online] Available at: http://www.secretariadoejecutivo.gob.mx [Accessed 1 Apr. 2017].

[17] Geonames.org. (2017). Geonames. [online] Available at: http://www.geonames.org [Accessed 17 Apr. 2017].

[18] 18. C.-H. Yu, W. Ding, P. Chen, and M. Morabito, “Crime forecasting using spatio-temporal pattern with ensemble learning,” in Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13–16, 2014. Proceedings, Lecture Notes in Computer Science, pp. 174–185.

[19] 19. T. Moon, S. Heo and S. Lee, Ubiquitous Crime Prevention System (UCPS) for a Safer City, Procedia Environmental Sciences, Open Access Journal Volume 22, 2014, ISSN 1878-0296.

[20] J. Ratcliffe, “Crime mapping: spatial and temporal challenges,” in Handbook of Quantitative Criminology, pp. 5–24, Springer, New York, NY, USA, 2010

[21] L. Lei, The GIS-based Research on Criminal Cases Hotspots Identifying, Procedia Environmental Sciences, Open Access Journal, Volume 12, 2012, Pages 957-963, ISSN 1878-0296,

[22] Hitachi Data Systems unveils new advancements in predictive policing to support safer, smart societies." Mena Report, 29/08/2015.

[23] F. Mata, et al., “A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City,” Mobile Information Systems, vol. 2016, Article ID 8068209, 11 pages, 2016. doi:10.1155/2016/806

[24] www3.inegi.org.mx. (2017). Clasificación Estadística de Delitos (CED). [online] Available at: http://www3.inegi.org.mx/sistemas/clasificaciones/delitos.aspx [Accessed 10 Jan. 2017].

[25] cs.waikato.ac.nz. (2017). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. [online] Available at: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 10 Jan. 2017

 ISBN: 978-1-61208-557-9